Complexity

Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022

Lead Guest Editor: Andrea Murari
Guest Editors: Jesus Vega, Gonzalo Farias, and Teddy Craciunescu

Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022

Complexity

Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022

Lead Guest Editor: Andrea Murari
Guest Editors: Jesus Vega, Gonzalo Farias, and Teddy Craciunescu

Contents

Contents

Hindawi
Complexity
Volume 2023, Article ID 8713812, 12 pages
https://doi.org/10.1155/2023/8713812

Research Article

An Alternative Statistical Model to Analysis Pearl Millet (Bajra) Yield in Province Punjab and Pakistan

Muhammad Zeshan Arshad,[1] **Muhammad Zafar Iqbal**,[1] **Festus Were**,[2] **Ramy Aldallal**,[3] **Fathy H. Riad**,[4,5] **M. E. Bakr**,[6] **Yusra A. Tashkandy**,[6] **Eslam Hussam**,[7] **and Ahmed M. Gemeay**[8]

[1]*Department of Mathematics and Statistics, University of Agriculture, Faisalabad 38000, Punjab, Pakistan*
[2]*Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya*
[3]*Department of Accounting, College of Business Administration in Hawtat Bani Tamim, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia*
[4]*Mathematics Department College of Science, Jouf University, P.O. Box 2014, Sakaka, Saudi Arabia*
[5]*Department of Mathematics, Faculty of Science, Minia University, Minia 61519, Egypt*
[6]*Department of Statistics and Operations Research, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia*
[7]*Helwan University, Faculty of Science, Department of Mathematics, Cairo, Egypt*
[8]*Department of Mathematics, Faculty of Science, Tanta University, Tanta 31527, Egypt*

Correspondence should be addressed to Festus Were; were.festus2022@students.jkuat.ac.ke

Received 11 June 2022; Revised 27 August 2022; Accepted 5 April 2023; Published 28 April 2023

Academic Editor: Teddy Craciunescu

Background. A country's agriculture reflects a backbone and performs a vital part in the betterment of the economy and individuals. Facts and figures of the agriculture sector offer a solid foundation and factual pathway intended for upcoming decisions in favor of a country. Accordingly, the probability models have a more significant influence not only in reliability engineering, hydrology, ecology, and medicine but also in agriculture sciences. *Objective*. The primary objective of this study is to propose a reliable and efficient model for pearl millet yield analysis, thereby empowering decision-makers to make informed decisions about their farming practices. With the successful implementation of this model, farmers can potentially increase their pearl millet yield, leading to higher incomes and improved livelihoods for the rural population of Pakistan. *Model*. This study proposes a novel probability model, namely, the alpha transformed odd exponential power function (ATOE-PF) distribution, for analyzing pearl millet yield in Punjab, Pakistan. *Data*. For data collection, two secondary data sets are explored that are electronically available on the site of the Directorate of Agriculture (Economics and Marketing) Punjab, Lahore, Pakistan. *Results*. The maximum likelihood estimation technique is used for estimating the model parameters. For the selection of a better fit model, we follow some accredited goodness of fit tests. The efficiency and applicability of the ATOE-PF distribution are discussed over the province of Punjab (with RMSE = 4.9176) and Pakistan (with RMSE = 4.5849). Better estimates and closest fit to data among the well-established neighboring models offer robust evidence in support of ATOE-PF distribution as well.

1. Introduction

Being an inhabitant of the agricultural country of Pakistan, our masses' primary source of income relies on agriculture. It has a dynamic role in developing the country's foreign exchange, economic growth, and employment. Over the last 40 years, it has had an outstanding contribution to the development of Pakistan's economy [1]. 65% fluctuating share of Pakistan's population, 18.9% gross domestic production (GDP), and

42.3% of the labor force ultimately dependent on agriculture [2]. The total land area of Pakistan is 196.72 million acres, and 66.97 million acres are harvested, along with 20.51 million acres not harvested [3]. Reference [4] categorized Pakistan's crops into food (wheat and rice) and cash food (cotton, maize, sugarcane), as both the crops have a 6.5% contribution to Pakistan's GDP.

One of the oldest cultivated food a pearl millet, which the locals call Bajra. It is a fifth-ranked crop in Pakistan after sorghum, maize, rice, and wheat. This crop is significant for fodder and grain, along with high nutritional contents for poultry and livestock. From 2010 to 2011, this crop yielded 346 thousand tons with a grown area of 548 thousand hectares. However, it was quite an impressive increase (by 18%) as compared to 2009-2010 production [5]. Worldwide, pearl millet's cultivation area is 31 million hectares [6], though, in Pakistan, 0.50 million hectares area along with 0.33 million tones production [7]. Pearl millet's low yield in Pakistan incorporates many factors, including nonstandard crop, inappropriate time of seeding, fluctuating weather intimidations, competitor cereals, and watering issues [8]. Reference [9] explored it as the feeding of the pet birds. It is expected that if Pakistan imports 61,000 tons of pearl millet by 2030, it will be considered the second leading importer country after China [10].

1.1. Probability Models Used for Different Field Crops. Several statistical techniques to model crop yield have been developed and discussed in the past. For this, one can extend his knowledge by reading from [11–32] and many others.

2. Materials and Methods

2.1. Punjab and Pakistan Area and Production. Crop pearl millet has a very high potential of growing with dry heat and drought tolerance along with the low rainfall area (less than 350 mm) circumstances. Consequently, Sindh (Sanghar, Hyderabad, Nawabshah, Kairpur, and Dadu); Punjab (Gujranwala, Bahawalnagar, Rawalpindi, Gujrat, Chakwal, Mianwali, and Attock); Balochistan (Sibbi, Lorali, and Khuzdar); and NWFP (Bannu, D. I. Khan, and Karak) are considered the most suitable and favorable districts (cities) for appropriate cultivation.

Table 1 provides valuable information on the coordinates and region of Pakistan and Punjab. It is a useful resource for researchers and other stakeholders who are interested in understanding the geography and location of the region, and can be used for various analytical and research purposes.

Figure 1: A graphic representation of the area and pearl millet output in Punjab and Pakistan. The figure uses a map of the region along with data on pearl millet production to provide an easy-to-understand overview of the cultivation of pearl millet in this area.

Figure 2: An illustration of the ultimate shape of the pearl millet crop. This figure provides a clear visual reference for the physical appearance of the crop, which can be useful for those who are not familiar with it.

Table 1: Pakistan and Punjab geography condition.

Coordinates	Region
Pakistan	
30.3753°N, 72.7097°E	South Asia
Punjab	
31.17040 N, 72.70970 E	Pakistan

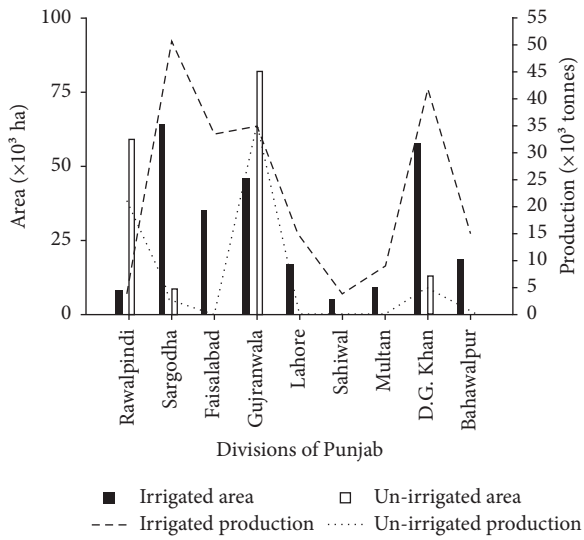
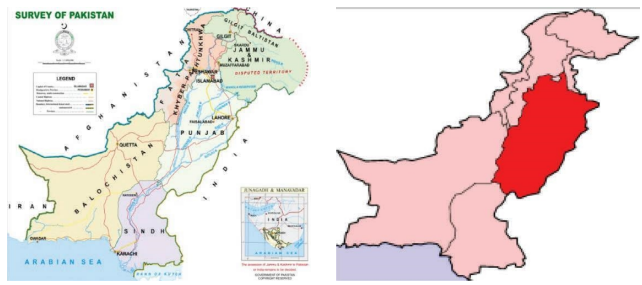


Figure 1: Area and production of pearl millet in Punjab and Pakistan.

Figure 2: Pearl millet picture.

Figure 3 includes two panels; the left panel displays a map of Pakistan, while the right panel displays a map of the Punjab province in Pakistan. The map of Punjab shows the major cities in the province, as well as the locations of pearl millet farms, providing valuable information on the geographic distribution of pearl millet cultivation in the region. The use of a map in this figure helps to provide a clear and visual representation of the information, making it easier for the audience to understand the distribution of pearl millet cultivation in the region.

2.2. Pakistan Climate Conditions. Pakistan experiences a significant amount of climatic variability. Despite the fact the summer months of April to September are fairly nice, the



(a) (b)

Figure 3: Map of Pakistan (a) map of Punjab (b)

winter is brutally chilly in the high mountains in the north and north west. The Indus Valley's plains experience sweltering heat in the summer and freezing conditions in the winter. The southern coastline region experiences a mild climate. Rainfall is generally insufficient. The lower Indus plain's northern regions receive an average annual rainfall of 16 centimeters, whereas the Himalayan area gets an average annual rainfall of 120 centimeters. Rainfall occurs late in the summer and has a monsoonal origin. Humidity is comparatively low because of the heavy rains and wide diurnal temperature fluctuation. High humidity only exists along the coastal strip.

2.3. Punjab Climate Conditions. In the majority of Punjab's regions, the winters are gloomy and frequently rainy. The weather turns springlike by mid-February and stays that way until mid-April, whenever the summer heat arrives. Punjab is expected to experience the start of the monsoon season around May, although the weather has been unpredictable since the early 1970s. Either as the spring monsoon missed the region or it rained so heavily that flooding occurred. It is very hot in June and July. Media sources indicate that the temperature exceeds 51°C and frequently publish stories about persons who have passed away from the heat, despite the fact that official measurements of the temperature seldom go over 46°C. When the temperature reportedly reached 54°C in Multan in June 1993, temperature records were smashed. The "bars" (monsoon season), which give comfort once it passes, interrupt the intense heat in August. Even though the hottest portion of the summer is passed, colder temperatures will not arrive until late October. One of the most frigid winters in the province's recent history dates back more than 70 years. Temperatures in the Punjab area average from −2° to 45°C; However they may get as high as 50°C (122°F) in the summer and as low as −10°C in the winter. Punjab experiences the following three distinct seasons:

(1) Hot weather (April to June), with temperatures reaching 123 degrees Fahrenheit (51 degrees Celsius).

(2) July to September is the rainy season. Average rainfall per year ranges between 96 cm in the submountain region and 46 cm in the plains.

(3) From October to March, the weather can be cold, foggy, or mild. The temperature drops to 35.6 degrees Fahrenheit (2.0 degrees Celsius).

It should be noted that September through October is the ideal time to harvest the crop known as Bajra.

2.4. Climate Prerequisite. It may be sown at low soil temperatures before reaching 23°C. It germinates best in ideal conditions (25–30°C). The vapor pressure deficit (VPD) caused by the daily maximum temperature of 42°C during blooming directly reduces the pearl millet's ability to set seeds [33]. At 40–45°C (base temperature of 10°C), tillering starts with the main tillers regions of the world depend on precipitation, which typically ranges from 150 to 750 mm (350 mm). Because of its resilience to very hot and dry weather conditions is becoming increasingly important in developing climate-resilient agricultural systems under changing climatic scenarios [34]. The pearl millet requires between 300 and 350 mm of rainfall to thrive. It is important to note that the water requirement of a crop can vary depending on various factors such as soil type, climate, and cultivation practices. The Figure 4 presented in the chart should therefore be considered as general guidelines rather than exact values.

2.5. Data Collection. For this study, we consider secondary data sets. For this, the first data presents the average yield of Bajra in Punjab (1947-48 to 2017-18) (Per Acre/000 Tonnes), and the second data relates to the average yield (Per Acre/000 Tonnes) of Bajra in Pakistan (1947-48 to 2017-18). The datasets are obtained from the agricultural statistics of Pakistan and are available at the electronic address provided in Appendix.

2.6. Model Description. In this paper, we develop a novel two-parameter probability model that performs so well not only in reliability engineering, hydrology, ecology, and medical sciences but has a vital role in agriculture sciences as well. We refer to it as the alpha transformed odd exponential power function (ATOE-PF) distribution. The associated cumulative distribution function (CDF) corresponding to the probability density function (PDF) along with the

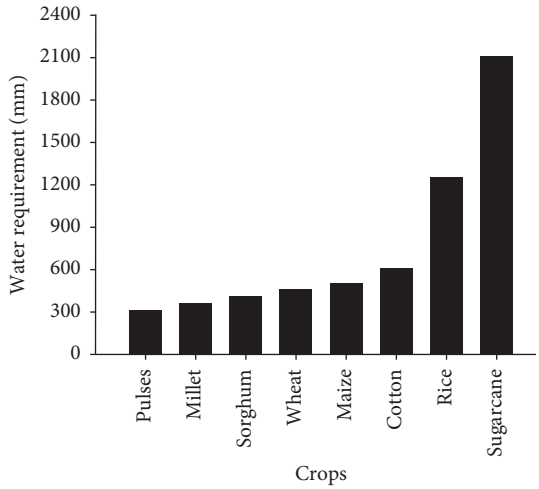


Figure 4: Water requirement of pearl millet in comparison with other crops.

quantile function is, respectively, given by the following equation:

$$F(x) = \frac{\alpha^{e^{\left(1-(g_n/x)^{\beta}\right)}} - 1}{\alpha - 1},$$

$$f(x) = \frac{(g_n)^{\beta}\beta \log \alpha}{\alpha - 1} x^{-(\beta+1)} e^{\left(1-(g_n/x)^{\beta}\right)} \alpha^{e^{\left(1-(g_n/x)^{\beta}\right)}}, \quad (1)$$

$$x_q = \frac{g_n}{\left[1 - \log\left[(1/\log\alpha)[\log[1 + q[\alpha - 1]]]\right]\right]^{1/\beta}},$$

where $0 < x \le g_n$ and $\alpha > 0, \alpha > 1, \beta > 0$, are two shape parameters.

Note that, the ATOE-PF distribution is one of the particular members of the ATOE-G class of distributions. Therefore, this paper uses ATOE-PF distribution as a modeling framework, and our ongoing project's advanced complementary mathematical and reliability measures are under-processed.

2.7. Parameter Estimation. We use the maximum likelihood estimation technique for the parameter estimation of the ATOE-PF distribution. For this, we suppose $X_1, X_2, ..., X_n$ be a random sample of size n taken from X, then the log-likelihood function (Log L) of X is given by the following equation:

$$\text{Log}\, L = \left[\begin{array}{l} n\beta \log(g_n) + n\log(\beta) + n\log[\log(\alpha)] - n\log(\alpha - 1) - \\ \\ (\beta + 1)\sum_{i=1}^{n} \log(x_i) + \sum_{i=1}^{n}\left[1 - \left(\frac{g_n}{x_i}\right)^{\beta}\right] + \log(\alpha)\sum_{i=1}^{n} e^{\left[1-\left(\frac{g_n}{x_i}\right)^{\beta}\right]} \end{array}\right]. \quad (2)$$

The partial derivatives of Log L for the parameters α and β are given by, respectively,

$$\frac{\partial l}{\partial \alpha} = \frac{n\alpha}{\log(\alpha)} - \frac{n}{(\alpha - 1)} + \frac{1}{\alpha}\sum_{i=1}^{n} e^{\left[1-(g_n/x_i)^{\beta}\right]}, \text{and}$$

$$\frac{\partial l}{\partial \beta} = n\log(x_n) + \frac{n}{\beta} - \sum_{i=1}^{n} \partial - \sum_{i=1}^{n}\left(\frac{g_n}{x_i}\right)^{\beta} log\left(\frac{g_n}{x_i}\right) + \left[e^{\left[1-(g_n/x_i)^{\beta}\right]}\right]\left(\frac{g_n}{x_i}\right)^{\beta} \log\left(\frac{g_n}{x_i}\right). \quad (3)$$

The ML estimates $(\widehat{\phi} = \widehat{\alpha}|^{\text{MLE}}, \widehat{\beta}|^{\text{MLE}})$ of the ATOE-PF distribution are derived by maximizing (2) or by solving the above nonlinear equations simultaneously. The following part has a detailed simulation with various parameter configurations to test the asymptotic capability of MLEs.

2.8. Simulation Study. The following algorithm discusses the performance of MLEs with the assistance of a simulation study:

Step-1: a random sample $x_1, x_2, x_3, \ldots, x_n$ of sizes $n = 100$, 150, 200, 250, 300, 350, 400, 450, and 500 are generated from $Q(q)$.

Step-2: the required results are obtained based on the different combinations of the model parameters for $g_n = 2$, placed in S-I ($\alpha = 1.9, \beta = 1.5$), S-II ($\alpha = 1.1, \beta = 2.5$), S-III ($\alpha = 1.5, \beta = 1.5$), S-IV ($\alpha = 1.2, \beta = 1.9$), S-V ($\alpha = 1.3, \beta = 1.7$), S-VI ($\alpha = 1.7, \beta = 3.9$), S-VII ($\alpha = 1.15, \beta = 4.75$), S-VIII ($\alpha = 1.25, \beta = 7.75$), and S-IX ($\alpha = 1.55, \beta = 5.95$)

Step-3: average estimate (AE), bias, mean square error (MSE), and variance (Var) are presented in Tables 2–4.

Step−4: each sample is replicated $N = 1000$ times.

Step-5: gradual decrease in AE(s), bias(es), MSE(s), and Var(s) with increases in the sample size is observed.

Step-6: finally, the estimates in Tables 2–4 help us specify that the method of maximum likelihood works consistently for the ATOE-PF distribution.

Note that, Figure 5 is a useful visual representation of the density function curves for various choices of model parameters for simulated data. The figure provides researchers with valuable insights into the impact of different parameter values on the shape of the distribution, enabling them to make more informed modeling decisions.

$$AE(\widehat{\Xi}) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\Xi}_i \quad \text{Bias}(\widehat{\Xi}) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\Xi}_i - \Xi\right),$$

$$\text{MSE}(\widehat{\Xi}) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\Xi}_i - \Xi\right)^2, \tag{4}$$

$$\text{Var}(\widehat{\Xi}) = \frac{1}{N}\sum_{i=1}^{N}\left(\Xi - \overline{\Xi}_i\right)^2.$$

3. Results and Discussions

Now, we report the application of the ATOE-PF distribution. For this, we focus on the agricultural sector and engage two suitable datasets. The ATOE-PF distribution is compared with well-known competitive models. The CDFs of competitive models are listed in Table 5. The parameter estimates and standard errors are presented in Tables 6 and 7 for both datasets, respectively. Some typical results from descriptive statistics for both datasets are shown in Tables 8 and 9. These descriptive statistics are minimum value, 1st quartile, mean, median, mode, standard deviation (SD), 3rd quartile, maximum value, 90%, 95%, and 99% confidence intervals.

The goodness-of-fit statistics for the ATOE-PF distribution and competing models are presented in Tables 10 and 11. A better fit model is one with the criteria of a minimum value of Anderson–Darling (AD), Cramer-von Mises (CVM), root mean square error (RMSE), and Kolmogorov–Smirnov (KS) with a higher P-value. Please note that a comprehensive list of standard measurement units and corresponding abbreviations can be found in Table 12 of this document.

Table 2: Bias, mean square error, variance, and average estimate.

n	Est	S-I		S-II		S-III	
		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
100	Bias	0.1591	0.0349	0.2141	0.0136	0.1573	0.0212
	MSE	1.6324	0.0237	0.3585	0.0376	0.8554	0.0178
	Var	1.6071	0.0225	0.3126	0.0374	0.8306	0.0173
	AE	2.0591	1.5349	1.3141	2.5136	1.6573	1.5212
150	Bias	0.0598	0.0288	0.1394	−0.0038	0.0735	0.0192
	MSE	0.9004	0.0168	0.1808	0.0221	0.4627f	0.0127
	Var	0.8968	0.0159	0.1613	0.0221	0.4573	0.0123
	AE	1.9598	1.5288	1.2394	2.4961	1.5735	1.5192
200	Bias	0.0412	0.0227	0.1122	−0.0091	0.0496	0.0166
	MSE	0.5669	0.0134	0.1113	0.0175	0.2968	0.0105
	Var	0.5652	0.0129	0.0987	0.0174	0.2943	0.0103
	AE	1.9412	1.5227	1.2122	2.4908	1.5496	1.5165
250	Bias	0.0165	0.0196	0.0898	−0.0089	0.0269	0.0153
	MSE	0.4386	0.0108	0.0835	0.0143	0.2315	0.0089
	Var	0.4383	0.0105	0.0754	0.0143	0.2308	0.0087
	AE	1.9165	1.5196	1.1898	2.4911	1.5269	1.5152
300	Bias	0.0135	0.0183	0.0790	−0.0055	0.0213	0.0134
	MSE	0.3532	0.0088	0.0675	0.0123	0.1894	0.0068
	Var	0.3530	0.0085	0.0613	0.0122	0.1889	0.0067
	AE	1.9135	1.5183	1.1790	2.4944	1.5213	1.5134
350	Bias	0.0188	0.0148	0.0733	−0.0048	0.0226	0.0112
	MSE	0.3017	0.0075	0.0589	0.0108	0.1647	0.0059
	Var	0.3014	0.0073	0.0535	0.0108	0.1642	0.0058
	AE	1.9188	1.5148	1.1733	2.4951	1.5226	1.5112
400	Bias	0.0116	0.0142	0.0679	−0.0022	0.0159	0.0110
	MSE	0.2731	0.0066	0.0525	0.0098	0.1503	0.0053
	Var	0.2730	0.0064	0.0479	0.0098	0.1501	0.0052
	AE	1.9116	1.5142	1.1679	2.4977	1.5159	1.5110
450	Bias	−0.0004	0.0137	0.5781	0.0012	0.0051	0.0110
	MSE	0.2386	0.0059	0.0453	0.0091	0.1322	0.0047
	Var	0.2386	0.0057	0.0419	0.0091	0.1322	0.0046
	AE	1.8995	1.5137	1.1578	2.5012	1.5051	1.5110
500	Bias	0.0029	0.0109	0.0531	0.0024	0.0064	0.0087
	MSE	0.2247	0.0051	0.0434	0.0085	0.1248	0.0042
	Var	0.2247	0.0050	0.0405	0.0085	0.1248	0.0042
	AE	1.9029	1.5109	1.1531	2.5024	1.5064	1.5087

The agriculture sector plays a crucial role in the economy of a country, and the ability to accurately predict crop yields is of utmost importance. In order to aid decision-makers in the farming industry, a new probability model was developed that is capable of accurately modeling agriculture data. This study utilized secondary data on pearl millet (Bajra) yields in Punjab Province, Pakistan and compared the alpha transformed odd exponential power function (ATOE-PF) distribution to its well-established rivals using various goodness of fit tests such as KS (P-value), AD, and CVM. The ATOE-PF distribution showed a better fit for the average yield of pearl millet (Bajra) in Punjab and Pakistan than any of its competitors. The P value (KS) was higher for the ATOE-PF distribution, indicating that it meets the minimal statistical value requirement for a better fit model.

Table 3: Bias, mean square error, variance, and average estimate.

n	Est	S-IV		S-V		S-VI	
		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
100	Bias	0.1932	0.0039	0.1763	0.0105	0.1536	0.0731
	MSE	0.4550	0.0203	0.5708	0.0188	1.2054	0.1398
	Var	0.4177	0.0203	0.5397	0.0187	1.1818	0.1345
	AE	1.3932	1.9039	1.4763	1.7105	1.8536	3.9731
150	Bias	0.1147	0.0035	0.0963	0.0109	0.0624	0.0627
	MSE	0.2349	0.0149	0.3008	0.0134	0.6633	0.1001
	Var	0.2217	0.0149	0.2915	0.0133	0.6594	0.0961
	AE	1.3147	1.9035	1.3963	1.7109	1.7624	3.9627
200	Bias	0.0538	0.0098	0.0690	0.0105	0.0424	0.0500
	MSE	0.0926	0.0087	0.1920	0.0112	0.4214	0.0804
	Var	0.0897	0.0086	0.1872	0.0110	0.4196	0.0779
	AE	1.2538	1.9098	1.3690	1.7105	1.7424	3.9500
250	Bias	0.0502	0.0080	0.0471	0.0112	0.0189	0.0463
	MSE	0.0811	0.0079	0.1480	0.0093	0.3280	0.0662
	Var	0.0785	0.0079	0.1458	0.0092	0.3276	0.0640
	AE	1.2502	1.9080	1.3471	1.7112	1.7189	3.9463
300	Bias	0.0444	0.0070	0.0371	0.0134	0.0154	0.0432
	MSE	0.0733	0.0069	0.1226	0.0083	0.2659	0.0545
	Var	0.0713	0.0069	0.1212	0.0082	0.2657	0.0527
	AE	1.2444	1.9070	1.3371	1.7134	1.7154	3.9432
350	Bias	0.0326	0.0055	0.0362	0.0082	0.0186	0.3610
	MSE	0.0645	0.0055	0.1069	0.0064	0.2295	0.0476
	Var	0.0634	0.0055	0.1056	0.0063	0.2291	0.0463
	AE	1.2326	1.9055	1.3362	1.7082	1.7186	3.9361
400	Bias	0.0035	0.0035	0.0287	0.0085	0.0120	0.0348
	MSE	0.0051	0.0051	0.0979	0.0057	0.2083	0.0423
	Var	0.0604	0.0051	0.0979	0.0057	0.2081	0.0411
	AE	1.2301	1.9035	1.3287	1.7085	1.7120	3.9348
450	Bias	0.0035	0.0035	0.0173	0.0086	0.0012	0.0331
	MSE	0.0051	0.0051	0.0866	0.0051	0.1819	0.0363
	Var	0.0604	0.0051	0.0863	0.0050	0.1819	0.0351
	AE	1.2301	1.9035	1.3173	1.7086	1.7012	3.9331
500	Bias	0.0035	0.0035	0.0174	0.0065	0.0032	0.0264
	MSE	0.0051	0.0051	0.0815	0.0047	0.1721	0.3221
	Var	0.0604	0.0051	0.0126	0.0046	0.1721	0.0315
	AE	1.2301	1.9035	1.3174	1.7065	1.7031	3.9264

Table 4: Bias, mean square error, variance, and average estimate.

n	Est	S-VII		S-VIII		S-IX	
		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
100	Bias	0.2029	0.0131	0.1842	0.0434	0.1552	0.0932
	MSE	0.4038	0.1241	0.5102	0.3326	0.9364	2.8667
	Var	0.3626	0.1239	0.4763	0.3308	0.9123	0.2778
	AE	1.3529	4.7630	1.4342	7.7934	1.7052	6.0432
150	Bias	0.1262	−0.0010	0.1047	0.0409	0.0696	0.0834
	MSE	0.2064	0.0766	0.2666	0.2558	0.5094	0.2144
	Var	0.1904	0.0766	0.2556	0.2541	0.5045	0.2075
	AE	1.2762	4.7489	1.3547	7.7909	1.6196	6.0334
200	Bias	0.0987	−0.0015	0.0773	0.0329	0.0469	0.0657
	MSE	0.1283	0.0666	0.1689	0.2141	0.3262	0.1728
	Var	0.1185	0.0666	0.1630	0.2130	0.3240	0.1685
	AE	1.2487	4.7484	1.3273	7.7829	1.5969	6.0157

Table 4: Continued.

n	Est	S-VII		S-VIII		S-IX	
		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
250	Bias	0.0761	0.0055	0.0551	0.0264	0.0241	0.0590
	MSE	0.0972	0.0561	0.1296	0.1675	0.2545	0.1394
	Var	0.0914	0.0561	0.1267	0.1669	0.2539	0.1359
	AE	1.2261	4.7555	1.3051	7.7764	1.5741	6.0090
300	Bias	0.0651	0.0148	0.0445	0.0361	0.0194	0.0567
	MSE	0.0793	0.0489	0.1071	0.1434	0.2074	0.1140
	Var	0.0751	0.0487	0.1051	0.1421	0.2070	0.1108
	AE	1.2151	4.7648	1.2945	7.7861	1.5669	6.0067
350	Bias	0.0605	0.0119	0.0421	0.0269	0.0211	0.0462
	MSE	0.0694	0.0449	0.0937	0.1249	0.1803	0.0972
	Var	0.0657	0.0448	0.0919	0.1242	0.1799	0.0950
	AE	1.2105	4.7619	1.2921	7.7769	1.5711	5.9962
400	Bias	0.0551	0.0088	0.0356	0.0294	0.0146	0.0465
	MSE	0.0622	0.0391	0.0852	0.1134	0.1640	0.0879
	Var	0.0591	0.0391	0.0839	0.1125	0.1638	0.8575
	AE	1.2051	4.7588	1.2856	7.7794	1.5646	5.9965
450	Bias	0.0439	0.0084	0.0236	0.0311	0.0038	0.0460
	MSE	0.0543	0.0333	0.0755	0.0967	0.1441	0.0772
	Var	0.0524	0.0332	0.0749	0.0958	0.1441	0.0750
	AE	1.1939	4.7584	1.2736	7.7810	1.5539	5.9960
500	Bias	0.0401	0.0026	0.0230	0.0235	0.0053	0.0371
	MSE	0.0519	0.0294	0.0713	0.0891	0.1361	0.0697
	Var	0.0503	0.0295	0.0707	0.0885	0.1361	0.0683
	AE	1.1900	4.7526	1.2730	7.7735	1.5553	5.9871

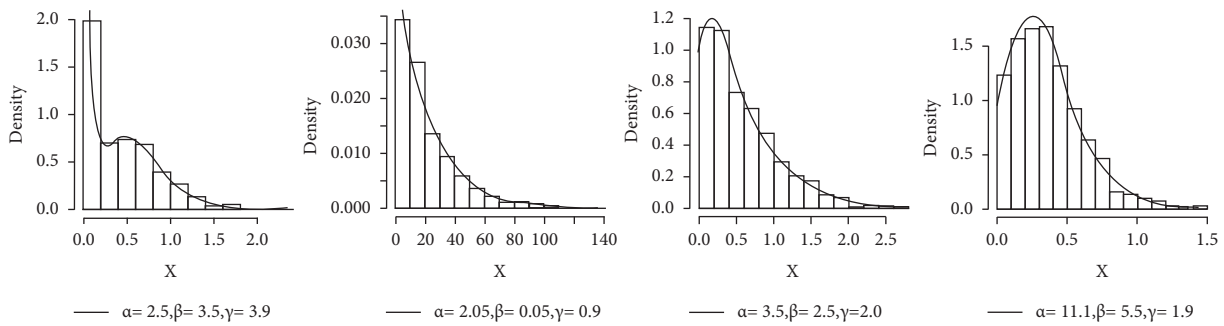


Figure 5: Density function curves for various choices of model parameters for simulated data.

Table 5: List of some competitive model's cumulative distribution functions.

Models	CDF's of model	Parameters	Support	Positions
HL-Exp	$P(x) = 1 - e^{-\alpha x}/1 + e^{-\alpha x}$	$\alpha > 0$	$0 < x < \infty$	Shape = α
Exp	$P(x) = 1 - e^{-\alpha x}$	$\alpha > 0$	$0 < x < \infty$	Shape = α
MO-Exp	$P(x) = 1 - \alpha e^{-x}/1 - (1-\alpha)e^{-x}$	$\alpha > 0$	$0 < x < \infty$	Scale = α
NH-Exp	$P(x) = 1 - e^{1-(1+\alpha x)^{\beta}}$	$\alpha, \beta > 0$	$0 < x < \infty$	Scale = α Shape = β
Exp-Exp	$P(x) = e^{(1-e^{-\alpha x})} - 1/e - 1$	$\alpha > 0$	$0 < x < \infty$	Shape = α
Alp-Exp	$P(x) = \alpha^{(1-e^{-\beta x})} - 1/\alpha - 1$	$\alpha, \beta > 0$	$0 < x < \infty$	Scale = α Shape = β
Pareto	$P(x) = 1 - (m_0/x)^{\alpha}$	$\alpha > 0$	$m_0 \le x < \infty$	Shape = α Reflected = m_0
Gompertz	$P(x) = 1 - e^{-\alpha(e^{\beta x}-1)}$	$\alpha, \beta > 0$	$0 < x < \infty$	Shape = α Shape = β
Normal	$P(x) = \Phi(x - \alpha/\beta)$	$\alpha, \beta > 0$	$-\infty < x < \infty$	Location = α Scale = β

Table 5: Continued.

Models	CDF's of model	Parameters	Support	Positions
Burr-XII	$P(x) = 1 - (1 + x^{\alpha})^{-\beta}$	$\alpha, \beta > 0$	$0 < x < \infty$	Shape = α Shape = β
PF	$P(x) = (x/g_n)^{\alpha}$	$\alpha > 0$	$0 < x \leq g_n$	Shape = α

Table 6: Parameter estimates and standard errors for average yield (per acre) of Bajra in province Punjab.

Models	$\widehat{\alpha}$		$\widehat{\beta}$	
	Estimate	Std. error	Estimate	Std. error
ATOE-PF	0.0324	0.0216	2.8176	0.1534
Normal	5.2154	0.0939	0.7914	0.0664
Gompertz	0.0027	0.0005	1.0561	0.0336
MO-Exp	178.85	32.008	—	—
PF	2.9609	0.3514	—	—
Pareto	2.5048	0.2973	—	—
Alp-Exp	143.58	56.144	0.4184	0.0284
NH-Exp	0.0045	0.0020	33.936	14.896
HL-Exp	0.2947	0.0270	—	—
Exp-Exp	0.2583	0.0253	—	—
Exp	5.2164	0.6192	—	—
B-XII	6.4565	12.036	0.0944	0.1764

Table 7: Parameter estimates and standard errors for average yield (per acre) of Bajra in Pakistan.

Models	$\widehat{\alpha}$		$\widehat{\beta}$	
	Estimate	Std. error	Estimate	Std. error
ATOE-PF	0.0371	0.0244	2.4341	0.1348
Normal	4.8428	0.1018	0.8580	0.0720
Gompertz	0.0048	0.0015	1.0144	0.0550
MO-Exp	122.02	22.020	—	—
PF	2.5878	0.3071	—	—
Pareto	2.1107	0.2505	—	—
Alp-Exp	263.67	110.42	0.4724	0.0309
NH-Exp	0.0047	0.0023	34.591	16.662
HL-Exp	0.3168	0.0291	—	—
Exp-Exp	0.2779	0.0273	—	—
Exp	4.8434	0.5749	—	—
B-XII	7.5194	19.042	0.0851	0.2158

Table 8: Descriptive statistics for average yield (per acre) of Bajra in province Punjab.

Data	Minimum	1st quartile	Mean	Median	Mode	SD	3rd quartile	Maximum
	3.510	4.635	5.216	5.100	5.700	0.797	5.585	7.230
		Confidence interval			Skewness		Kurtosis	
Punjab		90%	(5.057, 5.373)					
		95%	(5.027, 5.404)		0.611		3.033	
		99%	(4.965, 5.466)					

The empirical fitted PDF, CDF, Probability-Probability, and box plots of the ATOE-PF distribution are presented in Figures 6 and 7, which visually demonstrate the model's adequacy. All numerical results and model estimates were obtained using the free statistical software R Studio version 1.2.5033 (cited therein) and its exclusive package AdequacyModel. This new probability model provides decision-makers in the farming industry with a reliable tool to aid in predicting crop yields. By utilizing the ATOE-PF distribution, farmers and related departments can begin implementing more effective predictive measures. The model's superiority over its competitors in accurately

Table 9: Descriptive statistics for average yield (per acre) of Bajra in Pakistan.

Data	Minimum	1st quartile	Mean	Median	Mode	SD	3rd quartile	Maximum
	3.020	4.305	4.843	4.780	6.380	0.863	5.075	7.020
		Confidence interval			Skewness		Kurtosis	
Pakistan	90%	(4.672, 5.014)						
	95%	(4.638, 5.047)			0.646		3.159	
	99%	(4.571, 5.114)						

Table 10: The goodness of fit statistics for average yield (per acre) of Bajra in province Punjab.

Models	CVM	AD	K-S	K-S (*P*-val)	RMSE
ATOE-PF	0.0763	0.5023	0.0766	0.7981	4.9176
Normal	0.1541	1.0181	0.0888	0.6305	4.9282
Gompertz	0.4668	2.8004	0.1615	0.0493	4.9946
MO-Exp	0.1379	0.9312	0.2264	0.0014	5.0594
PF	1.9546	10.6645	0.3287	0.0012	5.0473
Pareto	0.1194	0.8793	0.3497	0.0009	5.0903
Alp-Exp	0.0887	0.6193	0.3906	0.0111	5.1469
NH-Exp	0.1352	0.9010	0.5483	0.0015	5.1801
HL-Exp	0.0989	0.6814	0.5107	0.0080	5.1921
Exp-Exp	0.0945	0.6542	0.5072	0.0050	5.1957
Exp	0.0886	0.6172	0.5153	0.0009	5.2064
B-XII	0.0454	0.3479	0.5498	0.0010	5.1197

Table 11: The goodness of fit statistics for average yield (per acre) of Bajra in Pakistan data.

Models	CVM	AD	K-S	K-S (*P*-val)	RMSE
ATOE-PF	0.1940	1.0670	0.1209	0.2497	4.5849
Normal	0.2885	1.7279	0.1515	0.0768	4.5966
Gompertz	0.6920	3.8658	0.2004	0.0067	4.6576
MO-Exp	0.2556	1.5727	0.2068	0.0046	4.7069
PF	1.9931	10.7507	0.3356	0.0051	4.7037
Pareto	0.2897	1.6829	0.3545	0.0161	4.7515
Alp-Exp	0.1966	1.1906	0.3530	0.0111	4.7824
NH-Exp	0.2629	1.5708	0.5265	0.0150	4.8169
HL-Exp	0.2106	1.2712	0.4931	0.0115	4.8295
Exp-Exp	0.2048	1.2361	0.4906	0.0111	4.8333
Exp	0.1973	1.1891	0.5012	0.0109	4.8446
B-XII	0.1500	0.8773	0.5352	0.0011	4.9182

Table 12: List of standard measurement units and other abbreviations.

Full names	Symbol/description
Half logistic exponential	HL-Exp
Nadarajah-Haghighi exponential	NH-Exp
Marshall-Olkin exponential	MO-Exp
Alpha power exponential	Alp-Exp
Exponential-exponential	Exp-Exp
Power function	PF
Exponential	Exp
Per acre, thousand tonnes	Per acre/000 tonnes
Anderson-darling test	AD = a test to detect the departure of sample distribution from normality
Kolmogorov-Smirnov test	K-S = a test to detect the departure of CDF from empirical CDF
Cramer-von mises test	CVM = a test to compare two empirical CDFs
Root mean square error	RMSE = a measure to describe data around the best fit line
Mean square error	MSE = a risk function, which measures the discrepancy between the estimated and real values

Table 12: Continued.

Full names	Symbol/description
Bias	Bias = the term "bias" refers to a consistent departure from the true value. It is the discrepancy between the parameter's actual value and its intended value
Variance	Var = the term "variance" refers to measuring how widely apart a group of numbers is from another
Average estimate	AE = a point estimate of a mean of an unknown distribution

Electronic address of data set https://www.amis.pk/Agristatistics/Data/HTML%20Final/Bajra/Production.html.

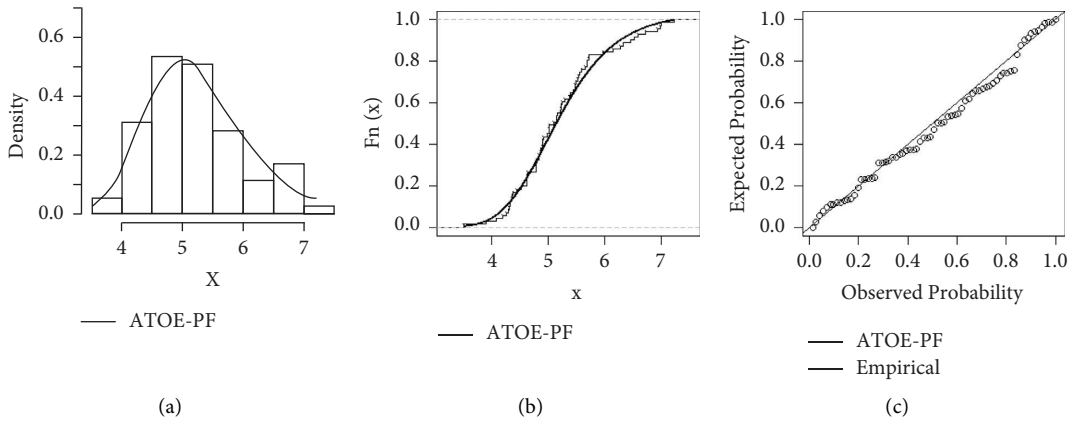


Figure 6: Empirically fitted plots for an average yield of Bajra in Punjab, Pakistan.

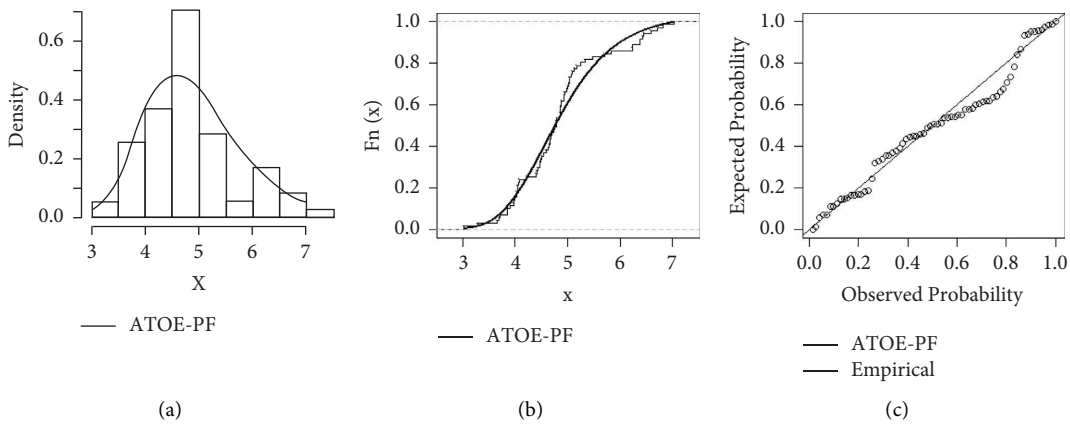


Figure 7: Empirically fitted plots for an average yield of Bajra in Pakistan.

modeling agriculture data provides valuable information for agriculture bodies. In addition, the use of various goodness of fit tests ensures that the model provides an adequate fit. Overall, the ATOE-PF distribution presents a promising solution for researchers and practitioners in the agriculture sector.

4. Conclusions

In this work, a novel model called the alpha transformed odd exponential power function (ATOE-PF) distribution was established, and we introduced its PDF and CDF. A simulation study was carried out using the maximum likelihood estimation technique. To prove the superiority of the proposed model, we fitted two pearl millet datasets. The ATOE-PF distribution was considered the best fit model among the well-known rivals after passing the various goodness of fit tests. Referring to Tables 10 and 11, we found that the (ATOE-PF) distribution has the lowest K-S value and the highest P value, proving the ATOE-PF distribution's superiority. The efficiency and applicability of the ATOE-PF distribution are discussed over the provinces of Punjab (with RMSE = 4.9176) and Pakistan (with RMSE = 4.5849). Furthermore, outperforming estimates made it more relevant and encouraging for pearl millet farm decision-makers and other agriculture agencies.

5. Future Directions

The proposed technique would hopefully be adopted by agriculture experts and concerned agencies and

implemented on maize, soybeans, rice, sugarcane, cotton, moong, mash, and jowar for a more appropriate prediction and a respectable predicted yield. Also, we have another critical future work: the study of COVID-19 infections and the mortality rate of the infected. Another expansion will be the competing risk resulting from death, whether it is from the disease or another cause.

Appendix

The first data presents the average yield of Bajra in Punjab (1947-48 to 2017-18) (Per Acre/000 Tonnes). 4.79, 4.64, 4.84, 4.92, 3.90, 3.51, 4.80, 4.30, 4.26, 4.31, 4.09, 4.44, 4.39, 4.62, 5.02, 5.48, 4.90, 5.16, 4.88, 4.91, 5.22, 4.63, 4.84, 5.02, 5.16, 5.24, 5.10, 4.99, 5.25, 5.23, 5.54, 5.30, 5.40, 5.38, 5.45, 5.48, 5.60, 5.66, 5.53, 5.70, 4.62, 4.20, 4.34, 4.37, 4.36, 4.34, 4.53, 4.63, 4.79, 4.81, 4.87, 4.91, 5.03, 5.57, 5.17, 5.50, 5.72, 5.70, 5.73, 5.98, 6.15, 6.47, 6.28, 6.94, 6.99, 7.00, 6.54, 6.59, 6.34, 6.73, 7.23.

The second data relates to the average yield (Per Acre/000 Tonnes) of Bajra in Pakistan (1947-48 to 2017-18). 3.7, 3.65, 3.91, 4.02, 3.28, 3.02, 4.47, 3.98, 3.86, 3.97, 3.71, 3.86, 4.07, 4.08, 4.43, 4.94, 4.86, 4.88, 4.39, 4.41, 4.51, 4.47, 4.76, 4.78, 4.79, 5.03, 4.85, 4.93, 4.99, 4.85, 5.02, 4.88, 5.00, 5.33, 4.93, 5.08, 4.69, 4.74, 4.66, 4.63, 4.68, 3.99, 4.04, 4.04, 4.49, 4.22, 4.59, 4.54, 4.02, 4.86, 4.65, 4.66, 5.03, 5.17, 5.25, 5.48, 5.13, 5.70, 5.07, 4.78, 5.82, 6.38, 6.23, 6.38, 6.71, 6.81, 6.42, 6.45, 6.24, 6.58, 7.02.

Data Availability

The data used to support the study are included in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported via funding from Prince sattam bin Abdulaziz University project number (PSAU/2023/R/1444).

References

[1] A. G. Awan and A. Alam, "Impact of agriculture productivity on economic growth: a case study of Pakistan," *Industrial Engineering Letters*, vol. 5, no. 7, pp. 27–33, 2015.

[2] A. Azam and M. Shafique, "Agriculture in Pakistan and its impact on economy," *Rev. Int. J. Adv. Sci. Tech*, vol. 103, pp. 47–60, 2017.

[3] F. Bhatti, "A model based study of rice production: a case study of district Lodhran," M. Phil thesis, Dep. Stat. Islamia Univ. Bahawalpur, Pakistan, 2015.

[4] M. A. Raza, L. Y. Fenga, A. Manaf, A. Wasaya, and M. Ansar, "Sulphur application increases seed yield and oil content in sesame seeds under rainfed conditions," *Field Crops Research*, vol. 51, p. 58, 2018.

[5] Anonymous, *Agricultural Statistics of Pakistan*, Food and Agriculture Division. Planning Unit, Islamabad, Pakistan, 2011.

[6] Icrisat, "International crops research institute for the semiarid tropics," 2016, https://www.cgiar.org/research/center/icrisat/.

[7] Gop (Government of Pakistan), *Economic Survey of Pakistan*, Ministry of Finance, Pakistan, 2015.

[8] M. Ayub, M. A. Nadeem, A. Tanveer, M. Tahir, and R. M. A. Khan, "Interactive effect of different nitrogen levels and seeding rates on fodder yield and quality of pearl millet," *Pakistan Journal of Agricultural Sciences*, vol. 44, pp. 592–596, 2007.

[9] S. R. Chughtai, J. Fateh, M. H. Munawwar, M. Aslam, and H. N. Malik, "Alternative uses of cereals-methods and feasibility: Pakistan perspective," *CFC and ICRISAT, 2004. Alternative uses of Sorghum and Pearl Millet in Asia: Proc. Expert Meeting, ICRISAT, Patancheru, Andhra Pradesh, India*, vol. 34, pp. 210–220, 2004.

[10] S. Nedumaran, P. Abinaya, and M. C. S. Bantila, "Sorghum and millets futures in asia under changing socio-economic and climatic scenarios," *ICRISAT Socio-Economic Discussion Paper Series*, International Crops Research Institute for the Semi-arid Tropics, Patancheru, India, 2013.

[11] R. H. Day, "Probability distributions of field crop yields," *Journal of Farm Economics*, vol. 47, no. 3, pp. 713–741, 1965.

[12] J. R. M. Hosking, J. R. Wallis, and E. F. Wood, "Estimation of the generalized extreme value distribution by the method of probability-weighted moments," *Technometrics*, vol. 27, no. 3, pp. 251–261, 1985.

[13] O. A. Ramirez, "Estimation and use of a multivariate parametric model for simulating heteroskedastic, correlated, non Normal random variables: the case of corn belt corn, soybean and wheat yields," *American Journal of Agricultural Economics*, vol. 79, no. 1, pp. 191–205, 1997.

[14] B. K. Goodwin and A. P. Ker, "Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts," *American Journal of Agricultural Economics*, vol. 80, no. 1, pp. 139–153, 1998.

[15] J. Atwood, S. Shaik, and M. Watts, "Are crop yields normally distributed? A reexamination," *American Journal of Agricultural Economics*, vol. 85, no. 4, pp. 888–901, 2003.

[16] M. H. Munawwar, J. Fateh, H. I. Javed, H. N. Malik, and M. Hussain, "Stability analysis of Millet varieties across diverse environments in Pakistan," *Sarhad Journal of Agriculture*, vol. 3, pp. 645–648, 2007.

[17] T. Górski, "A probability distribution for crop yields in Poland," *Geographia Polonica*, vol. 82, pp. 61–67, 2009.

[18] D. A. Hennessy, "Crop yield skewness and the normal distribution," *Journal of Agricultural and Resource Economics*, vol. 1, pp. 34–52, 2009.

[19] K. M. Lee, P. R. Armstrong, J. A. Thomasson, R. Sui, M. Casada, and T. J. Herrman, "Application of Binomial and Multinomial probability statistics to the sampling design process of a global grain tracing and recall system," *Food Control*, vol. 22, no. 7, pp. 1085–1094, 2011.

[20] W. Xiang, L. Yunxian, Q. Zhenwei, and S. Zeliang, "Estimation of crop yield distribution: implication for crop engineering risk," *Systems Engineering Procedia*, vol. 3, pp. 132–138, 2012.

[21] A. Holzkämper, P. Calanca, and J. Fuhrer, "Statistical crop models: predicting the effects of temperature and precipitation changes," *Climate Research*, vol. 51, no. 1, pp. 11–21, 2012.

[22] M. J. Yusuf, G. Nabi, A. Basit, S. K. Husnain, and L. H. Akhtar, "Development of high yielding Millet variety "Sargodha Bajra-2011" released for general cultivation in Punjab Province of Pakistan," *Pakistan Journal of Agricultural Sciences*, vol. 3, pp. 275–282, 2012.

[23] M. Hassan, A. Ahmad, S. Zamir et al., "Growth, Yield and Quality Performance of Pearl Millet (<i>Pennisetum americanum</i> L.)," *American Journal of Plant Sciences*, vol. 5, no. 15, pp. 2215–2223, 2014.

[24] K. Cristian and G. Camelia, "Use of maximum entropy in estimating production risks in crop farms," in *Proceedings of the 6th Edition of the International Symposium*, The Research Institute for Agricultural Economy and Rural Development (ICEADR), Bucharest, Romania, November 2015.

[25] M. M. Rahman and A. J Robson, "A novel approach for sugarcane yield prediction using landsat time series imagery: a case study on bundaberg region," *Advances in Remote Sensing*, vol. 5, no. 2, pp. 93–102, 2016.

[26] U. Asmat, A. Ahmad, T. Khaliq, and J. Akhtar, "Recognizing production options for Pearl Millet in Pakistan under changing climate scenarios," *Journal of Integrative Agriculture*, vol. 4, pp. 762–773, 2017.

[27] Z. I. Muhammad, M. Riaz, and W. Nasir, "Multivariate outlier detection: a comparison among two clustering techniques," *Pakistan Journal of Agricultural Sciences*, vol. 1, pp. 227–231, 2017.

[28] A. Ullah, N. Salehnia, S. Kolsoumi, A. Ahmad, and T. Khaliq, "Prediction of effective climate change indicators using statistical downscaling approach and impact assessment on pearl millet (Pennisetum glaucum L.) yield through Genetic Algorithm in Punjab, Pakistan," *Ecological Indicators*, vol. 90, pp. 569–576, 2018.

[29] O. P. Yadav, S. K. Gupta, M. Govindaraj et al., "Genetic gains in pearl millet in India: insights into historic breeding strategies and future perspective," *Frontiers of Plant Science*, vol. 12, Article ID 645038, 2021.

[30] T. I. Masenya, V. Mlambo, and C. M. Mnisi, "Complete replacement of maize grain with sorghum and pearl millet grains in Jumbo quail diets: feed intake, physiological parameters, and meat quality traits," *PLoS One*, vol. 16, Article ID e0249371, 2021.

[31] S. Gul, J. Ren, N. Xiong, and M. A. Khan, "Design and analysis of statistical probability distribution and nonparametric trend analysis for reference evapotranspiration," *PeerJ*, vol. 9, Article ID e11597, 2021.

[32] A. M. C. Heureux, J. Alvar-Beltrán, R. Manzanas et al., "Climate trends and extremes in the Indus river basin, Pakistan: implications for agricultural production," *Atmosphere*, vol. 13, no. 3, p. 378, 2022.

[33] V. Vadez, "Root hydraulics: the forgotten side of roots in drought adaptation," *Field Crops Research*, vol. 165, pp. 15–24, 2014.

[34] R. K. Uppal, S. P. Wani, K. K. Garg, and G. Alagarswamy, "Balanced nutrition increases yield of pearl millet under drought," *Field Crops Research*, vol. 177, pp. 86–97, 2015.

Hindawi
Complexity
Volume 2023, Article ID 4140594, 13 pages
https://doi.org/10.1155/2023/4140594

Research Article

Resource Allocation in Multicore Elastic Optical Networks: A Deep Reinforcement Learning Approach

Juan Pinto-Ríos,[1] **Felipe Calderón**,[1] **Ariel Leiva**,[1] **Gabriel Hermosilla**,[1] **Alejandra Beghelli**,[2] **Danilo Bórquez-Paredes**,[3] **Astrid Lozada**,[4] **Nicolás Jara**,[4] **Ricardo Olivares**,[4] **and Gabriel Saavedra**[5]

[1]*School of Electrical Engineering, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2950, Valparaíso 2362804, Chile*
[2]*Optical Networks Group, Department of Electronic and Electrical Engineering, University College London, WC1E 7JE, London, UK*
[3]*Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Av. Padre Hurtado 750, Viña Del Mar 2562, 340, Chile*
[4]*Department of Electronic Engineering, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso 2390123, Chile*
[5]*Electrical Engineering Department, Universidad de Concepción, Víctor Lamas 1290, Concepción 4070409, Chile*

Correspondence should be addressed to Juan Pinto-Ríos; juan.pinto.r@pucv.cl

Received 17 June 2022; Revised 16 December 2022; Accepted 27 January 2023; Published 2 March 2023

Academic Editor: Ning Cai

A deep reinforcement learning (DRL) approach is applied, for the first time, to solve the routing, modulation, spectrum, and core allocation (RMSCA) problem in dynamic multicore fiber elastic optical networks (MCF-EONs). To do so, a new environment was designed and implemented to emulate the operation of MCF-EONs - taking into account the modulation format-dependent reach and intercore crosstalk (XT) - and four DRL agents were trained to solve the RMSCA problem. The blocking performance of the trained agents was compared through simulation to 3 baselines RMSCA heuristics. Results obtained for the NSFNet and COST239 network topologies under different traffic loads show that the best-performing agent achieves, on average, up to a four-times decrease in blocking probability with respect to the best-performing baseline heuristic method.

1. Introduction

Due to the ever-growing number of users, devices, and networking applications, Internet traffic keeps on increasing, more than doubling every two years, to levels that will lead to an eventual capacity crunch of the current core optical networks [1, 2]. Big technological companies, such as Google, Meta, Amazon, Netflix, Apple, and Microsoft now account for more than half of Internet traffic, and the introduction of 5G is expected to accelerate the growth of emerging heavy app users consuming 1 terabyte per month [3].

Various solutions to deal with this constant traffic growth have been proposed, ranging from greater efficiency in using currently deployed optical resources to expanding the capacity of the optical transport network. Examples of the former and latter are Elastic Optical Networks (EONs) [4] and multicore optical fiber (MCF) [5], respectively. EONs [6] divide the spectrum into narrow slots called frequency slot units (FSU), usually of 12.5 GHz width [7]. In EON communication, each connection uses as many adjacent slots as needed, thereby improving the spectral usage efficiency [8]. Under dynamic operation, EONs [9] can establish and release connections on-demand. MCF extends the fiber capacity by adding multiple cores within the same cladding. Thus, the capacity of a single fiber is significantly increased given that each core can be considered as an extra optical medium [10].

One of the first cases for the support of elastic optical networks was given by data-intensive applications running on multidata center systems [11]. Later, the need for elastic optical networks was highlighted for applications such as

cloud-based IoT services [12], cloud-fog computing [13] as well as critical support for the 5G communication infrastructure [14] and the applications associated with it, such as Ultra High Definition videos, Telemedicine, and Smart City/Industry/Factory/Home [15]. Similarly, MCF has been identified as a complement of elastic optical networks to deliver the high capacity required by current and future applications, as well as the driving force to provide cost-efficient solutions for high-capacity submarine cables [16, 17], a key infrastructure underpinning Internet. Currently, applications requiring the combination of MCF [18] with the efficient use of the spectrum offered by dynamic elastic optical networks [19] have also been identified in scenarios such as intradata center networks [20, 21]. Going beyond the current technological state, expected Multimedia 3D Services for 6G networks such as Tactile/Haptic Internet, Video Games/Streaming as a 3D Service, and Deep-Sea Sightseeing [15] will certainly require a network capacity that only will be provided by the combination of MCF and dynamic EONs, termed as dynamic MCF-EON from now on.

One of the main challenges of dynamic MCF-EONs is the design of efficient routing, modulation, spectrum, and core assignment (RMSCA) strategies for establishing optical connections with as low blocking probability as possible. Most RMSCA proposals use heuristic approaches that consider the impact of intercore crosstalk (intercore XT) on optical signal quality, as described in [19, 22–25]. Although rule-based heuristics are computationally simple, their performance depends on the ability of the designer to detect the best set of rules defining the heuristic behavior [26]. In recent years, it has been shown that in most cases, deep reinforcement learning (DRL) techniques applied to solve resource allocation problems in dynamic elastic optical networks outperform rule-based systems [27, 28]. DRL has the ability to explore solutions other than those detected by the expert knowledge of the human designer. As a result, it has the potential of generating new nonobvious policies from the experience gained after training in a relevant environment [29].

1.1. Related Work. In dynamic scenarios, DRL was applied to solve the routing, modulation, and spectrum assignment (RMSA) problem in single-domain EONs [27, 28, 30, 31], multidomain EONs [32], multiband EONs [33, 34] ,and survivable EONs operating under shared protection [35]; the problem of energy-efficient traffic grooming in fog-cloud EONs [36], the problem of establishing and reconfiguring multicast sessions in EONs [37], the fragmentation mitigation problem [38], and the resource allocation problem with advanced reservation (AR) in EONs for cloud-edge computing [39]. Only one previous work has studied the application of DRL on MCF networks [40], but this work focused on fixed-grid networks. In this paper, we extend the work reported in [27, 30, 31] by applying DRL to dynamic MCF-EONs for the first time.

In the context of dynamic MCF or MCF-EON networks, with the exception of [40], only supervised machine-learning techniques have been applied so far. These consist of techniques for making inferences based on expert-labeled data. Thus, instead of taking actions, supervised learning algorithms perform estimations or classifications [41]. For example, the authors of [42, 43] used supervised learning to predict future connection requests in dynamic MCF-EONs to perform a crosstalk-aware resource allocation in advance. Instead, the authors in [44] used machine learning to estimate the intercore XT to then execute a crosstalk-aware allocation algorithm. All these studies have used machine learning as an auxiliary process to improve the heuristic allocation, either by predicting future traffic or transmission quality. In none of them, machine learning had direct participation in the decision-making related to resource allocation.

1.2. Paper Contribution. To the best of our knowledge, there are no previous studies on applying DRL to solve the RMSCA problem in dynamic MCF-EONs. In this paper, we present, for the first time, the implementation and testing of a new dynamic MCF-EON environment where four different DRL agents are trained to solve the RMSCA problem. The results obtained by the best-performing agent are then compared to 3 baseline heuristics.

The rest of this article is organized as follows: Section 2 presents the DRL system developed, Section 3 describes the performance evaluation experiments, and Section 4 concludes the paper.

2. DRL for Dynamic MCF-EONs

A DRL system can be summarized as an agent (an entity equipped with a learning algorithm) that—during its training phase—learns to make good decisions by interacting with an environment [45, 46].

In the context of RMSCA, the agent must learn to allocate optical resources to connection requests such that they are not blocked. Blocking can happen due to physical impairments or lack of spectral continuity or contiguity in the chosen route. A good allocation decision makes the environment give the agent a high-value reward.

Formally, a DRL system can be modeled as a Markov Decision Process (MDP) described by the 6-tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, s_0, \gamma\}$ [29], where the following takes place:

(i) $\mathcal{S}$ (States): Set of possible states describing the status of the system. In this work, the state s_t is described by the link spectrum utilization, at time step t, of each candidate route per core between the source and destination node of connection request cr_t. The latter is defined by the tuple $\{o, d, h, b\}$, where o is the source node, d is the destination node, h is the holding time of the request, and b is the bitrate of the demand.

(ii) $\mathcal{A}$ (Actions): Set of actions the agent can take. In this work, an action a_t at time step t is a triplet (k, c, j), where k is the selected route (out of K pre-computed routes) c the identifier of the core (out of C cores), and j the identifier of a block of contiguous slots that

can accommodate the demand of cr_t (out of J blocks).

(iii) $\mathcal{T}(s_{t+1}|s_t, a_t)$ (Transition probability): Probability distribution that the system transits to state s_{t+1}, given the system is in state s_t and the agent takes action a_t when receiving connection request cr_t.

(iv) $\mathcal{R}(s_t, a_t, s_{t+1})$ (Reward): The reward function that defines the immediate reward (r_t) received when transiting to state s_{t+1} due to action a_t while in state s_t. In this work, the reward is designed to be equal to 1 if the request is accepted and −1 if it is rejected.

(v) s_0 (Initial state): The state of the network at the start of the decision process. In this work, this state corresponds to all routes having all spectrum slots available in all links and cores.

(vi) γ (Discount factor): A parameter $\in [0, 1)$ that sets the importance of current and future rewards. This factor adjusts the process of exploration and exploitation of agents in the environment [29].

The evolution of the DRL system defined above is as follows: During a training episode—made of a finite amount of time steps—an agent learns to make good decisions by interacting with the environment at each time step t [45, 46]. To do so, upon receiving a connection request cr_t with the system in state s_t, the agent generates an action a_t. Such action makes the environment transit to the state s_{t+1} with probability $\mathcal{T}(s_{t+1}|s_t, a_t)$ and the agent receives a reward $R_t(s_t, a_t, s_{t+1})$. The objective of the agent is maximizing the expected future discounted reward. Thus, by repeating this process during the training episode, the agent will learn a policy $\pi^*(a|s)$ that leads to maximizing the return function, Γ_t, defined as

$$\Gamma_t = \sum_{t' \in [t,\infty)} \gamma^{t'-t} \cdot R_{t'}. \tag{1}$$

The details of the state modeling are as follows: We extend the state defined in [27] by considering the different cores. Thus, the state is represented as an array of $1 \times (2|V| + 1 + (2J + 3) \cdot K \cdot C)$ elements, where $|V|$ is the number of nodes of the optical network. The extended state is then given by

$$s_t = \left\{ o, d, h, b, \left\{ \left\{ z_{k,c}^{1,j}, z_{k,c}^{2,j} \right\}\Big|_{j \in \{1,\dots,J\}} \right\}, \right. \\ \left. z_{k,c}^{3}, z_{k,c}^{4}, z_{k,c}^{5} \right\}_{k \in \{1,\dots,K\}}\Big|_{c \in \{1,\dots,C\}}, \tag{2}$$

where a one-hot encoding is used to identify the origin and destination nodes. Each subcomponent z in the vector mentioned above is as follows: For each core $c \in C$ and route $k \in K$ between o and d nodes, $z_{k,c}^{1,j}$ is the size of $j-th$ block that can accommodate the connection request. $z_{k,c}^{2,j}$ is the index of the first slot of each block j. The third component, $z_{k,c}^{3}$, is the number of FSUs required to establish the connection (given the modulation format used). Finally, $z_{k,c}^{4}$, the average number of FSUs available in all J blocks in route k and core c is included, and $z_{k,c}^{5}$ is the total number of FSUs available in the route k in core c.

In our resource allocation problem, the environment is programmed to represent the operation and constraints of a dynamic MCF-EON. When a connection request arrives during the training phase, the agent decides what resources to allocate. At the beginning of its training, the agent makes random decisions (exploration process). Then, the environment determines whether the set of resources identified by the agent is feasible and gives the agent feedback about the quality of its decision. This information, stored in the experience buffer of the agent, allows the agent to learn. As a result, it starts to select better actions (exploitation process) for future requests. Better actions result in the agent earning a high cumulative reward. After an agent has finished its training stage, it can be evaluated (testing stage) by having it to process a new set of connection requests.

The implementation of any DRL system is done in two stages as follows:

(i) *Stage 1: Environment Design and Implementation*. The environment is a program that receives the agent's action, processes it, and sends back feedback. The specific feedback depends on the results of the agent's action on the environment. The environment must consider the characteristics and constraints of the existing system to process the action. In the case of an optical network, the environment must manage information about the network topology and status and model the network operation (including physical phenomena related to the signal transmission and spectrum allocation constraints).

(ii) *Stage 2: Agent Training*. The agent must first acquire knowledge about the environment. This training is done by exploration and exploitation. When exploring, the agent selects random actions to learn how the environment reacts and stores such knowledge. When exploiting stored knowledge, the agent makes informed decisions to select the following action: During exploration and exploitation, the agent receives feedback from the environment, which the agent uses to update its knowledge (policy). In this way, the agent's training progresses.

In the following section, these two stages are described in detail in the context of dynamic MCF-EONs.

2.1. Stage 1: Environment Design and Implementation. In this work, the toolkit *Optical RL-Gym*, developed by Natalino and Monti [31] to facilitate the implementation and replicability of deep reinforcement learning environments for optical networks was extended by creating a new environment: DeepRMSCAEnv. Such an environment encapsulates all the necessary functions to simulate an MCF-EON.

The right part of Figure 1 shows a schematic of the implemented environment, including its main components and interactions. Dashed and thick lines modules are modules from the Optical RL-Gym toolkit that had to be modified and developed from scratch, respectively, to model an MCF-EON environment correctly.

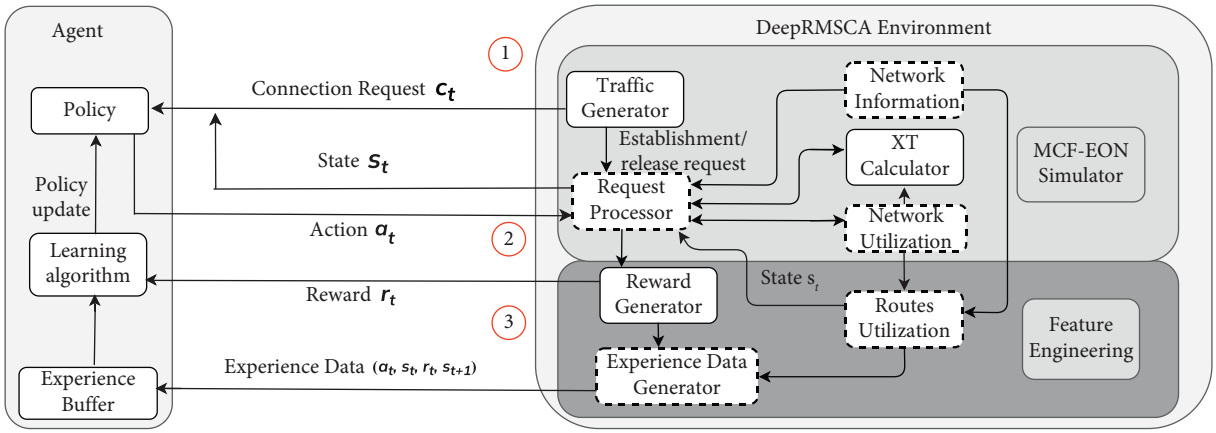


Figure 1: Interaction between a DRL agent and the MCF-EON environment developed: DeepRMSCAEnv.

The environment can be considered made of an event-driven dynamic MCF-EON simulator and a feature engineering module. The former is responsible for processing the connection requests according to the agent's action and sending the relevant information to the feature engineering module. The latter is responsible for preparing and sending feedback to the agent (reward and observation).

The dynamic MCF-EON simulator consists of five components. Two of these store data about the network as follows:

(i) *Network Information.* This component stores the graph representation of the network and the link capacity, considering the multicore nature of links. It also stores the K alternatives routes for each source-destination pair, the modulation format used as a function of the route length distance, and the network information, coded as in [27].

(ii) *Network Utilization.* This component stores the utilization of each slot (available or used) for each network link and core.

The remaining 3 components perform specific tasks

(i) *Traffic Generator.* This component is responsible for the random generation of connection establishment and release requests. At time step t, connection request cr_t is sent to the agent and the request processor component. Connection release requests are sent only to the request processor.

(ii) Request Processor. This component receives several inputs. The first two are the connection establishment or release request and the action of the agent (in the case of a connection request). When connection requests cr_t are received at time step t, the Request Processor module receives s_t from the Route Utilization module and sends it to the agent, then the Request Processor module waits for the action of the agent. Once the action, a_t, is received, the Request Processor first determines the number of slots the connection requires. To do so, the most efficient modulation format that ensures a QoT [47] is first selected (QoT has been transformed into a maximum reach, as shown in Table 1). The calculation of the number of slots is the same described in Section 2 of [27]. Next, it checks the network topology (input from the network information module) and the network utilization (input from the network utilization module) to evaluate the availability of the resources selected by the agent. It also obtains information from the XT calculator component regarding the feasibility of the allocation in terms of crosstalk. If resources are available and a positive answer is received from the XT calculator, then resources are allocated, and the corresponding information is updated on the network utilization module. Information about a successful establishment is also sent to the Reward Generator module. If resources cannot be allocated, information about the failed establishment is sent to the Reward Generator component only. When a connection release is received, the Request Processor component updates the network utilization module to make the released resources available.

(iii) *XT Calculator.* This component calculates the intercore crosstalk (XT), defined as the interference between optical connections in neighboring cores using the same frequency slots. It receives information about the resources selected by the agent's action a_t (length of the links composing the route and core) and the route-level utilization information from the network utilization module and evaluates the XT. For generic MCF systems, with any number of cores in any geometric arrangement, the steps to calculate the mean XT affecting a connection established in core x are as follows:

(a) Calculate the mean XT per unit of length between core x and adjacent core y, $w_{x,y}$ as

$$w_{x,y} = \frac{2g^2 q}{\beta \Lambda_{x,y}}, \quad (3)$$

where g, q, β, and Λ are the coupling coefficient, radius of curvature (or bending), constant propagation, and the distance between cores x and y, respectively.

(b) Calculate the total mean XT affecting core x, XT_x, by adding the crosstalk contribution of all its adjacent cores. That is,

$$\mathrm{XT}_x = \sum_{y=1}^{n} w_{x,y} \cdot L, \quad (4)$$

where n is the number of cores adjacent to core x and L the length of the link.

For the specific case where cores follow a triangular or hexagonal geometric arrangement and different pairs of cores are equidistant, equation (5) has been found to be a better approximation to calculate XT_x [19], as given as follows:

$$\mathrm{XT}_x = \frac{n - n \cdot \exp\left[-(n+1) \cdot wL\right]}{1 + n \cdot \exp\left[-(n+1) \cdot wL\right]}, \quad (5)$$

where, as in equation (4), n represents the number of cores neighbouring x, and L is the length of the link. The term w is given by equation (3) (subindices have been dropped since the distance between all core pairs is assumed to be the same).

An XT threshold value for different modulation formats is defined in [48, 49] such that the signal quality is acceptable. If XT exceeds this predefined threshold (summarized in Tables 1 and 2), a negative answer is sent to the request processor (−1). Otherwise, a positive answer is sent (1).

The feature engineering module in Figure 1 is responsible for preparing the information to be sent back to the agent. It is made of three components as follows:

(i) *Reward Generator.* This component calculates the numerical reward to be sent to the agent depending on the information received from the Request Processor component. In this work, a successful resource allocation returns a reward equal to 1 and a failed allocation equal to −1. Connections can be rejected due to a lack of spectrum resources along the route selected by the agent, because of crosstalk among cores exceeding the predefined threshold, or because the length of the route selected by the agent is longer than the maximum optical reach of any modulation format (such limit depends on the modulation format and a bit-error-rate threshold, as in Table 1 of [50].

(ii) *Routes Utilization.* This component receives the routing information from the network information component and the utilization state of the slots in the K shortest routes between the origin and destination nodes of connection request cr_t from the network utilization module. This information is then consolidated in a 1D vector made of $(K \cdot C \cdot J)$ elements, where K is the number of alternative routes, C is the number of cores, and J is the number of blocks with enough available slots to establish the connection request being processed.

(iii) *Experience Data Generator.* This component builds the information to be stored in the Experience Buffer which is a collection of tuples (a_t, s_t, r_t, s_{t+1}) generated during the training process.

Table 1: Maximum reach for each modulation format [50].

Modulation format	Max. reach (km)
64QAM	250
32QAM	500
16QAM	1000
8QAM	2000
QPSK	4000
BPSK	8000

Table 2: XT threshold for each modulation format [49].

Modulation format	XT threshold (dB)
64QAM	−34
32QAM	−27
16QAM	−25
8QAM	−21
QPSK	−18
BPSK	−14

2.2. Stage 2: Agent Training. The left side of Figure 1 shows the interaction between the agent and the DeepRMSCAEnv environment during the training stage.

The agent aims to maximize its long-term reward. That is, selecting actions leads to the highest number of connection requests established. To achieve this goal, the agent is built considering two main components.

2.2.1. Policy. This component is where the knowledge of the behavior of the agent is embedded. At a given time, t receives a connection establishment request, cr_t as input along with state s_t, and action a_t is outputted. The action is defined by 3 integer numbers, namely, a route identifier k (selected out of K possible precomputed routes), a core identifier c (selected out of C possible cores), and the identifier j of the block selected. These values define which route, core, and spectrum resources should be assigned to each request. As the agent successfully allocates more connection requests, the policy becomes better. At the end of the training, the policy is expected to allow the agent to define which action has the highest probability of not being blocked. Figure 2 shows a simplified example of two possible actions that might be taken by the agent, given a specific s_t.

On the left part of the figure, a 5-node network topology and a connection establishment request of 2 slots between nodes 5 and 3 are shown. The demand is represented by the

Table 3: Network, traffic and training parameters.

Parameters	Value
Network parameters	
Topologies	NSFNet [50] and COST239 [51]
Number of cores	3
Number of FSU by link	100
Modulation formats	BPSK, QPSK, 8-QAM, 16-QAM
Traffic parameters	
Bit rates (Gb/s)	Uniformly distributed in [25–100] Gbps
Agent training parameters	
Precomputed candidate routes	5
Number of connection requests per episode	50 [27]
Simulated requests per training	160,000
Agent's learning algorithm parameters	By-default [52]
Agent's hyperparameters	By-default [31]

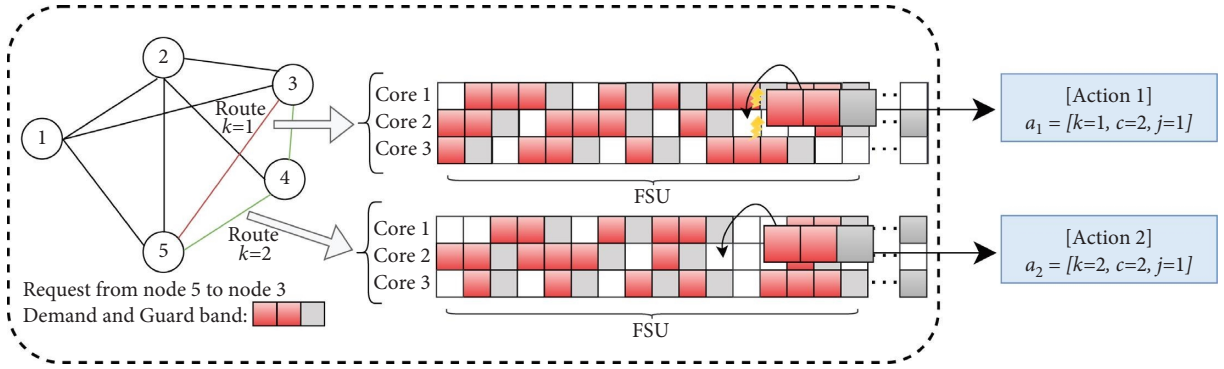


Figure 2: Example of a connection request from node 5 to 3, requesting 3 slots (2 for data, 1 as guard band). $K = 2$, meaning 2 routes. The routes spectral use is represented by white blocks (available FSUs) and red and grey ones (occupied for data and as guard bands, respectively).

red boxes (2 slots in this case) plus the grey box (1 slot used as a guard band). The number of slots required to serve the connection (red squares) is determined by the modulation format, using the same method presented in [27]. One guard band of 1 slot is considered for each connection request to achieve a good trade-off between the quality of transmission and the blocking probability [51].

Let us assume that the network is equipped with three cores per link, and the agent can select either route 1 ($k = 1$), represented by the red link in the topology, or route 2 ($k = 2$) by the green links. In addition to a route, the agent must also select a core and a slot. On the right side of the figure, the spectrum utilization of both routes is shown. Red and grey squares represent used FSUs. A row of squares represents the slot utilization in a specific core for a specific route. Thus, the three rows on the upper and lower part of the figure represent the slot utilization on the three cores of the first and second routes, respectively.

If the agent selects Action 1, depicted in the upper part of the figure, then action $a_t = [1, 2, 1]$ is sent back to the environment, signaling that the agent selects slot 11 as the initial slot on route 1 in core 2 to establish the connection. The thunderbolt symbol in route 1 represents the presence of crosstalk exceeding the acceptable threshold. In this case, the request will be rejected, and a reward of −1 will be sent to the agent. Instead, if the agent selects Action 2, depicted on the lower part of the figure, then action $a_t = [2, 2, 1]$ is sent back to the environment. This action leads to a successful connection establishment, and the agent receives a reward equal to 1. During the training stage, the policy component should be updated to select Action 2 over Action 1 (for this state s_t), leading to a higher reward.

2.2.2. Learning Algorithm. This component receives the Experience Data from the environment and, based on that information, updates the policy to produce actions that maximize the expected cumulative long-term reward. In this study, we consider learning algorithms compatible with the action space. The action space used has a multidiscrete nature because the action is defined by multiple discrete values (route, core, and slot identifier). Thus, the learning algorithms available in the Stable-Baselines [52] library that was compatible with a multidiscrete space state were selected (as also done in [27, 31]). These are as follows:

(i) *Advantage Actor-Critic (A2C)* [53] *and Actor-Critic using Kronecker-Factored Trust Region (ACKTR)* [54]. These are approaches based on the actor-critic algorithm [53], which has two interacting neural networks. The actor uses a dense neural network to

process and update the policy obtained. The critic uses a separated neural network to evaluate the quality of the policy by calculating the "value function" [45]. Both algorithms differ in how they update their neural networks' weights. A2C does that by using the feedback the critic's network gives to the actor's network, whilst ACKTR uses a Kronecker-factored approximation [56], which is a method that optimizes the stochastic gradient descent.

(ii) *Proximal Policy Optimization (PPO2)* [55] *and Trust Region Policy Optimization (TRPO)* [56]. These learning algorithms use only one neural network, whose weights are updated based on the policy gradient descent. They differ in the way the policy gradient descent is approached. TRPO avoids sudden changes in the neural network weights, updating only those that do not differ by a greater distance than what the Kullback–Leibler restriction (relative entropy) [56] allows. Instead, PPO2 does not impose limits on the neural network weights' changes to optimize the policy's descent curve.

3. Performance Evaluation

Table 3 lists the values of the main parameters used to train the agents. In terms of network parameters, we consider two topologies, namely, the NSFNet (mills [57]) and the COST239 (Batchelor [58]). For each one, we assume 100 FSUs and 3 cores arranged in a triangular geometry per link, and the available modulation formats are BPSK, QPSK, 8-QAM, and 16-QAM. These simplifications have been considered due to memory constraints. The same number of slots was considered in [33]. As in [59], we use (5) to calculate the XT.

Regarding the traffic characteristics, we assume a fully dynamic behavior, where connection establishment requests arrive as a Poisson process and connection holding times follow a negative exponential distribution. The bitrate associated to each connection is uniformly selected from the range [25–100] Gbps, as in [27]. Finally, regarding the agents (one per learning algorithm), they will select one out of 5 precomputed routes, one out of 3 cores, and the identifier j of the FSU block for the connection considering a total of 100 FSUs. Agents will be trained in episodes made of 50 connection requests each (to simplify backpropagation in the dense neural network used by the agent by delivering small batches of data continuously), and the whole training session will consider a total of 160,000 connection requests. The parameters of the four agents will be the ones set by default in the agent's library *Stable Baselines* [52]. The DRL system developed is available in a Git repository (The new environment, under the name DeepRMSCAEnv, is available at https://gitlab.com/IRO-Team/deeprmsca-a-mcf-eon-enviroment-for-optical-rl-gym/).

3.1. Preliminary Training Results. Results for the training process of 4 agents are presented. The following discussion about the results obtained is valid only for the hyper-parameters used for each agent defined in Table 3.

The agents TRPO, PPO2, A2C, and ACKTR were trained with a traffic load of 250 Erlang, as in [27].

Figures 3 and 4 show the reward accumulated by the different agents during their training in the NSFNet and COST239 topologies, respectively. Given that each episode is made of 50 connection requests, the maximum reward achievable by an agent is 50. It can be seen that the A2C and TRPO agents are the only ones reaching values close to the maximum expected reward in both topologies with an average reward of 49 and 47, respectively, with TRPO exhibiting slightly better performance. On the other hand, the PPO2 and ACKTR agents did not perform well using the default parameters. PPO2 performed well on the COST239 topology (reward oscillated around 42) but not on NSFNet (reward oscillated around 30). In both topologies, the agent got stuck to the same value of reward from the very beginning, showing no signs of learning. In the case of ACKTR, the default parameters were not suitable for this task either. Not only the agent got low values of reward in both topologies, but in the COST239 topology, the reward obtained decreased during several periods of the training process, never again exceeding the value obtained in the first 10,000 timesteps.

The A2C (Actor-Critic) learning algorithm filters those agents' actions leading to a low reward. Such filtering is possible thanks not only to the feedback received from the environment but also to the feedback given to the Actor (neural network in charge of applying the policy) by the Critic (neural network in charge of evaluating the quality of the policy used through the Value function). As a result, the agent starts with low values of reward (exploration phase) to then quickly increasing its reward per episode (exploitation phase) as the training progresses. Such behavior can be observed in Figures 3 and 4, where the A2C agent requires a few episodes to achieve a reward close to 50 and exhibits one of the best results in both topologies.

ACKTR (Actor-Critic using Kronecker-Factored Trust Region) is a trust-region optimization algorithm for actor-critic methods with gradient update sped up by means of the Kronecker-factored approximation. The effectiveness of the trust-region method is highly dependent on the learning algorithm's parameters. In practice, using the by-default parameters of the Stable Baselines led to the following: (a) the weights of the actor's neural network not being updated, trapping the agent in a local optimum, as seen in Figure 3 (NFSNet topology) and (b) not finding the trust region, resulting in random actions, as seen in Figure 4 (COST 239 topology).

PPO2 uses a different approach by updating the gradient more frequently than other methods. As a result, it can find a good policy more quickly than other methods, as shown in Figures 3 and 4. However, Figure 3 shows that it also gets

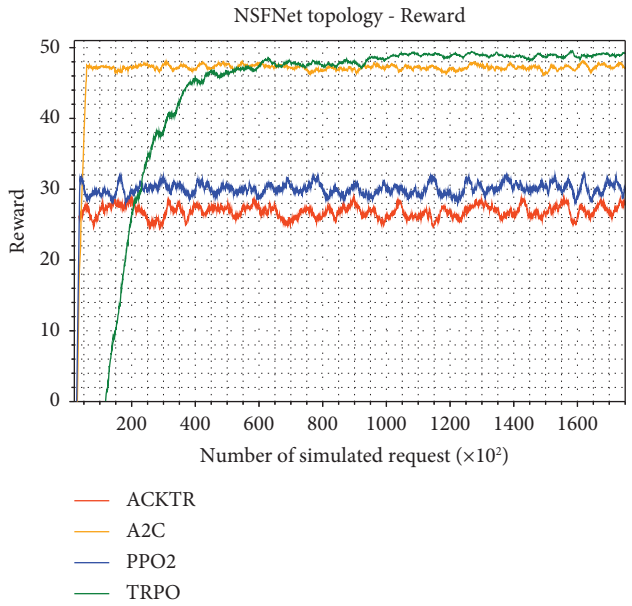


Figure 3: Accumulated reward for the A2C, PPO2, TRPO, and ACKTR agents in the NSFNet topology.

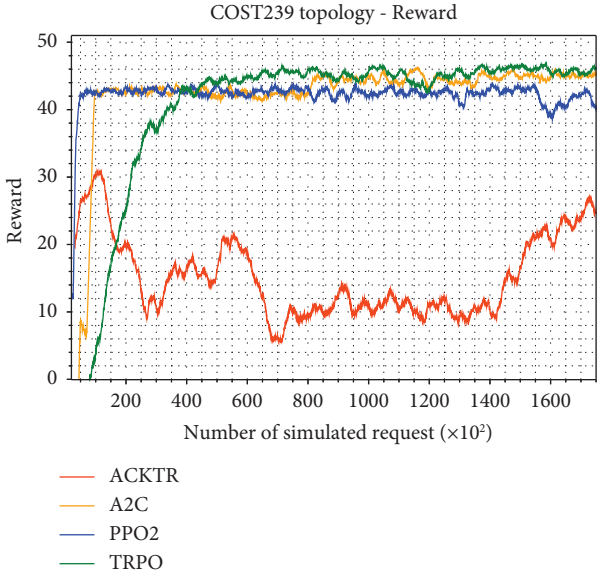


Figure 4: Accumulated reward for the A2C, PPO2, TRPO, and ACKTR agents in the COST239 topology.

stuck in a local optimum. Most probably this is due to the use of the by-default learning algorithm's parameters of Stable Baselines.

Finally, TRPO combines the policy gradient method of PPO2, but it also uses a trust region to avoid radical changes in the update of the neural network weights. The size of the trust region is aimed to avoid increasing the relative entropy of information based on the factor Kullback–Lieber As a result, it improves slowly and monotonically, as seen in Figures 3 and 4. For the problem studied here, this agent achieved the highest cumulative reward.

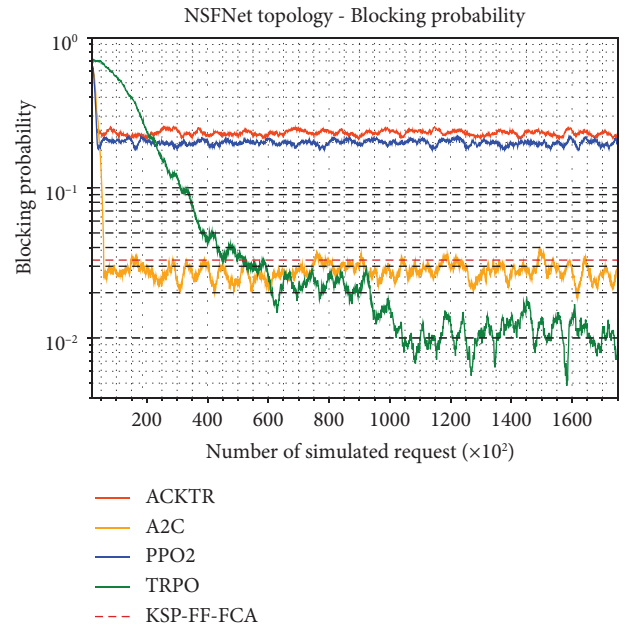


Figure 5: Blocking probability for A2C, PPO2, TRPO, and ACKTR in NSFNet Topology.

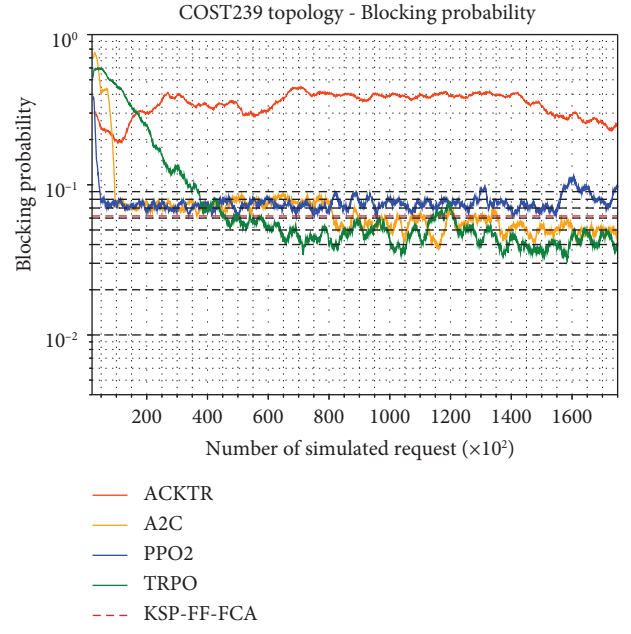


Figure 6: Blocking probability for A2C, PPO2, TRPO, and ACKTR in COST239 Topology.

Please notice that parameter tuning is out of the scope of this work, as our aim was to show the potential of DRL as a solution for the dynamic resource allocation in MCF-EONs.

Figures 5 and 6 show the evolution of the blocking probability during the training process of the same agents for the NFSNet and COST239 topologies, respectively. For comparison, the dashed red line shows the blocking probability obtained by one of the baseline heuristics, kSP-FF-FCA. This heuristic has a list of 5 precomputed routes, sorted

from shortest ($k = 1$) to longest ($k = 5$). When a connection request arrives, the heuristic attempts to establish the connection in the shortest path of the list ($k = 1$), applying the first-fit policy for spectrum allocation and first-fit crosstalk-aware for core allocation, as described in [60]. The same procedure is repeated for the following route in the list if unsuccessful: After attempting all paths, the connection is rejected if there are no available resources.

From the figure, we can see that once the agents are in steady-state, TRPO and A2C agents outperform the heuristic, improving blocking of 24.3% and 73.9% for the NSFNet topology and 14.51% and 38.71% for the COST239 topology, respectively.

Given the excellent performance of the TRPO agent in both topologies, in the following section, this agent will be trained for different traffic loads, and then its performance will be contrasted with that of the heuristics selected in [61].

3.2. TRPO Training Results. The TRPO agent was trained for traffic loads between 500 and 3000 Erlang, in steps of 500. Figures 7 and 8 show the evolution of the blocking probability achieved by the TRPO agent as the training process progresses for different traffic loads for the NSFNet and COST239 topologies, respectively. It can be seen that the agent exhibits consistent behavior, with the blocking probability increasing with the traffic load, as expected. It can also be seen that at the beginning of the training process, the agent obtains a high blocking probability due to the exploration process. When the exploitation process starts, the blocking probability is reduced until it converges to a steady value. This happens after 150 thousand timesteps for the NFSNet and 130 thousand timesteps for COST239, irrespective of the traffic load. Given this steady value, we assume training has finished and the trained agent can now be evaluated in a testing setting.

3.3. TRPO Agent VS. Heuristic: Blocking Performance. Figures 9 and 10 show the blocking probability achieved by the trained TRPO agent and the same heuristics, selected for blocking evaluation in the survey [50], namely, KSP-FF-FCA [61], KSP-RF-RCA [61], and KSP-SCMA XT/demand-aware [22]. Results assume operation in the C-band (320 FSU) for the NSFNet and COST239 topologies, respectively. The three heuristics apply alternated routing. KSP-FF-FCA uses the First Fit policy to select core and spectrum, KSP-RF-RCA applies a random policy to select core and spectrum, and KSP-SCMA XT/demand aware allocates different parts of the spectrum and core depending on the bitrate of the connection request. If the connection request's demand is below a bitrate's threshold, a First-Fit allocation policy is applied for spectrum and core assignment as long as the cross-talk levels are not exceeded; otherwise, a Last-Fit policy is applied if the connection request's demand is above the threshold.

Compared to the best-performing heuristic, KSP-SCMA XT/demand-aware, a significant improvement in the blocking performance of the DRL approaches is observed. For example, in the NFSNet topology, at the highest load

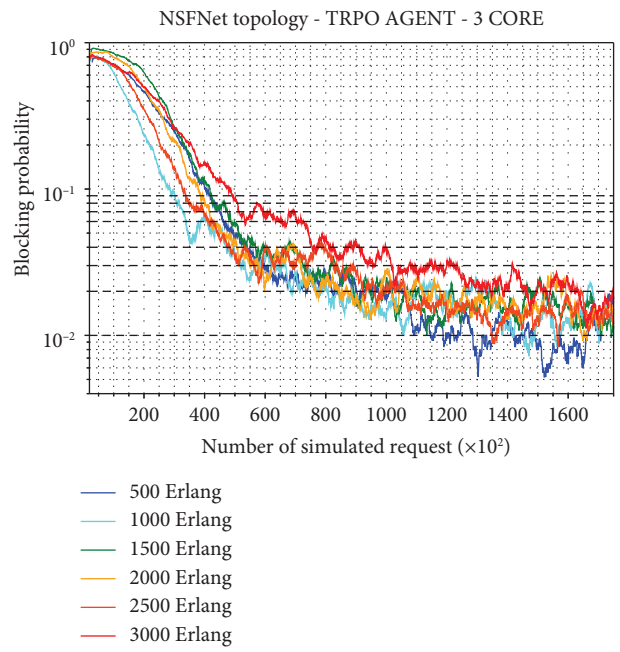


Figure 7: Blocking probability progress for TRPO agent training in NSFNet Topology.

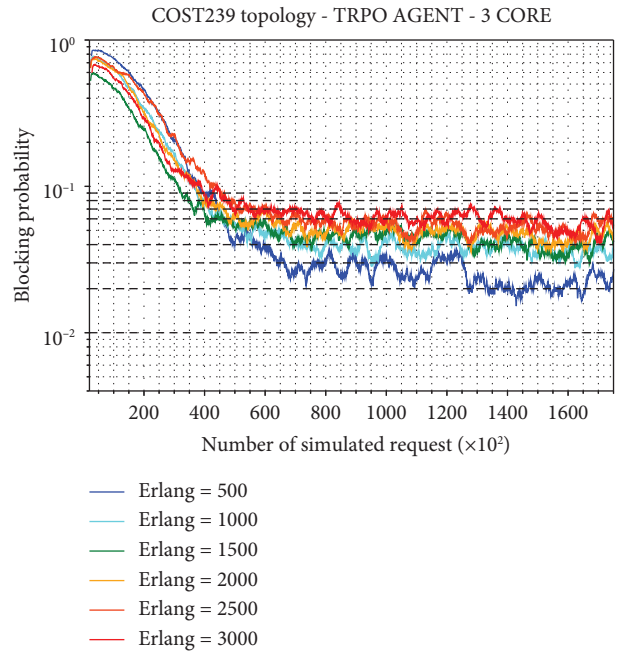


Figure 8: Blocking probability progress for TRPO agent training in COST239 Topology.

studied, the TRPO agent exhibits a blocking probability of about $1.9 \cdot 10^{-2}$, about four times slower than the blocking of $8.5 \cdot 10^{-2}$ achieved by the heuristic. On average, considering both topologies and loads over 2000 Erlang, TRPO achieves a 4-times decrease in blocking concerning the best heuristic, being ideal for the future scenario of demand for connection requests [64] and highlighting the benefits of applying DRL techniques to the RMSCA problem.

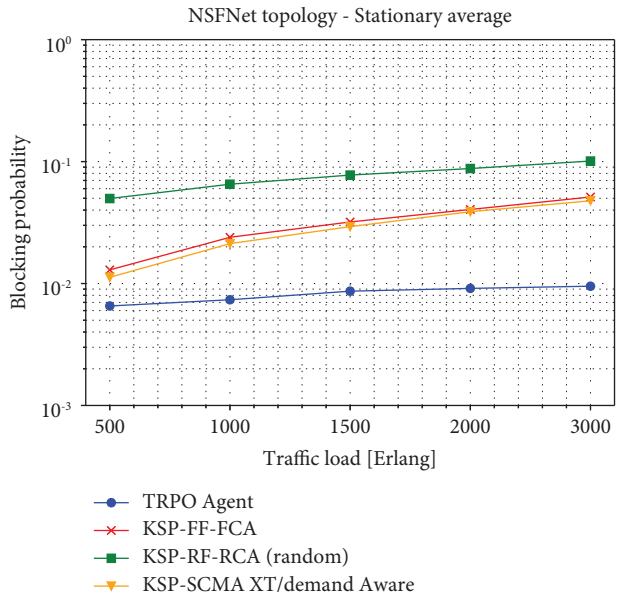


Figure 9: Blocking probability steady average of TRPO agent trained in NSFNet topology.

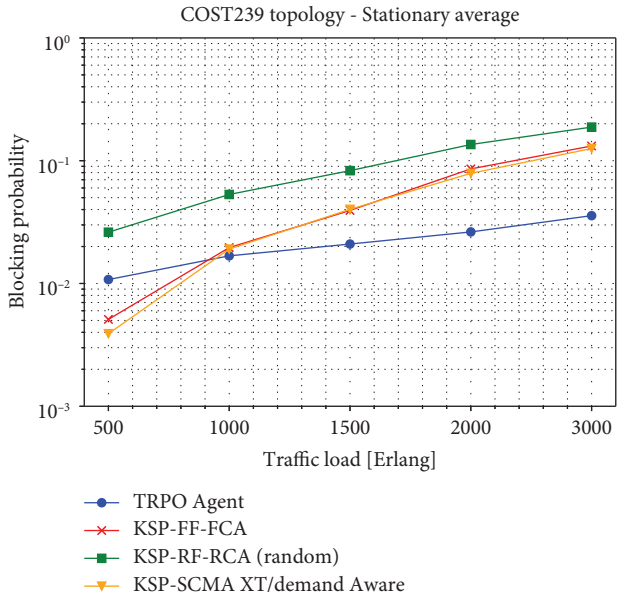


Figure 10: Blocking probability steady average of TRPO agent trained in COST239 topology.

Finally, our results show that the trained agent can generalise policies for different traffic loads and spectrum resources and outperform the rule-based heuristics. The improved performance comes from the ability of the DRL to explore solutions other than those detected by the expert knowledge of the human designer of the heuristics. We have also observed that training in adverse conditions achieves good results. That is, training the agent at high traffic loads makes the agent to perform well at lower traffic loads whereas training the agent using links with reduced capacity leads to the agent to perform better in links with increased capacity. In line with previous research [33], such generalisation was not observed in terms of topology: The agent trained in the NFSNet topology did not perform well in the COST 239 topology and vice versa. Studying the benefit of using Graph Neural Networks to overcome the lack of topology generalisation is part of current research [63].

4. Conclusion

This paper presents a deep reinforcement learning approach applied for the first time in the literature to solve the routing, modulation format, spectrum, and core allocation problem in dynamic multicore elastic optical networks. Simulation results show that the deep reinforcement learning approach offers a significant performance advantage over the best heuristic strategy studied.

Further research on improving the DRL approach performance should focus on hyperparameter tuning, applying transfer learning techniques or graph neural networks to cover a broader range of topologies with decreased computational effort, increasing the size of the data to be processed to study fibers with more cores and investigating different reward schemes that differentiate the reward according to the cause of blocking (e.g. crosstalk, capacity unavailability, fragmentation, or optical reach).

Additionally, we would like to explore explainability techniques that might help understand how the agent makes its decisions to improve current heuristics.

We expect these results and the code made available in the Git repository to help the research community study the benefits of deep reinforcement learning in the area of optical networks.

Data Availability

The data used to support the findings of this study are available at https://gitlab.com/IRO-Team/deeprmsca-a-mcf-eon-enviroment-for-optical-rl-gym.

Disclosure

The preprint is available in Arxiv at https://doi.org/10.48550/arXiv.2207.02074 and in DeepAI blog at: https://deepai.org/publication/resource-allocation-in-multicore-elastic-optical-networks-a-deep-reinforcement-learning-approach.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Financial support from DI-PUCV (039.437/2020, 039.382/2021); ANID FOVI 210082; ANID FONDECYT Iniciación (11201024, 11220650, 11190710); ANID Magister Nacional (/2020-22201418, /2021-22210736); ANID Doctorado Nacional/2022-21220867 is gratefully acknowledged.

References

[1] C. W. Paper, "Cisco visual networking index: global mobile data traffic forecast update," 2018, http://media.mediapost.com/uploads/CiscoForecast.pdf.

[2] TeleGeography, "State of the networks," 2022, https://www2.telegeography.com/hubfs/LP-Assets/Ebooks/state-of-the-network-2022.pdf.

[3] A. Weissberger, "Sandvine: Google, facebook, microsoft, apple, amazon & netflix generate almost 57% of internet traffic," 2021, https://tinyurl.com/2uerpfv2.

[4] M. Jinno, H. Takara, B. Kozicki, Y. Tsukishima, Y. Sone, and S. Matsuoka, "Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies," *IEEE Communications Magazine*, vol. 47, no. 11, pp. 66–73, 2009.

[5] T. Mizuno, H. Takara, A. Sano, and Y. Miyamoto, "Dense space-division multiplexed transmission systems using multi-core and multi-mode fiber," *Journal of Lightwave Technology*, vol. 34, no. 2, pp. 582–592, 2016.

[6] Y. Ujjwal and J. Thangaraj, "Review and analysis of elastic optical network and sliceable bandwidth variable transponder architecture," *Optical Engineering*, vol. 57, no. 11, pp. 1–18, 2018.

[7] R. Zhou, M. D. Gutierrez Pascual, P. M. Anandarajah, T. Shao, F. Smyth, and L. P. Barry, "Flexible wavelength de-multiplexer for elastic optical networking," *Optics Letters*, vol. 41, no. 10, pp. 2241–2244, 2016.

[8] J. Wu, S. Subramaniam, and H. Hasegawa, "Comparison of oxc node architectures for wdm and flex-grid optical networks," in *Proceedings of the 2015 24th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–8, Las Vegas, NV, USA, August 2015.

[9] C. Politi, T. Orphanoudakis, E. Kosmatos, and H. C. Leligou, "Dynamic resource allocation in elastic optical networks," in *Proceedings of the 2015 17th International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4, Budapest, Hungary, July 2015.

[10] Y. Awaji, K. Saitoh, and S. Matsuo, "Chapter 13 - transmission systems using multicore fibers," in *Optical Fiber Telecommunications*, I. P. Kaminow, T. Li, and A. E. Willner, Eds., pp. 617–651, Academic Press, Boston, MA, USA, 6th edition, 2013.

[11] P. Lu, L. Zhang, X. Liu, J. Yao, and Z. Zhu, "Highly efficient data migration and backup for big data applications in elastic optical inter-data-center networks," *IEEE Network*, vol. 29, no. 5, pp. 36–42, 2015.

[12] W. Wei, H. Gu, K. Wang, X. Yu, and X. Liu, "Improving cloud-based iot services through virtual network embedding in elastic optical inter-dc networks," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 986–996, 2019.

[13] R. Zhu, S. Li, P. Wang, M. Xu, and S. Yu, "Energy-efficient deep reinforced traffic grooming in elastic optical networks for cloud–fog computing," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12410–12421, 2021.

[14] S. Miladić-Tešić, G. Marković, D. Peraković, and I. Cvitić, "A review of optical networking technologies supporting 5g communication infrastructure," *Wireless Networks*, vol. 28, no. 1, pp. 459–467, 2022.

[15] A. A. Barakabitze and R. Walshe, "Sdn and nfv for qoe-driven multimedia services delivery: the road towards 6g and beyond networks," *Computer Networks*, vol. 214, Article ID 109133, 2022.

[16] J. D. Downie, X. Liang, and S. Makovejs, "Modeling the techno-economics of multicore optical fibers in subsea transmission systems," *Journal of Lightwave Technology*, vol. 40, no. 6, pp. 1569–1578, 2022.

[17] C. Papapavlou, K. Paximadis, D. Uzunidis, and I. Tomkos, "Toward sdm-based submarine optical networks: a review of their evolution and upcoming trends," *Tele.com*, vol. 3, no. 2, pp. 234–280, 2022.

[18] K. Saitoh and S. Matsuo, "Multicore fiber technology," *Journal of Lightwave Technology*, vol. 34, no. 1, pp. 55–66, 2016.

[19] I. Brasileiro, L. Costa, and A. Drummond, "A survey on crosstalk and routing, modulation selection, core and spectrum allocation in elastic optical networks," 2019, https://arxiv.org/abs/1907.08538.

[20] Z. Luo, S. Yin, L. Jiang, L. Zhao, and S. Huang, "Routing, spectrum and core assignment based on auxiliary matrix in the intra data center networks using multi-core fibers with super channel," in *Proceedings of the 2020 Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC)*, pp. 1–3, Beijing, China, October 2020.

[21] R. Llorente, V. Fito, and M. Morant, "Optical combs and multicore fiber as technology enablers for next-generation datacenter infrastructure," in *Metro and Data Center Optical Networks and Short-Reach Links V*, A. K. Srivastava, M. Glick, and Y. Akasaka, Eds., vol. 12027, Bellingham, WA, USA, International Society for Optics and Photonics, Article ID 120270E, 2022.

[22] H. Tode and Y. Hirota, "Routing, spectrum and core assignment for space division multiplexing elastic optical networks," in *Proceedings of the 2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, pp. 1–7, Funchal, Portugal, September 2014.

[23] S. Fujii, Y. Hirota, T. Watanabe, and H. Tode, "Dynamic spectrum and core allocation with spectrum region reducing costs of building modules in aod nodes," in *Proceedings of the 2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, pp. 1–6, Funchal, Portugal, September 2014.

[24] H. M. N. S. Oliveira and N. L. S. da Fonseca, "Protection, routing, spectrum and core allocation in eons-sdm for efficient spectrum utilization," in *Proceedings of the 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, May 2019.

[25] A. Samuel, Y. Zhang, and R. Zhu, "Deadline-aware multicast resource allocation in sdm-eons with fluctuating delay-sensitive traffic," *Journal of Lightwave Technology*, vol. 40, no. 16, pp. 5355–5368, 2022.

[26] J. Žerovnik, "Heuristics for np-hard optimization problems: simpler is better," *Logistics & Sustainable Transport*, vol. 6, no. 1, pp. 1–10, 2015.

[27] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. J. B. Yoo, "Deeprmsa: a deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4155–4163, 2019.

[28] B. Tang, Y.-C. Huang, Y. Xue, and W. Zhou, "Heuristic reward design for deep reinforcement learning-based routing, modulation and spectrum assignment of elastic optical networks," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2675–2679, 2022.

[29] T. Panayiotou, M. Michalopoulou, and G. Ellinas, "Survey on machine learning for traffic-driven service provisioning in optical networks," 2022, https://arxiv.org/abs/2209.05080.

[30] X. Chen, R. Proietti, C.-Y. Liu, Z. Zhu, and S. J. B. Yoo, "Exploiting multi-task learning to achieve effective transfer deep reinforcement learning in elastic optical networks," in *Optical Fiber Communication Conference (OFC) 2020*Optical Society of America, Washington, DC, USA, 2020.

[31] C. Natalino and P. Monti, "The optical rl-gym: an open-source toolkit for applying reinforcement learning in optical networks," in *Proceedings of the 2020 22nd International Conference on Transparent Optical Networks (ICTON)*, pp. 1–5, Bari, Italy, July 2020.

[32] B. Li and Z. Zhu, "Deepcoop: leveraging cooperative drl agents to achieve scalable network automation for multi-domainsd-eons," in *2020 Optical Fiber Communications Conference and Exhibition*, pp. 1–3, OFC), 2020.

[33] P. Morales, P. Franco, A. Lozada et al., "Multi-band environments for optical reinforcement learning gym for resource allocation in elastic optical networks," in *Proceedings of the 2021 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, Gothenburg, Sweden, June 2021.

[34] N. E. D. E. Sheikh, E. Paz, J. Pinto, and A. Beghelli, "Multi-band provisioning in dynamic elastic optical networks: a comparative study of a heuristic and a deep reinforcement learning approach," in *Proceedings of the 2021 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–3, Gothenburg, Sweden, June 2021.

[35] X. Luo, C. Shi, L. Wang, X. Chen, Y. Li, and T. Yang, "Leveraging double-agent-based deep reinforcement learning to global optimization of elastic optical networks with enhanced survivability," *Optics Express*, vol. 27, no. 6, pp. 7896–7911, 2019.

[36] R. Zhu, S. Li, P. Wang, L. Li, A. Samuel, and Y. Zhao, "Deep reinforced energy efficient traffic grooming in fog-cloud elastic optical networks," in *Proceedings of the 2020 Optical Fiber Communications Conference and Exhibition (OFC)*, pp. 1–3, San Jose, CA, USA, March 2020.

[37] X. Tian, B. Li, R. Gu, and Z. Zhu, "Reconfiguring multicast sessions in elastic optical networks adaptively with graph-aware deep reinforcement learning," *Journal of Optical Communications and Networking*, vol. 13, no. 11, pp. 253–265, 2021.

[38] R. Li, R. Gu, W. Jin, and Y. Ji, "Learning-based cognitive hitless spectrum defragmentation for dynamic provisioning in elastic optical networks," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1600–1604, 2021.

[39] R. Zhu, G. Li, P. Wang, M. Xu, and S. Yu, "Drl-based deadline-driven advance reservation allocation in eons for cloud–edge computing," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21444–21457, 2022.

[40] C. Wang, N. Yoshikane, F. Balasis, and T. Tsuritani, "Deepcms3: a deep reinforcement learning framework for core, mode and spectrum sequential scheduling over optical transport network," in *Proceedings of the 2020 European Conference on Optical Communications (ECOC)*, pp. 1–4, Brussels, Belgium, December 2020.

[41] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: an overview," in *Proceedings of the SAI Intelligent Systems Conference (IntelliSys) 2016*, Y. Bi, S. Kapoor, and R. Bhatia, Eds., Springer International Publishing, London, UK, pp. 426–440, September 2018.

[42] Y. Xiong, Y. Yang, Y. Ye, and G. N. Rouskas, "A machine learning approach to mitigating fragmentation and crosstalk in space division multiplexing elastic optical networks," *Optical Fiber Technology*, vol. 50, pp. 99–107, 2019.

[43] Y. Xiong, Y. Ye, H. Zhang, J. He, B. Wang, and K. Yang, "Deep learning and hierarchical graph-assisted crosstalk-aware fragmentation avoidance strategy in space division multiplexing elastic optical networks," *Optics Express*, vol. 28, no. 3, pp. 2758–2777, 2020.

[44] Q. Yao, H. Yang, R. Zhu et al., "Core, mode, and spectrum assignment based on machine learning in space division multiplexing elastic optical networks," *IEEE Access*, vol. 6, pp. 15898–15907, 2018.

[45] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[46] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An Introduction to Deep Reinforcement Learning*, vol. 1, Now Foundations and Trends, Hanover; MA, USA, 2018.

[47] B. Kozicki, H. Takara, Y. Sone, A. Watanabe, and M. Jinno, "Distance-adaptive spectrum allocation in elastic optical path network (slice) with bit per symbol adjustment," in *Proceedings of the 2010 Conference on Optical Fiber Communication (OFC/NFOEC)*, pp. 1–3, San Diego, CA, USA, March 2010.

[48] A. Muhammad, G. Zervas, and R. Forchheimer, "Resource allocation for space-division multiplexing: optical white box versus optical black box networking," *Journal of Lightwave Technology*, vol. 33, no. 23, pp. 4928–4941, 2015.

[49] Y. Zhao, Y. Zhu, C. Wang et al., "Super-channel oriented routing, spectrum and core assignment under crosstalk limit in spatial division multiplexing elastic optical networks," *Optical Fiber Technology*, vol. 36, pp. 249–254, 2017.

[50] I. Brasileiro, L. Costa, and A. Drummond, "A survey on challenges of spatial division multiplexing enabled elastic optical networks," *Optical Switching and Networking*, vol. 38, Article ID 100584, 2020.

[51] C. Chen, M. Ju, S. Xiao, F. Zhou, and X. Yang, "Minimizing total blocking by setting optimal guard band in nonlinear elastic optical networks," in *Proceedings of the 2017 19th International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4, Bari, Italy, July 2017.

[52] A. Hill, A. Raffin, M. Ernestus et al., "Stable baselines," 2018, https://github.com/hill-a/stable-baselines.

[53] V. Mnih, A. P. Badia, M. Mirza et al., "Asynchronous methods for deep reinforcement learning," 2016, https://arxiv.org/abs/1602.01783.

[54] Y. Wu, E. Mansimov, S. Liao, R. Grosse, and J. Ba, "Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation," 2017, https://arxiv.org/abs/1708.05144.

[55] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, https://arxiv.org/abs/1707.06347.

[56] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 1889–1897, Lille, France, December 2015.

[57] D. L. Mills and H. Braun, "The nsfnet backbone network," in *Proceedings of the ACM Workshop on Frontiers in Computer Communications Technology, SIGCOMM '87*, pp. 191–196, Association for Computing Machinery, Stowe, Vermont, August 1987.

[58] P. Batchelor, B. Daino, P. Heinzmann et al., "Study on the implementation of optical transparent transport networks in the european environment—results of the research project

cost 239," *Photonic Network Communication*, vol. 2, pp. 15–32, 2000.

[59] M. Klinkowski and G. Zalewski, "Dynamic crosstalk-aware lightpath provisioning in spectrally-spatially flexible optical networks," *Journal of Optical Communications and Networking*, vol. 11, no. 5, pp. 213–225, 2019.

[60] G. M. Saridis, D. Alexandropoulos, G. Zervas, and D. Simeonidou, "Survey and evaluation of space division multiplexing: from technologies to optical networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2136–2156, 2015.

[61] S. Fujii, Y. Hirota, and H. Tode, "Dynamic resource allocation with virtual grid for space division multiplexed elastic optical network," in *Proceedings of the 39th European Conference and Exhibition on Optical Communication (ECOC 2013)*, pp. 1–3, London, UK, September 2013.

[62] A. A. Saleh and J. M. Simmons, "Technology and architecture to enable the explosive growth of the internet," *IEEE Communications Magazine*, vol. 49, no. 1, pp. 126–132, 2011.

[63] J. Suárez-Varela, P. Almasan, M. Ferriol-Galmés et al., "Graph neural networks for communication networks: context, use cases and opportunities," 2021, https://arxiv.org/abs/2112.14792.

Hindawi
Complexity
Volume 2022, Article ID 3616163, 16 pages
https://doi.org/10.1155/2022/3616163

Research Article

Predicting the Robustness of Large Real-World Social Networks Using a Machine Learning Model

Ngoc-Kim-Khanh Nguyen,[1] **Quang Nguyen**,[2,3,4] **Hai-Ha Pham**,[5] **Thi-Trang Le**,[4] **Tuan-Minh Nguyen**,[4] **Davide Cassi**,[6,7] **Francesco Scotognella**,[8,9] **Roberto Alfieri**,[6,7] **and Michele Bellingeri**[6,7,8]

[1]*Faculty of Basic Science, Van Lang University, Ho Chi Minh, Vietnam*
[2]*Institute of Fundamental and Applied Sciences, Duy Tan University, Ho Chi Minh 700000, Vietnam*
[3]*Faculty of Natural Sciences, Duy Tan University, Da Nang 550000, Vietnam*
[4]*John von Neumann Institute, Vietnam National University Ho Chi Minh City, Ho Chi Minh, Vietnam*
[5]*Vietnam National University, International University, Department of Mathematics, Thu Duc, Ho Chi Minh, Vietnam*
[6]*Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università di Parma, Parco Area Delle Scienze 7/A 43124, Parma, Italy*
[7]*INFN, Gruppo Collegato di Parma, I-43124 Parma, Italy*
[8]*Dipartimento di Fisica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*
[9]*Center for Nano Science and Technology PoliMi, Istituto Italiano di Tecnologia, Via Giovanni Pascoli 70/3, 20133 Milan, Italy*

Correspondence should be addressed to Quang Nguyen; nguyenquang29@duytan.edu.vn

Received 30 June 2022; Revised 24 September 2022; Accepted 3 October 2022; Published 9 November 2022

Academic Editor: Andrea Murari

Computing the robustness of a network, i.e., the capacity of a network holding its main functionality when a proportion of its nodes/edges are damaged, is useful in many real applications. The Monte Carlo numerical simulation is the commonly used method to compute network robustness. However, it has a very high computational cost, especially for large networks. Here, we propose a methodology such that the robustness of large real-world social networks can be predicted using machine learning models, which are pretrained using existing datasets. We demonstrate this approach by simulating two effective node attack strategies, i.e., the recalculated degree (RD) and initial betweenness (IB) node attack strategies, and predicting network robustness by using two machine learning models, multiple linear regression (MLR) and the random forest (RF) algorithm. We use the classic network robustness metric R as a model response and 8 network structural indicators (NSI) as predictor variables and trained over a large dataset of 48 real-world social networks, whose maximum number of nodes is 265,000. We found that the RF model can predict network robustness with a mean squared error (RMSE) of 0.03 and is 30% better than the MLR model. Among the results, we found that the RD strategy has more efficacy than IB for attacking real-world social networks. Furthermore, MLR indicates that the most important factors to predict network robustness are the scale-free exponent α and the average node degree $<k>$. On the contrary, the RF indicates that degree assortativity a, the global closeness, and the average node degree $<k>$ are the most important factors. This study shows that machine learning models can be a promising way to infer social network robustness.

1. Introduction

The study of the social network from a complexity science perspective has attracted much interest recently [1]. Especially, the study of dynamic processes that take place in these complex networks can have various applications. For example, the study of network robustness, i.e., "network robustness" is the capacity of a network to hold its functionality when a proportion of nodes/edges are removed, can help attack a network efficiently, or inversely

design a more robust network structure in practice [2–7]. On the other hand, the study of epidemic processes that take place in the network can be used to spread the news [8–12], optimize vaccination strategy [13–15], or define a better social-distancing rule [16–19].

Besides a few simple model networks where analytical models can be developed [20–24], most of the studies rely on computer simulations. For example, for the study of the network's robustness, node/edge removal Monte-Carlo simulations are usually employed. In such a process, nodes/edges are sequentially removed from the network using computer simulations. A "robustness" metric is then recorded at each step of the removal process. The most commonly used robustness metric is the largest connected component (LCC) of the remaining network [25].

The way nodes/edges are selected to be removed is called the removal strategy or attack strategy. One can classify attack strategies into two types, initial and recalculated attack strategies. For an initial attack strategy, nodes/edges are removed according to a node/edge ranking that is computed ahead of the removal simulation. In contrast for a recalculated attack strategy, the ranking is updated after each node/edge removal [4].

For node removal attack strategies, the node ranking is usually computed using node centrality measures such as degree [26, 27], closeness [4], and betweenness [7, 30]. It was found that for social networks, the recalculated betweenness node attack strategy (RB) is, on average, the most effective node attack strategy to dismantle the network [2, 7, 28, 29]. Other effective strategies are the recalculated degree (RD) and the initial betweenness (IB) [7, 28, 30].

Because of the sequential nature of the removal process, the node removal simulation is computationally costly, especially for recalculated strategies. For example, a simulation using an RD attack strategy has a time complexity of $O(N \times E)$, where N is the number of nodes and E is the number of edges of the network. The reason is that the node removal process has an N step, and at each step, a degree ranking is computed taking a time that scales with E. However, for RB, the computation of the whole network's betweenness is known to be very computationally costly, due to the definition of the network's node betweenness [31, 32]. The most efficient known algorithm for calculating network betweenness is the Brandes algorithm [33], which has a time complexity of O $(N \times E)$. In consequence, the whole node removal process using IB and RB attack strategies can have a time complexity of O $(N \times E)$ and O $(N^2 \times \mathrm{E})$, respectively. Although the IB attack strategy has the same time complexity as the RD attack strategy, the RB's time complexity is much higher. For illustration, in Figure 1, we present the total simulation time tIB and tRD for the corresponding attack strategies IB and RD, respectively, for all our studying social networks (48 networks see Section 2). In addition, we present the total simulation time tRB for the attack strategy RB for 4 networks (insert graph) as an example, as a function of the product $N \times E$. We found a good linear relationship between t_{IB} and t_{RD} and $N \times E$ for all networks as expected, and tRB is about two orders of magnitude higher than t_{IB} and t_{RD}, for networks of equal $N \times E$.

The simulation time can become an issue for the cases of social networks because their size can be extremely large. In fact, to our knowledge, most studies of dynamic processes on social networks that use an RB attack strategy only consider small-size real-world social networks of less than 100,000 nodes [7, 28, 30]. For very large social networks, the RB node attack strategy can take an unrealistic amount of time. Therefore, RB is not suitable for large social networks for an average computer station. One possibility is to use the alternative betweenness-based attack strategy with only one betweenness calculation, namely, the initial betweenness attack strategy IB, together with other recalculated strategies that use another node centrality metric that is less computationally costly. In consequence, in this work, we consider two candidate attack strategies for breaking large real-world social networks, IB and RD attack strategies. Besides the comparative study between different network node attack strategies, other works focused on the relationship between network robustness and network structural indicators (NSIs). Iyer et al. [4] studied network robustness as a function of the node clustering coefficient (or node transitivity). The research on model networks with tunable clustering coefficients demonstrates that networks with higher clustering coefficients are more robust, with the most important effect for the node degree and node betweenness attack [4]. Nguyen and Trang [34] studied Facebook social networks and found that those networks with higher modularity Q have lower robustness to node removal. The modularity indicator Q introduced by Newman and Girvan [35] measures how well a network breaks into communities, (i.e., a community or module in a network is a well-connected group of nodes that have sparser connections with nodes outside the group). In [29], the authors empirically analyzed how the modularity of scale-free models and real-world social networks affects their robustness and the relative efficacy of different node attack strategies. The abovementioned studies analyzed the relationship between network robustness and a single NSI.

On the other hand, machine learning (ML) is a technique that has seen a huge breakthrough in the last decade, beating state-of-the-art results in many prediction applications [36]. It initially solved technical problems in computer vision and natural language processing [37–39] and then expanded into many other fields such as health care, finance, manufacturing, energy, and environment. The key characteristic of an ML model is the ability to intelligently learn nonlinear relationships between the input and output without explicitly knowing them.

In this work, given such a complex relationship between network robustness and NSIs, we adopted a method from machine learning in order to learn such a complexity. Our main contribution is the application of the ML model to predict real-world social network robustness with acceptable errors. We develop ML models to predict network robustness under two main attack strategies, the IB and RD attack strategies, independently. We also implemented three popular ML models, single-variable linear regression, multiple-variable linear regression, and random forest models. Our results demonstrate that a data-driven method

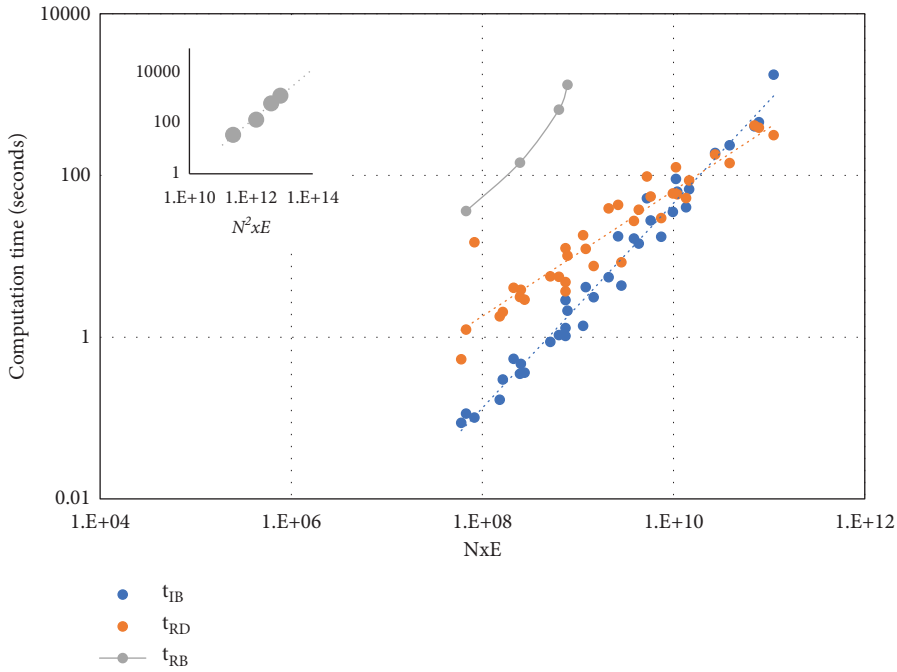


Figure 1: Computation time of a complete Monte Carlo network node attack simulation for all studied real-world social networks (using initial betweenness (IB) and recalculated degree (RD) attack strategy) and for 4 networks (using recalculated betweenness strategy (RB)) as a function of the product $N \times E$ (node number (N) edge number (E). We found that tIB and tRD scale approximately linearly with respect to the product $N \times E$, while tRB scales linearly with respect to the product $N^2 \times E$ (insert graph). From this result, we can estimate that the RB simulation time for the largest networks in our dataset will take more than 50 days using the same hardware.

such as ML can be an efficient way to study the network's complexity.

Our work comprises three steps: (1) collect a real-world network dataset and compute NSIs; (2) run Monte Carlo node attack simulations to estimate network robustness; (3) build and evaluate a model that predicts network robustness from their NSIs. The paper is organized as follows: in Section 2, we describe our dataset of 48 real-world social networks. In Section 3, we describe the network robustness Monte Carlo simulation method and three ML models for predicting the network robustness, i.e., simple and multiple linear regression (SLR and MLR, respectively) and random forest (RF) model. Section 4 presents the main results, and finally, we discuss and conclude in Section 5.

2. Real-World Social Network Datasets and Robustness Estimation

Real-world social networks are downloaded from two sources: the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/) and the Network Repository social networks (https://networkrepository.com/soc.php). We select 48 social networks with a node number (N) ranging over five orders of magnitude. The smallest network is the "Twitch user-user network of gamers who stream in Portugal" having $N = 1{,}914$, and the largest network is the "e-mail network from an EU research institution" with $N = 265{,}216$. However, the network with the largest number of edges (E) is the "BlogCatalog social blog" with $E = 4{,}186{,}390$. The social networks used in this study are unweighted (i.e., we do not take into account edge weights) and undirected (we do not consider edge directionality).

Table 1 summarizes 48 real-world social networks and their NSIs. Besides N and E, we also compute the following NSIs:

(i) Network density <k> is the average node degree, i.e., the average number of edges per node.

(ii) Fitted scaled-free exponent (α): we assume that all social network degree distributions follow a power law of $P(k) \sim k{-}\alpha$ where k is the node degree. The power exponent value α is fitted using the ordinary least squared method. From this fitting, we also extract the fitting variance of α, denoted by α^2.

(iii) Assortativity (a): the assortativity coefficient is a Pearson correlation coefficient of the degree between pairs of linked nodes [40], which varies between −1 and 1. A positive value of a indicates a preferential connection between nodes of a similar degree, while negative values indicate that nodes of different degree have more change to connect.

(iv) Modularity (Q): The modularity indicator Q calculates how a network can be partitioned into subnetworks (modules or communities):

$$Q = \frac{1}{2E}\sum_{i,j}\left(a_{ij} - \frac{k_i k_j}{2E}\right)\delta\left(c_i, c_j\right), \quad (1)$$

Table 1: Structural statistics of real-world social networks: node (N), edge (E), average node degree $<k>$, fitted power-law exponent α, the fitting variance of the power law exponent α^2, assortativity coefficient a, modularity Q, global clustering coefficient C, and average node closeness Cl.

Nb	Network description	Shortname	N	E	$<k>$	α	α^2	a	Q	C	Cl
1	Blue verified Facebook page networks of artist category	Artist	50,515	819,306	32.4	1.937	5.437	0.002	0.177	0.053	0.275
2	Blue verified Facebook page networks of athlete category	Athlete	13,868	86,859	12.5	2.130	4.778	0.005	0.547	0.129	0.237
3	Blue verified Facebook page networks of company category	Company	14,115	52,311	7.4	1.995	4.191	0.022	0.632	0.153	0.193
4	Blue verified Facebook page networks of government category	Government	7,059	89,456	25.3	1.829	4.401	0.004	0.478	0.224	0.270
5	Blue verified Facebook page networks of new site category	new_sites	27,919	206,260	14.8	2.097	5.082	0.009	0.509	0.114	0.233
6	Blue verified Facebook page networks of politician category	Politician	5,910	41,730	14.1	2.058	4.474	0.005	0.660	0.301	0.219
7	Blue verified Facebook page networks of public figure category	public_figure	11,567	67,115	11.6	1.841	4.212	0.009	0.441	0.167	0.221
8	Blue verified Facebook page networks of tV show category	Tvshow	3,894	17,263	8.9	1.622	3.279	0.037	0.770	0.591	0.166
9	Citation NW of arXiv High Energy Physics (phenomenology) paper	Cit-HepPh	34,548	421,579	24.4	2.528	6.310	0.003	0.472	0.146	0.237
10	Citation NW of arXiv High Energy Physics (theory) paper	Cit-HepTh	27,772	352,808	25.4	1.916	5.006	0.006	0.424	0.120	0.000
11	Collaboration network of arXiv astro physics	CA-AstroPh	18,774	396,161	42.2	2.174	6.021	0.050	0.412	0.316	0.000
12	Collaboration network of arxiv condensed matter	CA-CondMat	23,135	186,937	16.2	2.584	6.235	0.074	0.649	0.258	0.245
13	Collaboration network of arxiv general relativity	CA-GrQc	5,244	28,981	11.1	2.290	4.895	0.192	0.781	0.611	0.000
14	Collaboration network of arxiv high energy physics	CA-HepPh	12,010	237,011	39.5	1.407	4.013	0.103	0.383	0.657	0.000
15	Collaboration network of arxiv high energy physics theory	CA-HepTh	9,879	51,972	10.5	3.306	6.907	0.080	0.708	0.272	0.000
16	Deezer's users friendship networks from Croatia	deezer_HR	54,575	498,203	18.3	3.461	7.954	0.005	0.525	0.115	0.224
17	Deezer's users friendship networks from Hungary	deezer_HU	47,540	222,888	9.4	4.435	8.525	0.008	0.580	0.093	0.189
18	Deezer's users friendship networks from Romania	deezer_RO	41,775	125,827	6.0	3.392	6.402	0.008	0.682	0.075	0.160
19	E-mail communication network from enron	Email-enron	36,694	367,663	20.0	1.446	4.213	0.036	0.333	0.085	0.307
20	E-mail network from a EU research institution	Email-EuAll	265,216	420,046	3.2	0.646	1.901	−0.039	0.047	0.007	0.000
21	Follower relationships network of European users from deezer	deezer_Europe	28,283	92,753	6.6	2.981	5.972	0.011	0.603	0.096	0.159
22	Network of trusting consumers from the review site Epinions.com	Soc-Epinions1	75,881	508,838	13.4	1.512	4.289	0.001	0.247	0.082	0.237
23	Page-page network of verified facebook sites	musae_facebook	22,472	171,003	15.2	2.029	4.945	0.011	0.630	0.232	0.206
24	Slashdot social network from February 2009	Slashdot0902	82,170	948,465	23.1	1.617	4.728	0.080	0.202	0.026	0.250
25	Slashdot social network from November 2008	Slashdot0811	77,362	905,469	23.4	1.603	4.685	0.083	0.207	0.026	0.252
26	Social network of github developers.	musae_git	37,702	289,004	15.3	1.267	3.482	−0.001	0.152	0.012	0.314
27	Social network of LastFM users from asia.	lastfm_Asia	7,626	27,807	7.3	1.807	3.730	0.006	0.679	0.179	0.195
28	Twitch user-user networks of gamers who stream in English	musae_ENGB	7,128	35,325	9.9	1.204	2.770	0.001	0.267	0.042	0.277
29	Twitch user-user networks of gamers who stream in French	musae_FR	6,551	112,667	34.4	1.303	3.392	−0.001	0.084	0.054	0.378
30	Twitch user-user networks of gamers who stream in German	musae_DE	9,500	153,139	32.2	1.305	3.476	−0.001	0.062	0.046	0.374
31	Twitch user-user networks of gamers who stream in Portugal	musae_PTBR	1,914	31,300	32.7	1.127	2.680	−0.003	0.081	0.131	0.402

Table 1: Continued.

Nb	Network description	Shortname	N	E	$<k>$	α	α^2	a	Q	C	Cl
32	Twitch user-user networks of gamers who stream in Russian	musae_RU	4,387	37,305	17.0	1.054	2.522	−0.003	0.101	0.049	0.337
33	Twitch user-user networks of gamers who stream in Spain	musae_ES	4,650	59,383	25.5	1.327	3.233	−0.001	0.109	0.084	0.352
34	Wikipedia page-to-page networks on chameleon topic	musae_chameleon	2,279	36,102	31.7	0.974	2.381	0.006	0.203	0.445	0.291
35	Wikipedia page-to-page networks on crocodile topic	musae_crocodile	11,633	180,021	31.0	0.892	2.483	−0.010	0.181	0.039	0.316
36	Wikipedia page-to-page networks on squirrel topic	musae_squirrel	5,203	217,074	83.4	0.748	2.245	−0.001	0.072	0.451	0.335
37	Wikipedia who-votes-on-whom network	Wiki-vote	7,117	103,690	29.1	1.412	3.644	−0.001	0.025	0.136	0.318
38	BlogCatalog social blog	BlogCatalog1	88,784	4,186,390	94.3	2.265	9.866	−0.001	0.018	0.060	0.331
39	BlogCatalog social blog version 2	BlogCatalog2	97,884	2,043,701	41.8	2.141	9.330	−0.001	0.006	0.057	0.355
40	BlogCatalog social blog version 3	BlogCatalog3	10,312	333,983	64.8	1.929	6.939	−0.001	0.026	0.091	0.424
41	Douban online social network	Douban	154,908	654,188	8.4	3.946	10.755	−0.048	0.093	0.010	0.195
42	Gowalla location-based social networking	Gowalla	196,591	950,327	9.7	1.386	5.337	0.006	0.501	0.023	0.221
43	TheMarker cafe online social network	TheMarker	69,413	1,644,849	47.4	2.547	9.799	0.000	0.022	0.046	0.332
44	Brightkite location-based online social network	Brightkite	58,228	214,078	7.4	2.330	6.802	0.005	0.539	0.111	0.224
45	The friendships network between users of the website http://www.hamsterster.com	Hamsterster	2,426	16,630	13.7	2.599	6.051	0.067	0.394	0.231	0.404
46	A google-plus subgraph	Soc-gplus	23,628	39,242	3.3	1.188	4.021	−0.068	−0.027	0.004	0.251
47	Anybeat online social network	Anybeat	12,645	67,053	10.6	0.922	3.294	−s0.009	0.133	0.018	0.323
48	Advogato online social network	Advogato	6,551	51,332	15.7	2.064	5.785	0.078	0.312	0.111	0.000

where E is the number of edges, a_{ij} is the element of the adjacency matrix A in the row i and column j, k_i is the degree of i, k_j is the degree of j, c_i is the module (or community) of i, c_j that of j, the sum goes over all i and j pairs of nodes, and $\delta(x, y)$ is 1 if $x = y$ and 0 otherwise [13].

(v) Global clustering coefficient (C): the global clustering coefficient (C) is based on triplets of nodes. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected edges. The global clustering coefficient is the number of closed triplets (or $3x$ triangles because a triangle comprises 3 overlapping triplets, each centered at one of the three nodes) over the total number of triplets (both open and closed). The formula is as follows:

$$C = \frac{\lambda_{\text{closed}}}{\lambda_{\text{total}}}, \tag{2}$$

where λ_{closed} is the number of closed triplets and λtotal is the total number of triplets in the network. The global clustering coefficient represents the overall probability for the network to have adjacent nodes interconnected, thus making more tightly connected modules [41].

(vi) Average closeness (Cl) is the average of all network nodes' closeness, where the closeness (or closeness centrality) of a node is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph [42, 43]:

$$Cl = \overline{Cl}_i = \frac{1}{N}\sum_{i=1}^{N} Cl_i, \text{with} Cl_i = \frac{1}{\sum_{j\neq i} d(i,j)}, \tag{3}$$

where N is the number of nodes and $d(i, j)$ is the length of the shortest path between nodes i and j.

2.1. Network Robustness Monte Carlo Simulation. For each network, we run two node removal processes using Monte-Carlo simulations. Nodes are removed consecutively following the ranking of initial betweenness (IB) and the ranking of the recalculated degree (RD). In the case of ties, e.g., nodes with an equal betweenness or degree score, we removed one of them at random. After each node removal, we compute the network robustness measure and the relative size of the largest connected component LCC, together with the accumulated proportion of nodes removed q. Finally, we obtain two curves LCC (q) corresponding to two node removal processes, IB and RD. The whole simulation is repeated 10 times, and the final curves LCC (q) are the average results.

In addition, we compute a single value defined as the network robustness (R), as performed by Bellingeri et al. [44], and the area below the normalized LCC curve during the removal process, $R = \overline{\text{LCC}(q)}$. R therefore can be between two theoretical extremes, $R \approx 0$ (absolute fragile network) and $R \approx 0.5$ (absolute robust network). We denote

RRD and RIB as the network robustness against RD and IB node attack strategies, respectively.

In summary, we collect 48 real-world social networks, and then, we compute 9 NSIs for each network as inputs. In parallel, we run Monte Carlo simulations and obtain the robustness represented by two metrics, RRD and RIB. The higher they are, the more robust the network is. Those two metrics are the output of each network and will be predicted using ML models.

3. Machine Learning Approach

This section presents the details of SLR, MLR, and RF models.

3.1. Simple Linear Regression Model (SLR). Linear regression is the simplest model for prediction. The SLR model between the network robustness R and an NSI x is expressed by the linear equation:

$$R = a_0 + a_1 x, \tag{4}$$

where a_0 is the intercept and a_1 is the slope. In (4), an ordinary least square (OLS) is applied for estimating coefficients by minimizing an appropriate loss function [45, 46]. Once the OLS process, which is also called the fitting process, is performed, we can use (1) to predict the robustness R of a new network for a given indicator x. In addition, we derive a statistics t-test from the OLS process with the null hypothesis H0: $a_1 = 0$. A rejection of H0 means that there is a significant linear relationship between R and the NSI x.

We run the SLR model fit for all NSIs listed in Table 1 excluding E because it can be expressed in terms of two other NSIs : $E = N < k > /2$.

3.2. Multiple Linear Regression Model. Multiple linear regression (MLR) is an extension of SLR for multidimension variables $x = (x_1, x_2, \ldots, x_n)$, where $x_1, x_2, \ldots, x_n$ are NSIs. The linear equation between network robustness R and NSIs is as follows:

$$R = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n, \tag{5}$$

where a_i are coefficients obtained from the OLS method.

3.3. Random Forest Model. The random forest (RF) belongs to the ensemble class of ML models, indicating that it aggregates the prediction from an ensemble of ML base models, here, decision tree regression (DTR) models. We briefly describe the DTR in the following section.

A DTR starts with the root of the tree containing all samples (48 networks in our case). It then splits into two different nodes by selecting samples whose value of a certain variable is higher or lower than a certain threshold value. Figure 2(a) represents a basic decision tree diagram for our dataset. The root node containing 48 networks splits into two other nodes by considering whether the variable (NSI in our case) scale-free exponent α is higher or lower than 2.5.

The DTR selects the variable, and its splitting value is based on information theory, in concrete considering the entropy concept. Entropy is a metric of uncertainty of a node. The DTR splits a node by maximizing the information gain, which is the weighted difference between the total entropy of two resulting nodes and the entropy of the initial node. The DTR successively splits until a stopping condition is reached, for example if the size of the current node is smaller than 20. The final node is also called a leaf node. In Figure 2(a), after the first split of the root, the left child node becomes a leaf node, while the right child node continues to split into two leaf nodes.

Once the final DTR is obtained, it can be used to predict the value of a new sample as follows. The new sample will be classified into one of the leaves, and its prediction value will be the average value of all the samples that are classified into the same leaf.

Finally, the RF model creates multiple decision trees randomly drawn from the data, usually several hundred, and averaging the results from all trees to output a new result often leads to strong predictions [47, 48].

The decision tree can fit nonlinear datasets because it can split the same NSI multiple times. However, decision tree is easy to be overfitting, i.e., it is too sensitive to the training data while failing to predict new coming (testing) data. In order to address this problem, a random forest (RF) model is obtained by creating multiple randomly drawn decision trees from data, usually several hundred. The final regression prediction will be the average prediction of all the decision trees [47–49] (in this work, we implement an RF with 300 DTRs). Using an RF, "feature importance" measurement can be derived to rank the NSI [50].

3.4. Data Preparation, Validation, and Performance Evaluation. All NSIs can be computed from the network's data, and thus, our dataset did not contain missing values. We also exclude E because of redundancy as mentioned above. The other 8 NSIs are normalized to avoid large differences in the indicators' range:

$$x'_{i,j} = \left(x_{i,j} - \overline{x}_i\right)\sigma\left(x_i\right), \tag{6}$$

where $x_{i,j}$ is the value of the NSI i for observation (network) j and $\overline{x}_i$ and $\sigma(x_i)$ are the mean and the standard deviation of the NSI i, respectively.

In the first step, we use the whole dataset to build ML models and compare the results between models and two target variables. However, due to overfitting problems in many ML models, the model's performance for new data is not always coherent as that in the training step, and we need to validate models in the second step. We choose the leave-one-out validation [51]. In this way, we train each of the above models 48 times: each time the whole dataset excluding one observation is used to train the model, and then, the model is used to predict the target value of the remaining (hold-out) observations and repeats for each of 48 hold-out observations. The overall evaluation result is the average across all 48 regressions.

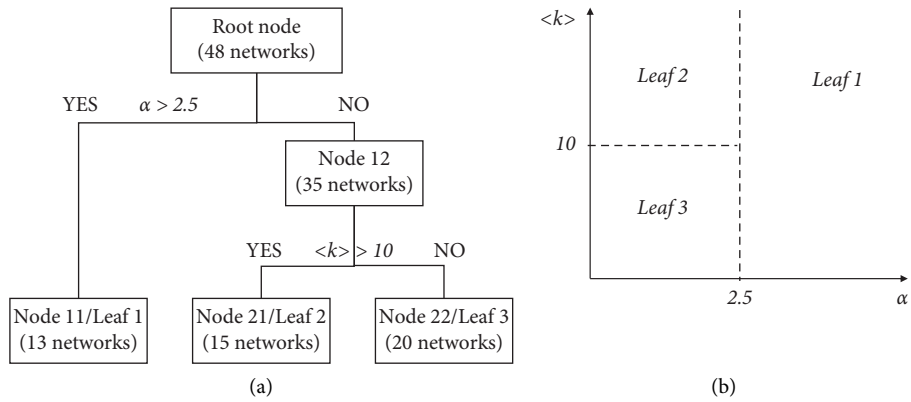


Figure 2: (a) An example of a decision tree: the root node containing 48 networks splits into two other nodes, Node 11 and Node 12, with 13 and 35 networks, respectively, according to the value of the scale-free exponent α. Then, Node 12 splits into two other nodes, Node 21 and Node 22, with 15 and 20 networks, respectively, according to the value of the network density $<k>$. We assume that at Node 11, Node 21, and Node 22, no split is possible because of a certain stopping rule, and thus, they become final leaves. In general, any NSI can be used to divide networks at any split, and the decision tree can be arbitrarily complex depending on the stopping rule. (b) An illustration of the same decision tree in the 2-dimension (α and $<k>$) space with final leaves.

It is noted that for the SLR model, we only consider regression coefficients in order to analyze the dependence of robustness metrics with respect to each NSI. However, for MLR and RF models, we analyze the prediction of robustness metrics using four common evaluation metrics for regression problems, the root mean square error (RMSE) and the coefficient of determination (also named the explained variance ratio, R^2) as analytical metrics and the frequency distribution and the Q-Q plot of residual errors as graphical metrics.

RMSE is the square root of the summation of the squared difference between observed and predicted data points. The RMSE has the same unit as the target feature and is generally considered the model error. A lower RMSE value represents superior prediction results. The formula of the RMSE is provided by

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}\left(R_j - R_{\text{predicted},j}\right)^2}{n}}, \tag{7}$$

where n is the number of observations, R_j denotes the empirical (simulated) network robustness, and $R_{\text{predicted},\,j}$ is the predicted value of robustness for the observation j.

R^2 is used to represent the general prediction performance of regression models. R^2 is one minus the ratio of the remaining variance and the original variance. The formula of R^2 is provided by

$$R^2 = 1 - \frac{\sum_{j=1}^{n}\left(R_j - R_{\text{predicted},j}\right)^2}{\sum_{j=1}^{n}\left(R_j - \overline{R}_j\right)^2}, \tag{8}$$

where n is the number of observations, R_j is the simulation robustness, $R_{\text{predicted},\,j}$ denotes the predicted value for observation j, and $\overline{R}_j$ is the average of all the simulation robustness. R^2 varies between 0 (the model has no prediction ability) and 1 (the model correctly predicts all values).

The residual error, $\varepsilon = R_{\text{empirical}} - R_{\text{predicted}}$, is simply an error between the empirical (simulated) network robustness and the predicted value of robustness. The distribution histogram of ε is expected to be close to the origin. Furthermore, the most important assumption of a linear regression model is that residual errors are independent, and consequently, these errors are expected to be normally distributed.

The network is analyzed using the "graph-tool" library in *Python*. All data preparation, model building, and evaluation are written using the *Python* code. The hardware for numerical simulations is a PC with an i9-10850 Intel processor and 32 GB RAM.

4. Results

4.1. Network Robustness as a Function of the NSIs and SLR T-Test. The simulation robustness of each network RIB and RRD is represented in Table 2. Overall, we found that RRD is slightly smaller than RIB for most networks (43 out of 48 networks), with an average of 0.148 vs. 0.173, respectively. It suggests that the RD strategy has more efficacy than IB for attacking real-world social networks. The largest and sparsest network, Email-EuAll ($N = 265{,}216$ and $<k \geq 1.58$), has the smallest robustness with an equal RIB and RRD of 0.001. In contrast, the gemsec_deezer_HR network, with $N = 54{,}575$ and $<k \geq 9.12$, has the strongest robustness with an RIB and RRD of 0.375 and 0.338, respectively.

In Figure 3, we plot RRD and RIB as a function of 8 independent NSIs, and we found that RRD and RIB behave similarly in all cases. The SLR unveils some significant relationships between R and NSIs (Figure 3 and Table 3). For example, in Figure 3(a), we can see that both RRD and RIB slightly decrease with the network size N. This linear dependence between robustness RRD and RIB and N is tested by using the SLR model, and we found that it is statistically significant, with a confidence level of 95% (p value <0.05, Table 3).

Table 2: Simulation result by IB and RD node attack strategies, represented by the network robustness metrics R_{IB} and R_{RD}, for all 48 real-world social networks (sort by networks' size from smallest to largest).

Short names	N	E	R_{IB}	R_{RD}
musae_PTBR	1,914	31,300	0.257	0.214
musae_chameleon	2,279	36,102	0.153	0.143
Hamsterster	2,426	16,630	0.134	0.133
Tvshow	3,894	17,263	0.139	0.153
musae_RU	4,387	37,305	0.209	0.149
musae_ES	4,650	59,383	0.248	0.202
musae_squirrel	5,203	217,074	0.298	0.184
CA-GrQc	5,244	28,981	0.057	0.069
Politician	5,910	41,730	0.198	0.195
musae_FR	6,551	112,667	0.288	0.240
Advogato	6,551	51,332	0.100	0.090
Government	7,059	89,456	0.311	0.283
Wiki-vote	7,117	103,690	0.136	0.144
musae_ENGB	7,128	35,325	0.180	0.132
lastfm_asia	7,626	27,807	0.171	0.137
musae_DE	9,500	153,139	0.282	0.229
CA-HepTh	9,879	51,972	0.097	0.091
BlogCatalog3	10,312	333,983	0.194	0.181
public_figure	11,567	67,115	0.204	0.168
musae_crocodile	11,633	180,021	0.124	0.071
CA-HepPh	12,010	237,011	0.138	0.162
Anybeat	12,645	67,053	0.039	0.028
Athletes	13,868	86,859	0.234	0.199
Company	14,115	52,311	0.175	0.150
CA-AstroPh	18,774	396,161	0.193	0.212
musae_facebook	22,472	171,003	0.228	0.206
CA-CondMat	23,135	186,937	0.113	0.113
Soc-gplus	23,628	39,242	0.002	0.001
Cit-HepTh	27,772	352,808	0.342	0.307
new_sites	27,919	206,260	0.264	0.228
deezer_Europe	28,283	92,753	0.186	0.153
Cit-HepPh	34,548	421,579	0.350	0.307
Email-Enron	36,694	367,663	0.048	0.039
musae_git	37,702	289,004	0.168	0.122
deezer_RO	41,775	125,827	0.261	0.200
deezer_HU	47,540	222,888	0.343	0.287
Artists	50,515	819,306	0.299	0.265
deezer_HR	54,575	498,203	0.375	0.338
Brightkite	58,228	214,078	0.107	0.083
TheMarker	69,413	1,644,849	0.113	0.100
Soc-Epinions1	75,881	508,838	0.066	0.054
Slashdot0811	77,362	905,469	0.093	0.073
Slashdot0902	82,170	948,465	0.103	0.077
BlogCatalog1	88,784	4,186,390	0.072	0.063
BlogCatalog2	97,884	2,043,701	0.016	0.014
Douban	154,908	654,188	0.026	0.024
Gowalla	196,591	950,327	0.160	0.115
Email-EuAll	265,216	420,046	0.001	0.001
Average	38,026	391,698	0.173	0.148
Std	51,490	687,601	0.098	0.085

Interestingly, RRD and RIB do not statistically linearly depend on the network density $<k>$ as found previously in [4, 52] (Figure 3(b) and Table 3). This contrasting observation would suggest that network robustness also depends on other NSIs and that the network density alone cannot predict the whole network's robustness as previously seen.

Besides N, the only other NSI that shows a significant linear relationship is the modularity Q (Figure 3(f)) in the case of RRD.

However, in Figure 3, we still observe some nonlinear dependencies. For example, in Figure 3(e), we show that network robustness decreases with the assortativity coefficient a when $a > 0$. However, it decreases faster when a is close to 0 and increases with a when $a < 0$.

Similarly, in Figure 3(g), we found that the relationship between RRD and RIB and the global clustering coefficient C follows an inverted u-shaped pattern. We ran a two-line statistical test [53] and found that two-line (or broken line) regression is significantly better than a single-line test. The breakpoint was found to be $C = 0.115$. Both RRD and RIB linearly increase with C (with a significance level of 95%) up to the breakpoint and linearly decrease with C (with a significance level of 95%). One possible explanation is that if the network is sparse, more triplets help increase the network's connectivity and thus increase its robustness. However, above a certain value (when $C = 0.115$), more triplets may denote the presence of hubs or central nodes, which are likely to be the target of intentional node removal strategies such as RD and IB, consequently lowering network robustness.

4.2. Machine Learning Prediction of Network Robustness. The results of the previous section suggest that the social network's robustness depends on multiple NSIs in a highly complex, multidimensional, and nonlinear manner. To improve the model prediction, in this section, we use two multiple variable ML models, MLR and RF, to predict network robustness.

The results of multiple linear regression MLR are shown in Table 4. We found that both R_{IB} and R_{RD} have a positive overall linear regression coefficient with respect to α, Q, Cl, and $<k>$ and a negative overall linear regression coefficient with respect to α^2, a, C, and N. Moreover, the MLR result indicate that α, α^2, and $<k>$ are the most significant coefficients. A positive linear regression coefficient for the average node degree $<k>$ suggests that networks are more robust when k is higher, while all other NSIs are fixed. This result agrees with previous outcomes demonstrating that denser networks may be more resistant to the attack [4, 52]. However, the different results between the MLR and SLR would suggest that there is a strong correlation between $<k>$ and other NSIs. In addition, the MLR model predicts R_{IB} better than R_{RD}, with an R^2 coefficient of 58.04% compared to 51.76%. Nevertheless, the RMSE was smaller for R_{RD}, with a value of 0.0657, compared to 0.0709 for RIB (this is because the standard deviation of RIB is higher than that of R_{RD}, as shown in Table 2 (bottom row)).

Because of the nonlinearity found in the previous section, we expect that the regression result using the RF model will be improved. Table 5 represents the regression result of the RF model. We found that R^2 increases to 92.24% and 91.88% for R_{IB} and R_{RD} regressions, respectively. Interestingly, the RF model predicts R_{RD} roughly as well as R_{RD}, while MLR predicts R_{IB} better than R_{RD}, suggesting that R_{RD}

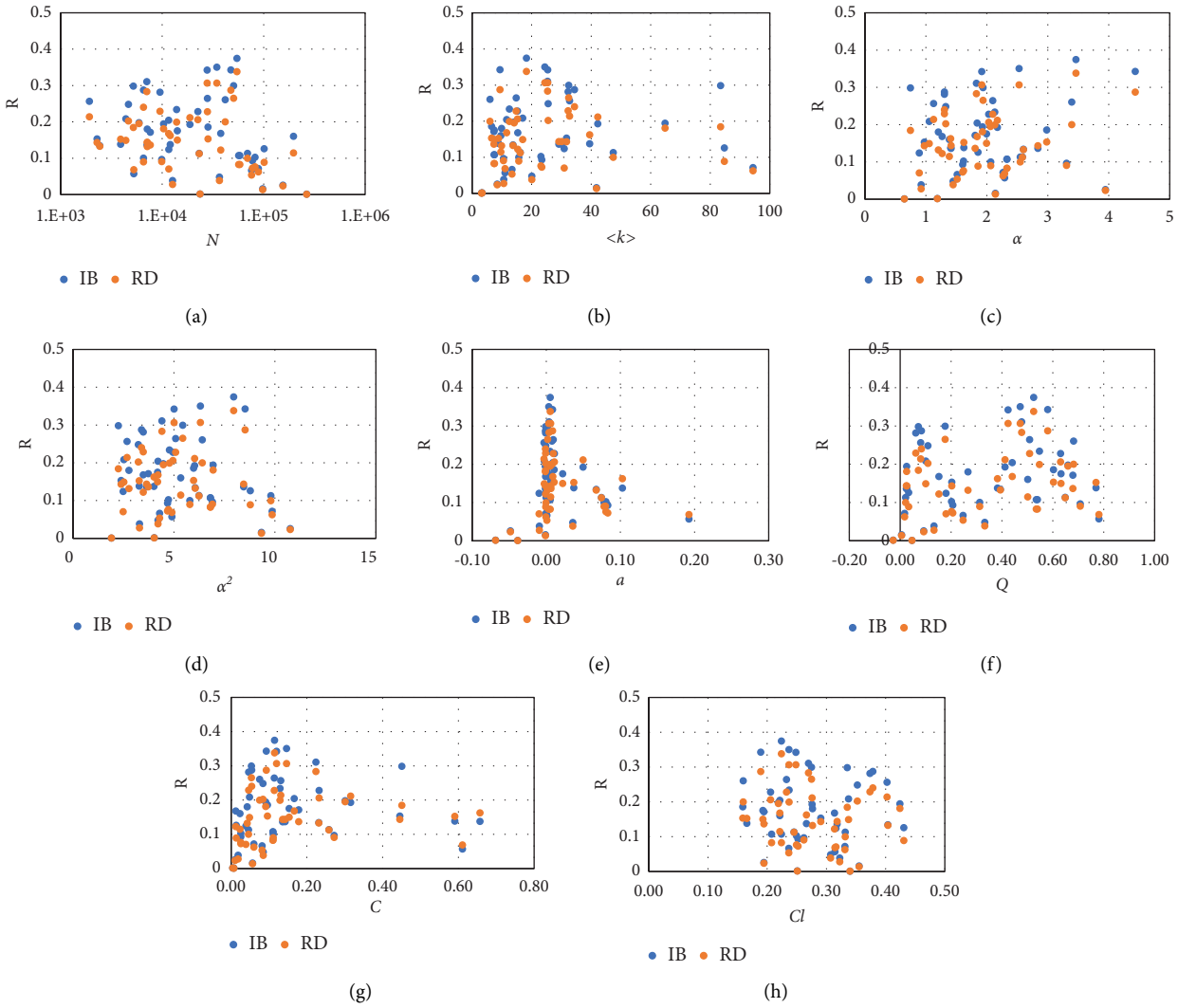


Figure 3: Simulation result by IB and RD node attack strategies, represented by the network robustness metrics R_{IB} and R_{RD}, for all the 48 real-world social networks as a function of 8 NSIs.

Table 3: The SLR results for 8 NSIs. The last two columns show the slope, and in parenthesis, the R^2 values of the SLR between the NSI and RIB or RRD. The bold character with an asterisk indicates a significant relationship, with a confidence level of 95%.

Nb	NSI	R_{IB}	R_{RD}
1	N	$\mathbf{-6.920{\cdot}10^{-7}\ (0.132)}^{*}$	$\mathbf{-6.377{\cdot}10^{-7}\ (0.150)}^{*}$
2	<k>	0.0006 (0.015)	0.0004 (0.010)
3	α	0.0215 (0.031)	0.0247 (0.056)
4	α^2	−0.0039 (0.007)	−0.0011 (0.000)
5	α	−0.3282 (0.019)	−0.1184 (0.003)
6	Q	0.0917 (0052)	$\mathbf{0.1022\ (0.087)}^{*}$
7	C	0.0274 (0.001)	0.0781 (0.021)
8	Cl	−0.1515 (0.010)	−0.1607 (0.016)

may follow a stronger nonlinear relationship with NSIs than RIB. Additionally, the RMSE improved both for R_{IB} and R_{RD}, with a value of 0.0272 and 0.0241, respectively. Interestingly, the feature importance ranking in Table 5 shows that with an RF model, the assortativity a, the global closeness C, and the node number N are the most important NSIs. This result agrees with the exploratory observations shown in Figure 3 as discussed above.

In Figure 4, we compare network robustness R_{IB} and R_{RD} with the prediction value given by MLR and RF using a scatter plot. The scatter plots indicate that RF fit data significantly better than MLR, where the predicted actual data points are closer to the diagonal line $y = x$. Meanwhile, for MLR regression, we still found nonlinear dependency between the actual and predicted values. As a matter of fact, the MLR model was not able to capture the inherent nonlinearity dependency in the actual data. We also analyzed the residual errors of the above regression using the frequency histogram and QQ-plot and found that they follow a normal distribution relatively well (Figures 5–8).

Finally, we run leave-one-out regression for both models MLR and RF in order to avoid overfitting. The result is summarized in Table 6, and the scatter plots are shown in Figure 9. We found that the prediction result is less accurate than the above "in-sample" training with lower RMSEs in both MLR and RF models. We obtained an RMSE of 0.0812

Table 4: Fit coefficients and the evaluation result given by the MLR. R_{IB} or R_{RD} columns show the slope coefficient, and in parenthesis, the standard error values for the NSI. The bold character with an asterisk indicates a significant relationship between the NSI and the robustness R, with a confidence level of 95%.

Nb	NSI	Regression coefficients	
		R_{IB}	R_{RD}
1	α	**0.113 (0.028)***	**0.088 (0.026)***
2	α^2	**−0.118 (0.028)***	**−0.083 (0.026)***
3	α	**−0.029 (0.013)***	−0.025 (0.012)
4	Q	**0.047 (0.02)***	0.038 (0.019)
5	C	**−0.034 (0.016)***	−0.016 (0.015)
6	Cl	0.005 (0.014)	0.007 (0.013)
7	N	−0.013 (0.013)	−0.014 (0.012)
8	$<k>$	**0.077 (0.017)***	**0.056 (0.016)***
9	Intercept	**0.173 (0.01)***	**0.148 (0.009)***
MLR results			
	RMSE	0.0709	0.0657
	R^2	58.04%	51.76%

Table 5: Feature importance of the NSI and the evaluation result given by RF.

NSI	Feature importance	
	R_{IB}	R_{RD}
α	0.0622	0.0654
α^2	0.0535	0.0463
a	0.2765	0.1912
Q	0.0823	0.0658
C	0.1114	0.1834
Cl	0.0581	0.0584
N	0.2683	0.2759
$<k>$	0.0873	0.1133
RMSE	0.0272	0.0241
R^2	92.24%	91.88%

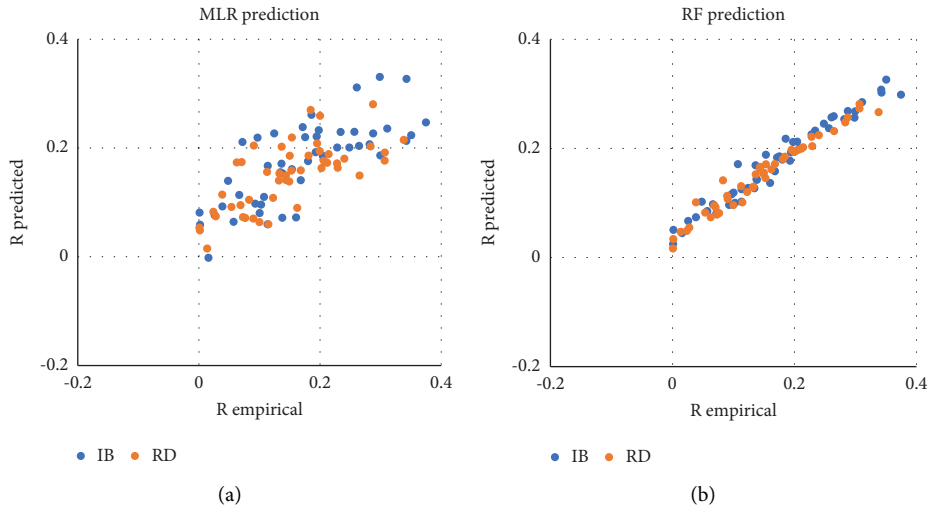


Figure 4: Scatter plots between the predicted value of robustness (R predicted) and simulated (R empirical) for MLR (a) and RF model (b). The model is trained using the whole dataset, and the predicted values are of the same dataset.

and 0.0760 for R_{IB} and R_{RD} predictions using MLR, respectively, and an RMSE of 0.0733 and 0.0636 for R_{IB} and R_{RD} predictions using RF, respectively. Even though the regression results are less effective because we predict the single sample which is independent of the remaining samples used for training (building the ML model), residual errors still fit well to a normal distribution as shown in the histogram and QQ-plots (Figures 10–13).

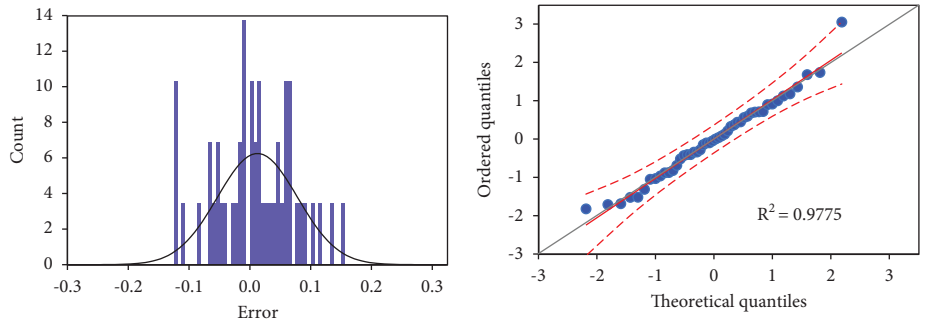


Figure 5: Histogram of residual errors for MLR prediction of the IB strategy for the whole dataset (a) and its QQ-plot (b).

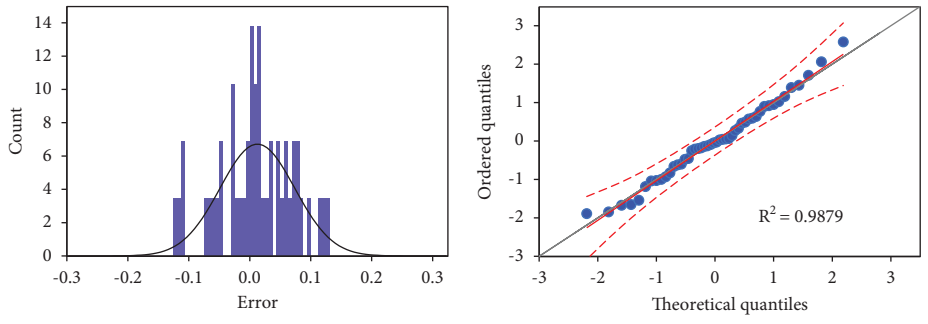


Figure 6: Histogram of residual errors for MLR prediction of the RD strategy for the whole dataset (a) and its QQ-plot (b).

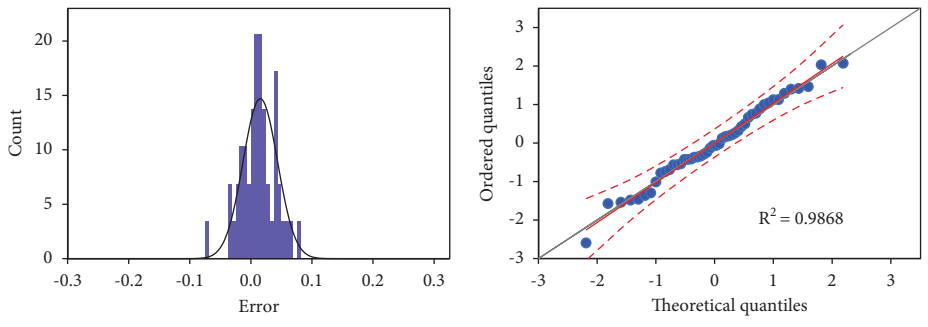


Figure 7: Histogram of residual errors for RF prediction of the IB strategy for the whole dataset (a) and its QQ-plot (b).

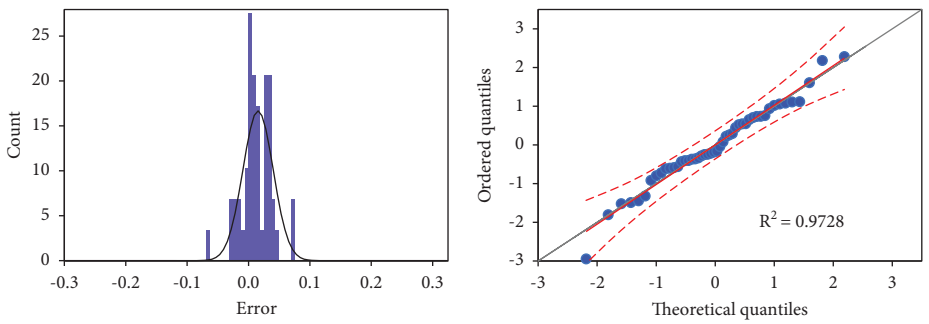


Figure 8: Histogram of residual errors for RF prediction of the RD strategy for the whole dataset (a) and its QQ-plot (b).

Table 6: MLR and RF evaluation results using the leave-one-out method.

	MLR	RF
R_{IB}		
RMSE	0.0812	0.0733
R^2	31.30%	43.87%
R_{RD}		
RMSE	0.0760	0.0636
R^2	19.30%	43.47%

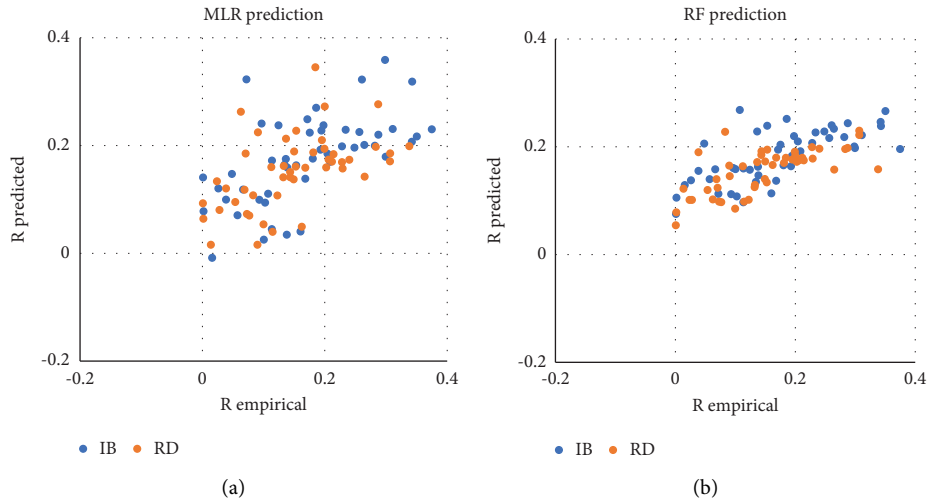


Figure 9: Scatter plots between the predicted value of robustness (R predicted) and simulated (R empirical) of the hold-out observation for MLR (a) and RF model (b). The model is trained using the whole dataset excluding one observation (hold-out observation) and is used to predict the outcome of the hold-out observation.

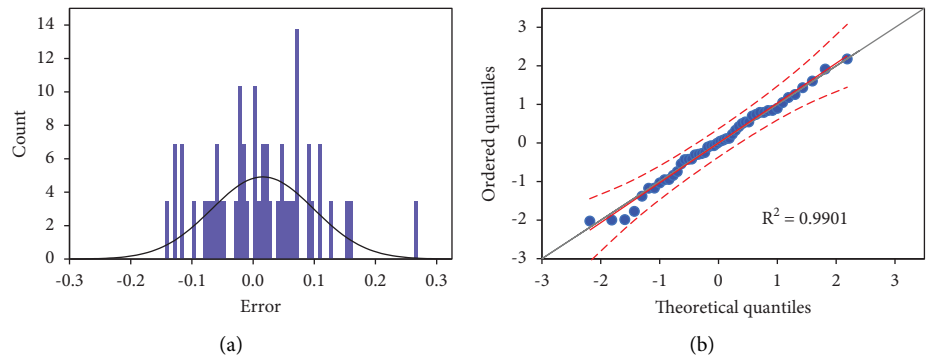


Figure 10: Histogram of residual errors for MLR prediction of the IB strategy for the leave-one-out sample (a) and its QQ-plot (b).

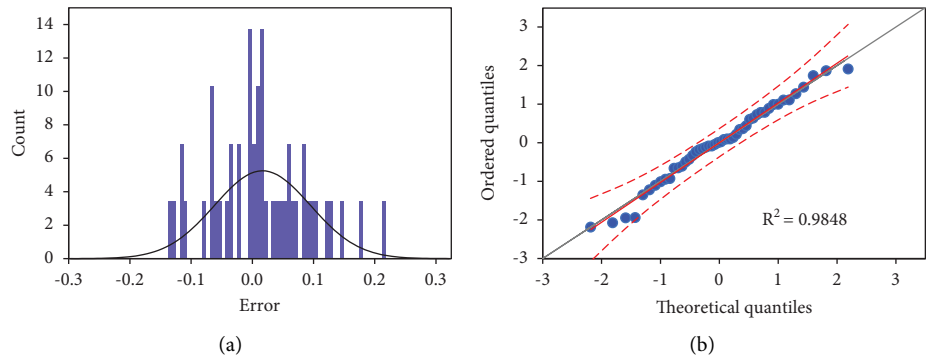


Figure 11: Histogram of residual errors for MLR prediction of the RD strategy for the leave-one-out sample (a) and its QQ-plot (b).

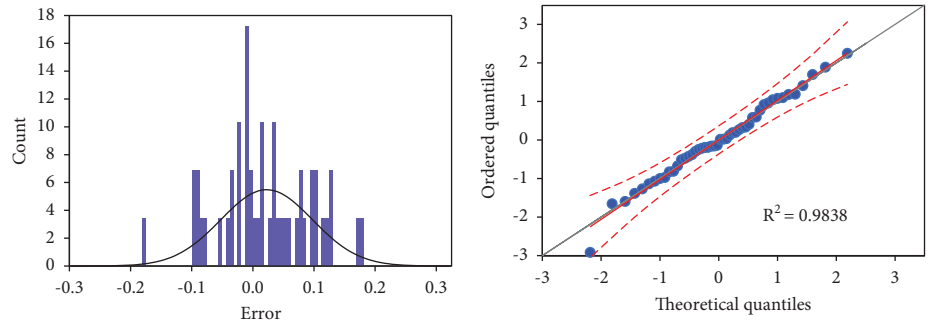


Figure 12: Histogram of residual errors for RF prediction of the IB strategy for the leave-one-out sample (a) and its QQ-plot (b).

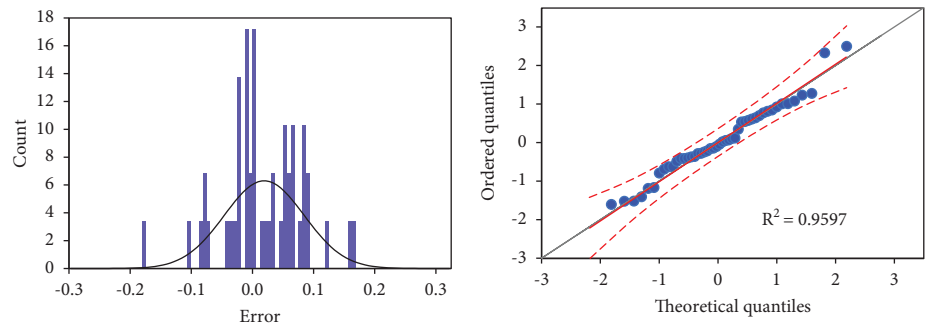


Figure 13: Histogram of residual errors for RF prediction of the RD strategy for the leave-one-out sample (a) and its QQ-plot (b).

5. Discussion and Conclusion

In this work, we have analyzed the robustness of 48 real-world social networks with the node number ranging over five orders of magnitude, from 1,914 to 265,216. Using Monte Carlo simulations, we have run two commonly used node attack strategies, IB and RD strategies, whose computation time is within our hardware capability. We found that their corresponding simulation time, t_{IB} and t_{RD}, scales linearly with the product of the network's node number and edge number, i.e., $N \times E$. We also found that the two attack strategies IB and RD present similar efficacy when evaluated by the unique robustness metric R, with RD slightly better than IB (average R_{RD} is slightly smaller than average R_{IB}). It suggests that for the social networks used in this study, the RD strategy is the most efficient strategy to dismantle (breakdown) networks, both in terms of computational cost and breakdown efficiency.

To understand how the structure of a social network determines its robustness, we investigate the relationship between the metric R and a set of network structural indicators (NSIs) from the literature. The simple linear regression (SLR) between R and NSIs shows low goodness of fitting, and it is overall not able to produce significant prediction models. The low goodness of SLR would indicate that network robustness depends on NSIs in a nonlinear manner.

To improve fitting, we have developed two machine learning models to predict two robustness metrics R_{RD} and R_{RD} from the combination of 8 NSIs, multiple linear regression (MLR), and random forest (RF) model. The latter one is chosen as it can handle nonlinear data well and is built on a collection of base models, decision tree classifiers. We found clearly that the random forest model can predict network robustness better than the multiple linear regression model. In concrete, the RF model predicts network robustness with an RMSE of 0.0272 and 0.0241 for R_{IB} and R_{RD}, respectively. This result is encouraging to predict real-world social network robustness, although the error is about 16% (for R_{IB}, the RMSE is 0.0272 compared to an average R_{IB} of 0.173, and for R_{RD}, the RMSE is of 0.0241 compared to an average R_{RD} of 0.148). Meanwhile, when the leave-one-out evaluation is applied, the RMSE increases to 0.0733 and 0.0636 for R_{IB} and R_{RD}, respectively, which is about one-third of the average value.

Finally, MLR indicates that the most important factors to predict RIB are the exponent α and the average node degree $<k>$, for both R_{IB} and R_{RD}. In particular, a higher value of α is correlated with higher R_{IB} and R_{RD}. Higher absolute values of the exponent α denote a network with fewer hub nodes (highly connected nodes) [35]. In consequence, the RD and IB attack strategies cannot find large hub nodes whose removal may disintegrate the network faster, resulting in higher values of R_{RD} and R_{IB}. Additionally, MLR indicates that $<k>$ is positively related to lower R_{IB} and R_{IB}. This last outcome agrees with previous results, demonstrating that networks with higher edge density may be more resistant to the attack [4, 52]. On the

other hand, it confirms that SLR, which focuses on a single NSI, may not be able to predict the robustness of real-world social networks.

Our work demonstrates that the ML model can be used to predict network robustness with acceptable results. Therefore, it alleviates the need to run a full Monte Carlo simulation on a network when only approximate robustness is needed. Meanwhile, more network datasets are expected to improve the accuracy of ML models. This work also contributes to the understanding of the relationship between real-world social network robustness and its structural indicators. Finally, we have proved that using a data-driven approach to predict the outcome of the nonlinear and complex dynamic process, such as network robustness, is an appropriate approach [54–60].

Abbreviations

RD: Recalculated degree node attack strategy
IB: Initial betweenness node attack strategy
RB: Recalculated betweenness node attack strategy
t_{IB}: Total simulation time for the attack strategy IB
t_{RD}: Total simulation time for the attack strategy RD
t_{RB}: Total simulation time for the attack strategy RB
SLR: Simple linear regression model
MLR: Multiple linear regression model
RF: The random forest model
DTR: Decision tree regression model
NSI: Network structural indicator
RMSE: Mean squared error
R^2: Coefficient of determination (also named the explained variance ratio)
a_0: Intercept coefficient of SLR
a_1: Slope coefficient of SLR
OLS: Ordinary least square method
ε: Error between the empirical (simulated) network robustness and the predicted value of robustness
α: Fitted scale-free exponent
k: Node degree
$<k>$: Average node degree
a: Degree assortativity
Cl: Global closeness
C: Global clustering coefficient
LCC: Largest connected component
N: Number of nodes
E: Number of edges
Q: Modularity indicator
α^2: Fitting variance of α
q: Accumulated proportion of nodes removed
R: Network robustness
R_{RD}: Network robustness against RD node attack strategies
R_{IB}: Network robustness against IB node attack strategies.

Appendix

The histogram and the QQ-plot of residual errors of all regressions are given in Figures 5–8 and Figures 10–13.

Data Availability

All the 48 real-world social networks are downloaded from the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/) and the Network Repository social networks (https://networkrepository.com/soc.php).

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

QN conceived analyses. NKKN, HHP, TTL, and TMN performed simulations. QN, NKKN, FS, RA, MB, and DC wrote the paper.

Acknowledgments

This work was supported by Vietnam's Ministry of Science and Technology (MOST) under the Vietnam-Italy Scientific and Technological Cooperation Program for the period of 2021–2023 and by the Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh city, Vietnam under grant nos. B2017-42-01. This research was funded by a grant from the Italian Ministry of Foreign Affairs and International Cooperation. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. (816313)). The authors are greatly thankful to Van Lang University, Vietnam, for providing the budget for this study.

References

[1] S. Lehmann and Y.-Y. Ahn, "Complex spreading phenomena in social Systems," *Computational Social Sciences)*, Springer International Publishing, Berlin, Germany, 2018.
[2] M. Bellingeri, D. Cassi, and S. Vincenzi, "Efficiency of attack strategies on complex model and real-world networks," *Physica A: Statistical Mechanics and its Applications*, vol. 414, pp. 174–180, 2014.
[3] M. Bellingeri, D. Bevacqua, F. Scotognella et al., "Link and node removal in real social networks: a Review," *Frontiers in Physiology*, vol. 8, p. 228, 2020a.
[4] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, "Attack robustness and centrality of complex networks," *PLoS One*, vol. 8, no. 4, Article ID e59613, 2013.
[5] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.
[6] K. Nguyen and Q. Nguyen, "Resilience of stock cross-correlation network to random breakdown and intentional attack," *Studies in Computational Intelligence*, vol. 760, pp. 553–561, 2018.
[7] S. Wandelt, X. Sun, D. Feng, M. Zanin, and S. Havlin, "A comparative analysis of approaches to network-dismantling," *Scientific Reports*, vol. 8, no. 1, Article ID 13513, 2018.
[8] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of Modern Physics*, vol. 87, no. 3, pp. 925–979, 2015.

[9] C. Stegehuis, R. van der Hofstad, and J. S. H. van Leeuwaarden, "Epidemic spreading on complex networks with community structures," *Scientific Reports*, vol. 6, no. 1, Article ID 29748, 2016.

[10] C. Li, L. Wang, S. Sun, and C. Xia, "Identification of influential spreaders based on classified neighbors in real-world complex networks," *Applied Mathematics and Computation*, vol. 320pp. 512–523, C, 2018.

[11] J. Wang, C. Li, and C. Xia, "Improved centrality indicators to characterize the nodal spreading capability in complex networks," *Applied Mathematics and Computation*, vol. 334, pp. 388–400, 2018.

[12] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Scientific Reports*, vol. 3, no. 1, p. 2522, 2013.

[13] P. Holme, "Efficient Local strategies for vaccination and network attack," *Europhysics Letters*, vol. 68, no. 6, pp. 908–914, 2004.

[14] L. K. Gallos, F. Liljeros, P. Argyrakis, A. Bunde, and S. Havlin, "Improving immunization strategies," *Physical Review E*, vol. 75 pp. 1–4, 2007.

[15] J. Hadidjojo and S. A. Cheong, "Equal graph partitioning on estimated infection network as an effective epidemic mitigation measure," *PLoS One*, vol. 6, no. 7, Article ID e22124, 2011.

[16] M. A. Amaral, M. Md Oliveira, and M. A. Javarone, "An spidemiological model with Voluntary Quarantine strategies Governed by evolutionary vame dynamics," *Chaos, Solitons & Fractals*, vol. 143, Article ID 110616, 2021.

[17] H. Amini and A. Minca, "Epidemic spreading and qquilibrium social distancing in geterogeneous networks," *Dyn Games Appl*, vol. 12, no. 1, pp. 258–287, 2022.

[18] M. Bellingeri, D. Bevacqua, F. Scotognella, R. Alfieri, and D. Cassi, "A comparative analysis of link removal strategies in real complex weighted networks," *Scientific Reports*, vol. 10, pp. 3911–3915, 2020b.

[19] M. Bellingeri, M. Turchetto, D. Bevacqua et al., "Modeling the consequences of social distancing over epidemics spreading in complex social networks: from link removal analysis to SARS-CoV- 2 Prevention," *Frontiers in Physiology*, vol. 9, Article ID 681343, 2021.

[20] D. Achlioptas, R. M. D'souza, and J. Spencer, "Explosive percolation in random networks," *Science*, vol. 323, no. 5920, pp. 1453–1455, 2009.

[21] G. Dong, J. Fan, L. M. Shekhtman et al., "Resilience of networks with community structure behaves as if under an external field," *Proceedings of the National Academy of Sciences*, vol. 115, no. 27, pp. 6911–6915, 2018.

[22] O. Riordan and L. Warnke, "Explosive percolation is continuous," *Science*, vol. 333, no. 6040, pp. 322–324, 2011.

[23] Y. Sun, C. Liu, C.-X. Zhang, and Z.-K. Zhang, "Epidemic spreading on weighted complex networks," *Physics Letters A*, vol. 378, no. 7-8, pp. 635–640, 2014.

[24] A. Majdandzic, B. Podobnik, S. V. Buldyrev, D. Y. Kenett, S. Havlin, and H. Eugene Stanley, "Spontaneous recovery in dynamical networks," *Nature Physics*, vol. 10, no. 1, pp. 34–38, 2014.

[25] S. Wandelt, X. Shi, X. Sun, and M. Zanin, "Community detection boosts network dismantling on real-world networks," *IEEE Access*, vol. 8, pp. 111954–111965, 2020.

[26] R. Albert, H. Jeong, and A.-L. Barabasi, "Diameter of the world-wide web," *Nature*, vol. 401, pp. 130-131, 1999.

[27] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," *Physical Review Letters*, vol. 85, no. 21, pp. 4626–4628, 2000.

[28] Q. Nguyen, T. V. Vu, H. D. Dinh et al., "Modularity affects the robustness of scale-free model and real-world social networks under betweenness and degree-based node attack," *Appl Netw Sci*, vol. 6, no. 1, p. 82, 2021b.

[29] X. Sun, V. Gollnick, and S. Wandelt, "Robustness analysis metrics for worldwide airport network: a comprehensive study," *Chinese Journal of Aeronautics*, vol. 30, no. 2, pp. 500–512, 2017.

[30] Q. Nguyen, N. K. K. Nguyen, D. Cassi, and M. Bellingeri, "New betweenness centrality node attack strategies for real-world complex weighted networks," *Complexity*, vol. 2021, pp. 1–17, Article ID 1677445, 2021a.

[31] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, p. 35, 1977.

[32] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge UK, 1994.

[33] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[34] Q. Nguyen and T. Trang Le, "Structure and robustness of Facebook's pages networks," in *Proceedings of the 2019 the 10th Conference on Network Modeling and Analysis (Marami 2019)*, Dijon, France, November 2019.

[35] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[36] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge UK, 2014.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, NV, USA, June 2016.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, J. M. ., N. S. Leibe and M. Welling, Eds., vol. 9908B, pp. 630–645, Springer International Publishing, Cham, 2016.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[40] M. E. J. Newman, "Spread of epidemic Disease on networks," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 66, no. 1, Article ID 016128, 2002.

[41] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[42] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and Fragility: percolation on random graphs," *Physical Review Letters*, vol. 85, no. 25, pp. 5468–5471, 2000.

[43] S. Wasserman and K. Faust, *Social Network Analysis: A Handbook*, Sage, CA, USA, 2000.

[44] M. Bellingeri, D. Bevacqua, F. Scotognella, and D. Cassi, "The heterogeneity in link weights may decrease the robustness of real-world complex weighted networks," *Scientific Reports*, vol. 9, no. 1, Article ID 10692, 2019.

[45] A. S. Goldberger, "Classical linear regression," *Econometric Theory*, p. 158, John Wiley & Sons, NY, USA, 1964.

[46] F. Hayashi, *Econometrics*, p. 15, Princeton University Press, NJ, USA, 2000.

[47] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, CRC Press, Boca Raton, Florida, 1984.
[48] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan kaufmann, Burlington, Massachusetts, USA, 2006.
[49] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001a.
[50] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.
[51] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, NY, USA, 2013.
[52] R. Albert and A.-L. Baraba´si, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.
[53] R. H. Jones and B. A. Molitoris, "A statistical method for determining the breakpoint of two lines," *Analytical Biochemistry*, vol. 141, no. 1, pp. 287–290, 1984 Aug 15.
[54] R. Albert, H. Jeong, and A.-L. Baraba´si, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.
[55] A. Bavelas, "Communication gatterns in task-oriented groups," *Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.
[56] M. Bellingeri and D. Cassi, "Robustness of weighted networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 489, pp. 47–55, 2018.
[57] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Physical Review A*, vol. 65, no. 5, Article ID 056109, 2002.
[58] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, Article ID 208701, 28 October 2002.
[59] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.
[60] Y. Sun, C. Liu, C.-X. Zhang, and Z.-K. Zhang, "Epidemic spreading on weighted complex networks," *Physics Letters A*, vol. 378, no. 7-8, pp. 635–640, 2014.

Hindawi
Complexity
Volume 2022, Article ID 2132005, 20 pages
https://doi.org/10.1155/2022/2132005

Research Article

An Optimized Design of New $XY\theta$ Mobile Positioning Microrobotic Platform for Polishing Robot Application Using Artificial Neural Network and Teaching-Learning Based Optimization

Minh Phung Dang,[1] Hieu Giang Le,[1] Ngoc Le Chau,[2] and Thanh-Phong Dao [3,4]

[1]*Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam*
[2]*Faculty of Mechanical Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam*
[3]*Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[4]*Faculty of Electrical & Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam*

Correspondence should be addressed to Thanh-Phong Dao; daothanhphong@tdtu.edu.vn

Received 13 August 2022; Accepted 6 October 2022; Published 7 November 2022

Academic Editor: Gonzalo Farias

Compliant mechanisms with flexure hinges have been widely applied for positioners, bioengineering, and aerospace. In this study, a new optimized design method for the mobile microrobotic platform was developed for the polishing robot system. A metaheuristic-based machine learning technique in combination with finite element analysis (FEA) was developed. The designed platform allows three degrees of freedom with two x-and-y translations and one z-axis rotation. A new hybrid displacement amplification mechanism was also developed using Scott-Russell and two-lever mechanisms to magnify the workspace of the platform. The leaf hinges were employed due to their large rotation, and the right circular hinges were adopted because of their high accuracy. In modeling the behaviors of the developed platform, the artificial neural network is formulated in combination with the teaching-learning-based optimization (TLBO) method. The ANN architecture was optimized through TLBO to a better approximation. And then, three optimized case studies were conducted by the TLBO. The data is collected through FEA simulation. The modeling results from the TLBO-based ANN were well established with excellent metrics of R, R^2, and MSE. The optimized results found that the proposed MPM platform achieves a max-y stroke of 1568.1 μm, max-x stroke of 735.55 μm, and max-θ rotation angle of 2.26 degrees. The proposed MPM platform can operate at a high displacement amplification ratio of over 9.

1. Introduction

Compliant mechanisms play a vital role in ultrahigh precision engineering, such as stable switch [1, 2], vibration-assisted cutting [3], manipulations/microgrippers [4], fast servo in precision machining, energy harvester [5], alignment of optics [6], robotics [7], and so on. Compared with rigid-link counterparts, compliant mechanisms can propose a high resolution with precise smooth motion due to the excellent advantages such as without backlash, no friction, reduced assembly, cheap manufacture, and monolithic structure.

Currently, many planar compliant mechanisms from one degree of freedom (DOF) to three-DOF motions have been developed by using series architecture, parallel chain, or hybrid series-parallel type. The one DOF mechanisms often have a high accuracy with a minimal parasitic motion, but these mechanisms have still limited in some applications, e.g., positioners [8]. Then, two DOF mechanisms have been designed to propose more complicated applications, i.e.,

scanner [9]. The two DOF mechanisms possess a decoupled property. Although one or two DOF mechanisms can achieve a wide stroke, simple control, and high accuracy but their applications are still limited. Therefore, three DOF mechanisms have been developed as alternatives for many planar applications as a positioner, manipulation, and so forth [10]. However, the workspace of two translations and one rotation of the existing three DOF mechanisms are still small. To overcome such drawbacks, a kinematic structure with better properties is needed to provide a high load capacity, large stroke, high safety factor, and high stiffness. Hence, three DOF mechanisms have attracted much attention and become a hot topic for researchers.

Generally speaking, complaint mechanisms, which are acted by piezoelectrical actuators (PZT), have limited workspace. To overcome this drawback, many displacement amplification mechanisms were proposed to amplify the stroke of PZTs, such as Scott–Russell mechanism, lever and bridge types [11]. In addition, a lot of other researchers have also designed many different types of three-DOF compliant positioning platforms with desired characteristics. A micropositioning stage with 3-DOF was designed [12]. In this study, the compliance matrix and finite element method were utilized to build the stiffness and the input coupling ratio of the stage. Besides, the parameters of the stage were optimized to minimize the input coupling ratio. A 3-DOF spatial precision manipulation was designed and analyzed [13]. The translational and angular displacements were analyzed in this article. Besides, a 3-DOF translational mechanism was proposed, and it was analyzed via the pseudo-rigid-body model (PRB) method [14]. By using the PRB technique, another 3-DOF mechanism with two translations and one rotation was designed and analyzed [15]. This type for nanopositioning application was analyzed by a compliance matrix [16].

Although the discussed 3-DOF stages have been designed with multiple excellent characteristics, but the structure is still complicated. Moreover, the workspaces are still limited. Considering an application of 3-DOF compliant mechanisms in the robots, a planar micropositioning platform was designed, and the manufacturing error was analyzed [17]. Almost the behavior analysis of the previous stages employed some popular analytical techniques, such as PRB and compliance matrix. With high nonlinear characteristic behavior, modeling of them has a large error. This causes a large manufacturing error, decreasing the practical positioning ability. To overcome this obstacle, a new approach based on machine-learning-based methods and metaheuristics is devoted in the present article. The artificial neural network (ANN) is combined with the teaching-learning-based optimization algorithm (TLBO) in modeling the behaviors of a new $XY\theta$ mobile positioning microrobotic platform. The developed microrobotic platform can basically be applied for vibration-based polishing robot applications.

Motivated by the gaps between the existing studies, this paper presents an optimized design method for a three-DOF mobile microrobotic platform for use in polishing robot application. The developed platform is able to provide a large workspace in the x-and-y translations and rotation around the z-axis. In modeling the behaviors of the proposed microrobotic platform, artificial neural network is adopted to resolve the stroke and safety factor. To overcome the ANN limitations, the TLBO algorithm was extended to optimize the ANN approximate accuracy. Then, the geometrical factors of the proposed microrobotic platform were optimized by adopting the TLBO algorithm. Finally, three case studies are considered to confirm the accuracy and effectiveness of the proposed methodology.

2. Conceptual Design of $XY\theta$ Mobile Positioning Microrobotic Platform

A basic application of the $XY\theta$ mobile positioning microrobotic (MPM) platform is used for manipulations and precise sample positioning from sub-micrometer to hundreds of micrometer scales. Figure 1 illustrates a design scheme of the MPM platform. The proposed MPM platform utilizes three piezoelectric stack actuators (PZT) to actuate an input displacement to three corresponding robotic legs (robotic leg #1, robotic leg #2, and robotic leg #3).

By arranging three robotic legs around a circle with 120 degrees and three PZTs located in a tripedalism, so-called tripedal topology, the MPM platform can generate a locomotion in three DOF on a planar surface. It means that the platform includes three main motions, such as two translations along the x-and-y axes and one rotation (θ_z) around the z-axis.

Overall, the MPM platform was manufactured with a monolithic flexure-based mechanism. The fabrication will be carried out via wire electrical discharged machining (WEDM). Each robotic leg was also a flexure structure that consists of a hybrid displacement amplification mechanism (HDAM) in combination with a leaf hinge. The robotic leg #1 was defined in a local coordinate of $O_1X_1Y_1$. The robotic leg #2 and the robotic leg #3 were defined in a local coordinate of $O_2X_2Y_2$ and $O_3X_3Y_3$, respectively. More details of the HDAM are presented in next section. Under actuating the three PZTs simultaneously, the mobile platform of the microrobot makes two translations δx_1 and δy_1, and a rotation θ_1.

Technical requirements and specifications of the MPM platform in the design phase are expected to achieve large strokes in the translations over 1000 (μm) or higher than 1 mm and a wide rotation. Furthermore, a high safety factor of over 1.8 is required. The mentioned importantly technical specifications of the MPM platform can fulfil the practical applications. In addition, Al 7075-T651 is chosen to manufacture the microrobotic platform. The properties of Al 7075-T651 are listed, including a density of 2810 kg/m^3, Poisson ratio of 0.33, yield stress of 503 MPa, and Young's modulus of 71.7 GPa.

Figure 2 illustrates the assembly scheme of $XY\theta$ mobile positioning microrobotic platform.

As shown in Figure 2, it includes the following key components:

(1) Preload crew,

(2) PZT mounting plate,

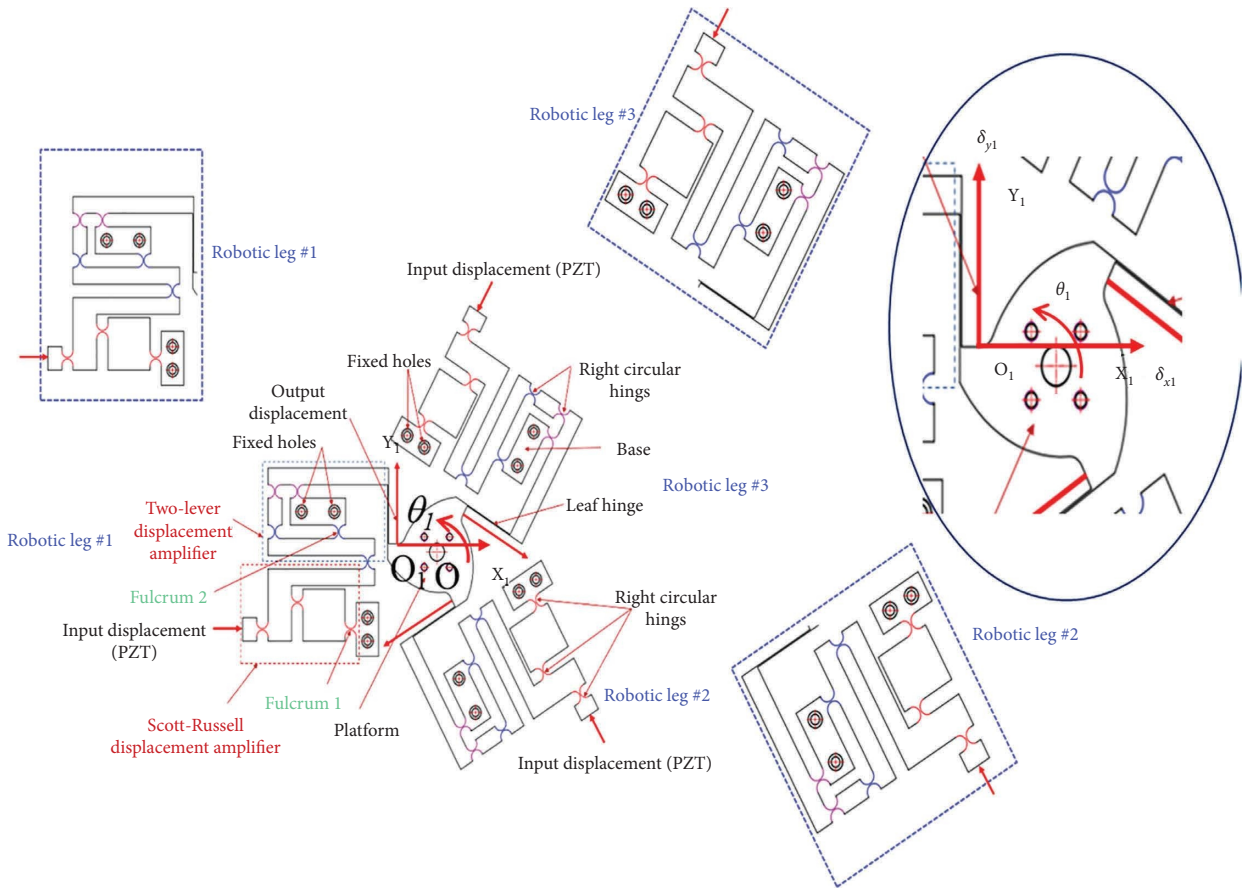


Figure 1: Design scheme of $XY\theta$ mobile positioning microrobotic platform.

(3) PZT actuator,

(4) Intermediate plate,

(5) Prototype,

(6) Anti-vibration fixing plate,

(7) Fixed hole.

As depicted in Figure 2, the prototype of the proposed microrobotic platform was mounted on the intermediate plate. The PZTs were fixed on the PZT mounting plate, and the preload screw was employed to adjust the PZT in contact with the input port of the platform. Finally, the whole of the system was put on the anti-vibration table.

A basic application of the proposed MPM platform is able to be employed for polishing robot system, as given in Figure 3. The proposed platform is mounted on the station. The polished sample is located on the mobile platform through fixing screws while the end-effector of the robotic arm brings the polishing tool.

When three PZTs act, the platform causes a micro-vibration for the sample. The micro-vibration is aimed to reduce the friction between the sample and the polishing tool. This leads to improvement of the surface roughness of the final workpiece. This machining process is considered as a vibration-assisted polishing process.

The dimensional scheme of the proposed MPM platform is provided in Figure 4, and the main dimensions are given in Table 1. The thickness of the platform in the out plane (z-axis) is 8 mm.

2.1. Analysis of Hybrid Displacement Amplification Module. Figure 5 provides a new hybrid displacement amplifier. The suggested HDAM is built by a combination of Scott–Russell mechanism (SRM) amplifier with a two-lever displacement (TLD) amplifier. The hybrid amplifier is moved based on the deformation of right circular hinges. In the beginning, an input displacement of 135 μm along the x-axis is acted to the SRM amplifier, and this displacement amplifier is rotated around the fulcrum (1) and then, the output motion of the SRM is transformed to the input port of the TLD amplifier, and this mechanism is rotated about the fulcrum (2) the output displacement is collected along the y-axis. Finally, the output motion of the proposed HDMA is kept to transfer to the leaf hinge (see Figure 1) so that the MPM platform is moved.

To illustrate the amplification ratio of the proposed HDMA, the proposed HDMA is meshed and simulated by finite element analysis (FEA) ANSYS 2019R1 software. The number of nodes and elements are about 29047 and 16867, respectively. The quality of the mesh is measured by the Skewness technique with an average value of 0.44906. The results of the HDMA are provided in Table 2.

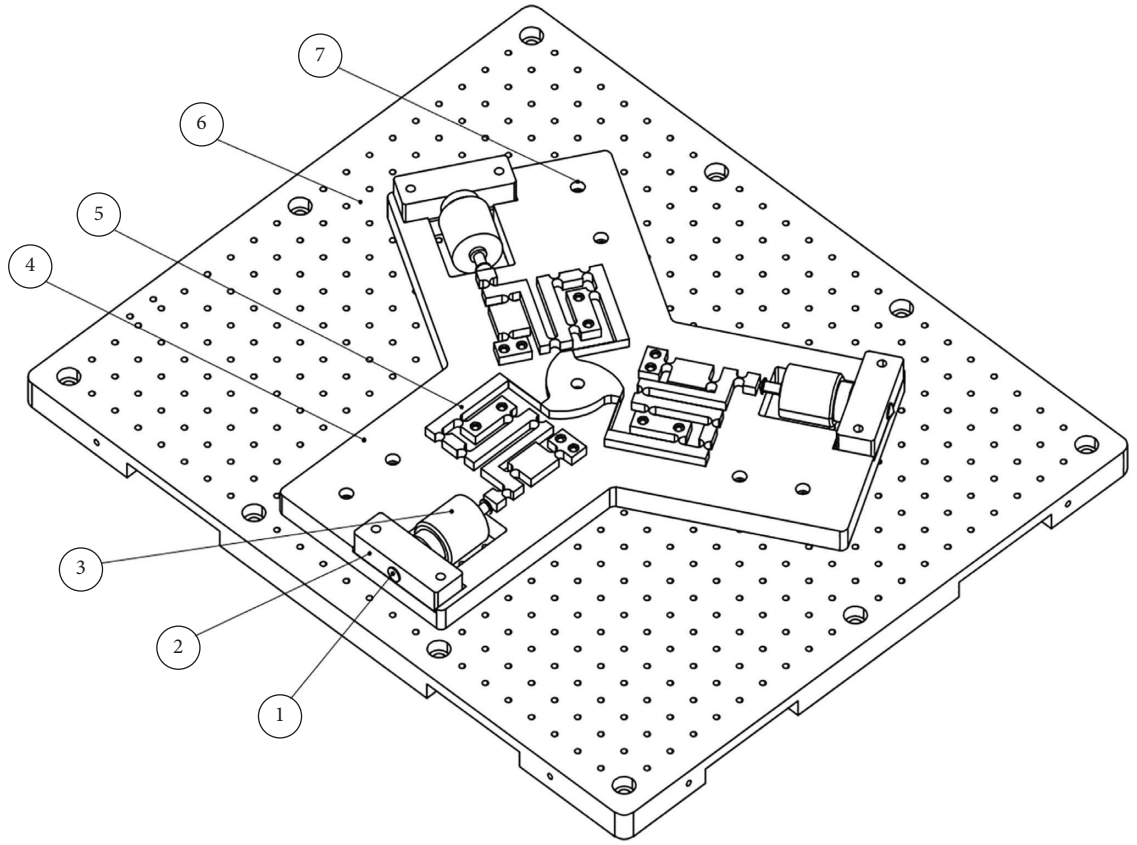


Figure 2: Assembly scheme of $XY\theta$ mobile positioning microrobotic platform: (1) preload crew, (2) PZT mounting plate, (3) PZT actuator, (4) intermediate plate, (5) prototype, (6) anti-vibration fixing plate, (7) fixed hole.

The results of Table 2 indicates that the amplification ratio of the proposed hybrid amplifier is about 12.43 with a high safety factor (SF) over 1.4 when the input displacement is from 90 μm to 145 μm. Besides, the stress is still lower than the yield stress of the material (503 MPa).

2.2. Initial Evaluation of Static and Dynamic Behavior of Microrobotic Platform. In order to evaluate the initial specifications of the proposed MPM platform, the static and dynamic behaviors are simulated by ANSYS software. The three PZTs are employed simultaneously with 135 μm, and the output stroke/displacement of the robotic leg #1 is measured. Figure 6(a) shows the boundary conditions for simulating the platform. The number of nodes is 71202, and the number of elements is 41045. Skewness average value is about 0.4877, as given in Figure 6(b).

Figure 7 depicts the stress concentration. It is found that the high stress appeared on the surfaces of leaf hinges and right circular hinge.

The deformation of the MPM platform is provided in Figure 8.

The initial evaluation showed that the amplification ratio of the proposed MPM platform is about 9.85, with a high safety factor (SF) over 1.7 when the input displacement is from 90 μm to 145 μm. Besides, the stress is still smaller than the yield stress of material (503 MPa), as depicted in Table 3.

The dynamic behavior is achieved by FEA simulations. The four natural frequencies for the first mode shapes include 102.036 Hz, 113.81 Hz, 113.9 Hz, and 154.84 Hz, respectively, as provided in Table 4. Considering a resonance of the proposed MPM platform with the PZTs and others, the first mode shape is a z-axis translation. The second mode shape is the x-axis translation. The third mode shape is the z-axis translation. Finally, the fourth mode shape is the z-axis rotation.

2.3. Formulation of Optimization Problems. The characteristics of the proposed MPM platform are desirable to gain the two main design targets, including a large stroke (δy_1) and a high safety factor.

When the stroke is enhanced, the rotation of the platform (θ_1) is also improved. A good SF over 1.8 can ensure a long working time. Based on the initial evaluations in the previous parts, it determined that the performances of the proposed MPM platform are strongly affected by varying the thickness values of right circular hinges (A, B, C, D) and the thickness of the leaf hinges (E).

Three optimization problems of the proposed MPM platform are considered as follows.

Case #1.: maximize the stroke

Find design variables: $\mathbf{X} = [A, B, C, D, E]$

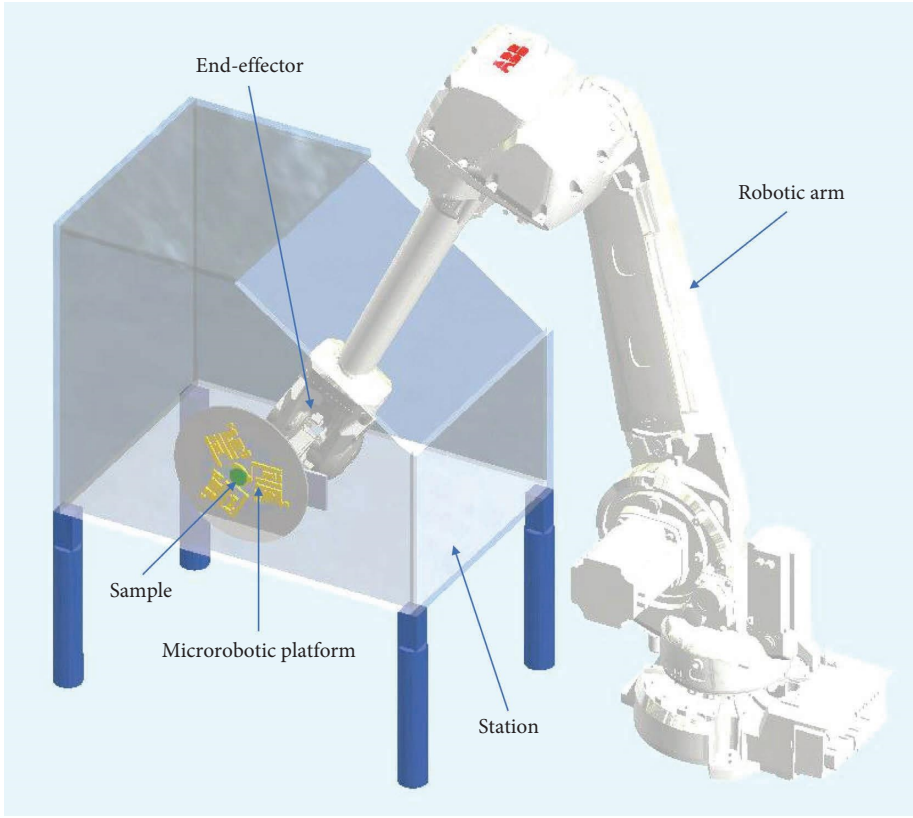


Figure 3: Application of microrobotic platform for polishing.

$$\text{Maximize} : f_1(\mathbf{X}). \tag{1}$$

Bounds of design variables (unit: mm):

$$\begin{cases} 0.8 \le A \le 0.9 \\ 0.7 \le B \le 0.8 \\ 0.6 \le C \le 0.7 \\ 0.55 \le D \le 0.6 \\ 45 \le E \le 50 \end{cases}. \tag{2}$$

Case #2.: maximize the safety factor

Find design variables: $\mathbf{X} = [A, B, C, D, E]$

$$\text{Maximize} : f_2(\mathbf{X}). \tag{3}$$

Bounds of design variables (unit: mm):

$$\begin{cases} 0.8 \le A \le 0.9 \\ 0.7 \le B \le 0.8 \\ 0.6 \le C \le 0.7 \\ 0.55 \le D \le 0.6 \\ 45 \le E \le 50 \end{cases}. \tag{4}$$

Case #3.: maximize the stroke and the safety factor simultaneously (multi-objective optimization problem)

Find design variables: $\mathbf{x} = [A, B, C, D, E]$

$$\begin{cases} \text{Maximize} : f_1(\mathbf{X}) \\ \text{Maximize} : f_2(\mathbf{X}) \end{cases}. \tag{5}$$

Bounds of design variables (unit: mm):

$$\begin{cases} 0.8 \le A \le 0.9 \\ 0.7 \le B \le 0.8 \\ 0.6 \le C \le 0.7 \\ 0.55 \le D \le 0.6 \\ 45 \le E \le 50 \end{cases}, \tag{6}$$

where X is a vector of design variables. Parameters A, B, C, and D are the thickness of right circular hinges. Parameter E is the thickness of leaf hinges. The stroke and safety factor are represented as f_1(X) and f_2(X), respectively.

3. Proposed Modeling and Optimization Method

As designed in Figure 1, the proposed MPM platform is a monolithic architecture with three robotic legs. The translations and rotation motions of the platform are totally based on the elastic motions of the leaf hinges and right circular hinges.

Because the MPM platform is built using the concept of flexure-based mechanism, so-called compliant mechanism,

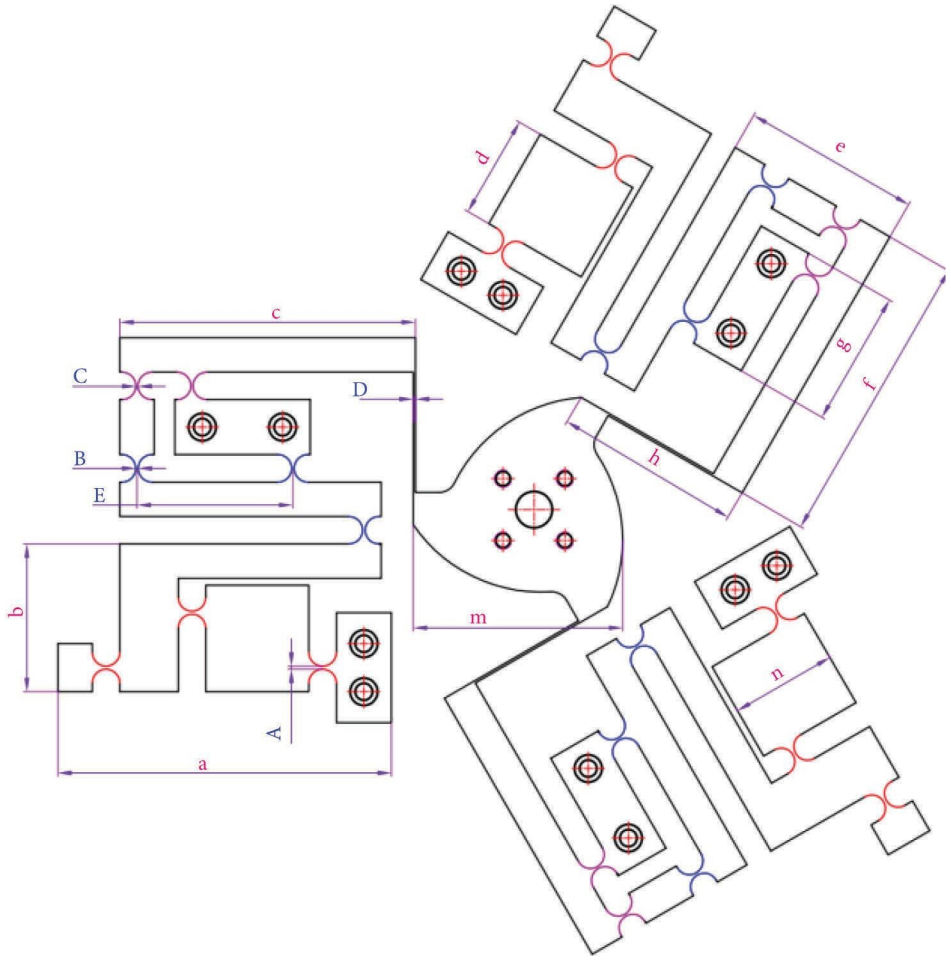


Figure 4: Mechanical scheme of proposed $XY\theta$ monolithic mechanism: (a) $XY\theta$ stage, (b) parameters.

Table 1: Dimensions of the $XY\theta$ microrobotic platform (unit: mm).

Par.	Value	Par.	Value	Par.	Value	Unit
a	97	f	86	A	$0.8 \le A \le 0.9$	mm
b	43	g	40	B	$0.7 \le B \le 0.8$	mm
c	86	h	54	C	$0.6 \le C \le 0.7$	mm
d	30	m	60	D	$0.55 \le D \le 0.6$	mm
e	52	n	32	E	$45 \le E \le 50$	mm

it inherits many excellent properties such as low weight, reduced assemble, simple fabrication, and without kinematic joints in comparison with rigid-link counterparts. Nevertheless, mathematical equations in modeling of the static behaviors of the MPM platform is difficult to exactly formulate because it has not kinematic joints. Therefore, the leaf hinges and right circular hinges are treated as virtual joints.

As a result, a modeling method based on ANN is chosen in approximating the stroke and the safety factor. In order to enhance the prediction ability of the ANN, the TLBO algorithm is employed. And then, the TLBO is extended to handle the three optimization cases of the MPM platform. The flowchart of the proposed modeling and optimization techniques is provided in Figure 9.

3.1. Simulation Technique for Microrobotic Platform. In order to collect the data of the performances of the MPM platform, the FEA implements are carried out, as seen in Figure 10. With five design variables, twenty-seven experimental samples are made.

(i) Build 3D model of the proposed MMP platform.

(ii) Design variables (A, B, C, D, and E) and output performances (stroke and safety factor) are parameterized.

(iii) Define properties of material Al 7075-T651.

(iv) Determine boundary conditions and a load/input displacement from PZT.

(v) Simulate the MPM platform by finite element method (FEM).

(vi) Collect the data.

(vii) If the data sets are not satisfied, it will return to adjust the range of variables.

3.2. ANN Optimization by TLBO. In this study, feedforward-learning ANN technique is selected to formulate the modeling of stroke and safety factor for the proposed MPM platform. Basically, ANN is operated based on human brain

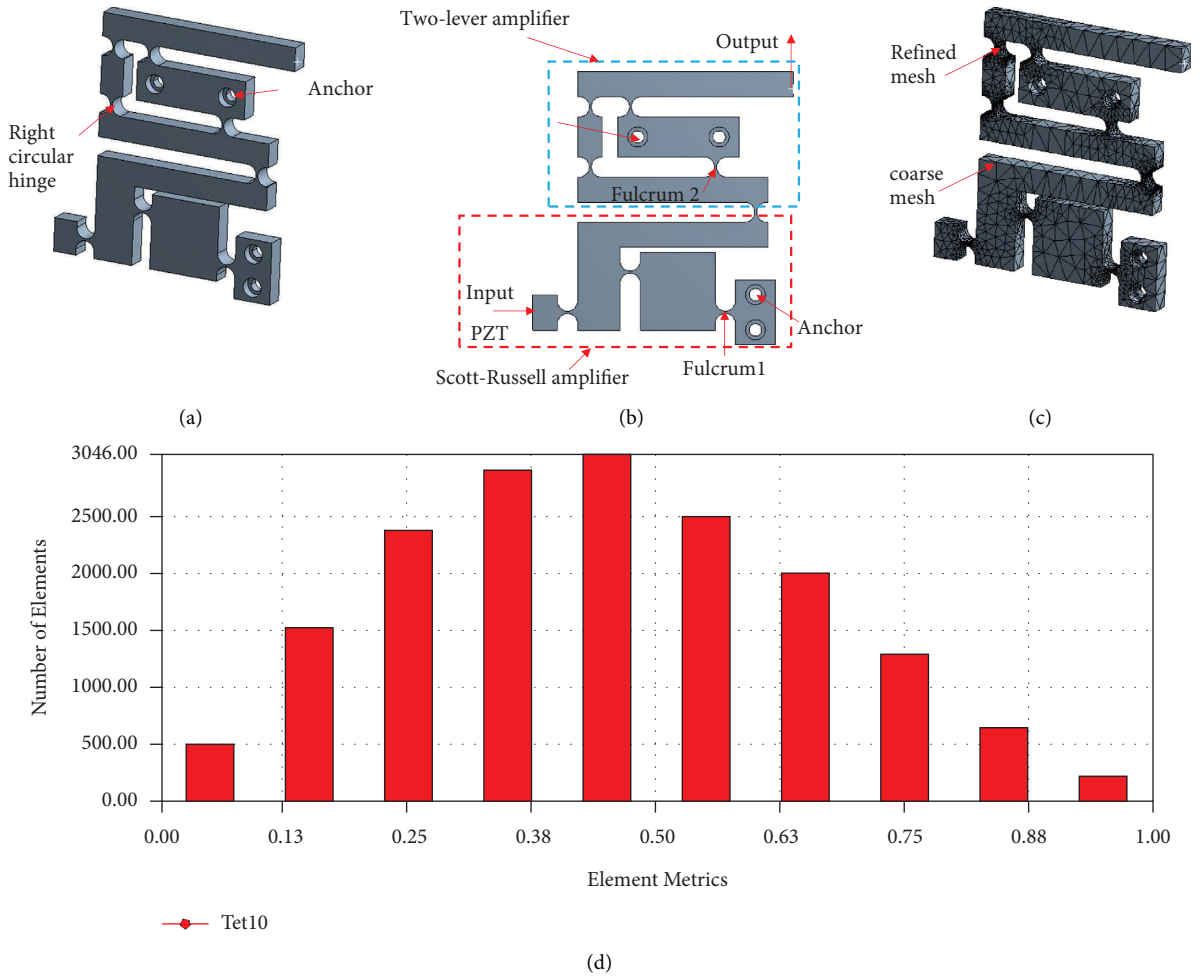


Figure 5: Proposed hybrid displacement amplifier: (a) 3D, (b) 2D, (c) meshing, (d) mesh quality.

Table 2: Results of amplification ratio of proposed HDMA.

Input (μm)	Output (μm)	Amplification ratio (μm/μm)	Stress (MPa)	Safety factor
90	1118.9	12.43	217.09	2.31
105	1305.4	12.43	253.27	1.98
125	1554.1	12.43	301.52	1.66
135	1678.4	12.43	325.64	1.54
145	1802.7	12.43	349.76	1.43

[18]. In the reasoning of ANN, the geometrical parameters and output responses of the MPM platform are embedded into the programming. An ANN programming includes three main signals such as input, hidden, and output layer. To effectively operate, the learning rate, momentum rate, bias, minimum error, and activation function should be appropriately defined. Operation of the ANN can gain a high effectiveness when it can ensure a minimal training error. This can be well done when the weight and bias are reasonably updated.

Although the ANN can build nonlinear behavior modeling but the accuracy is still strongly dependent on its controllable factors. To solve this limitation, the TLBO [19] is applied to optimize the ANN architecture. One of the most problems is how to define exactly the number of hidden nodes in hidden layer. The following equation is utilized to resolve this problem.

$$\text{Number of hidden nodes} = (2 * \text{inputs}) + \text{outputs}. \quad (7)$$

With five design variables corresponding to one output performance, the hidden layer is 11 nodes. An optimization of ANN by TLBO is provided in Figure 11.

In the optimization problem, the objective function is mean square error (MSE) which is defined as below:

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(t_i - \hat{t}_i)^2, \quad (8)$$

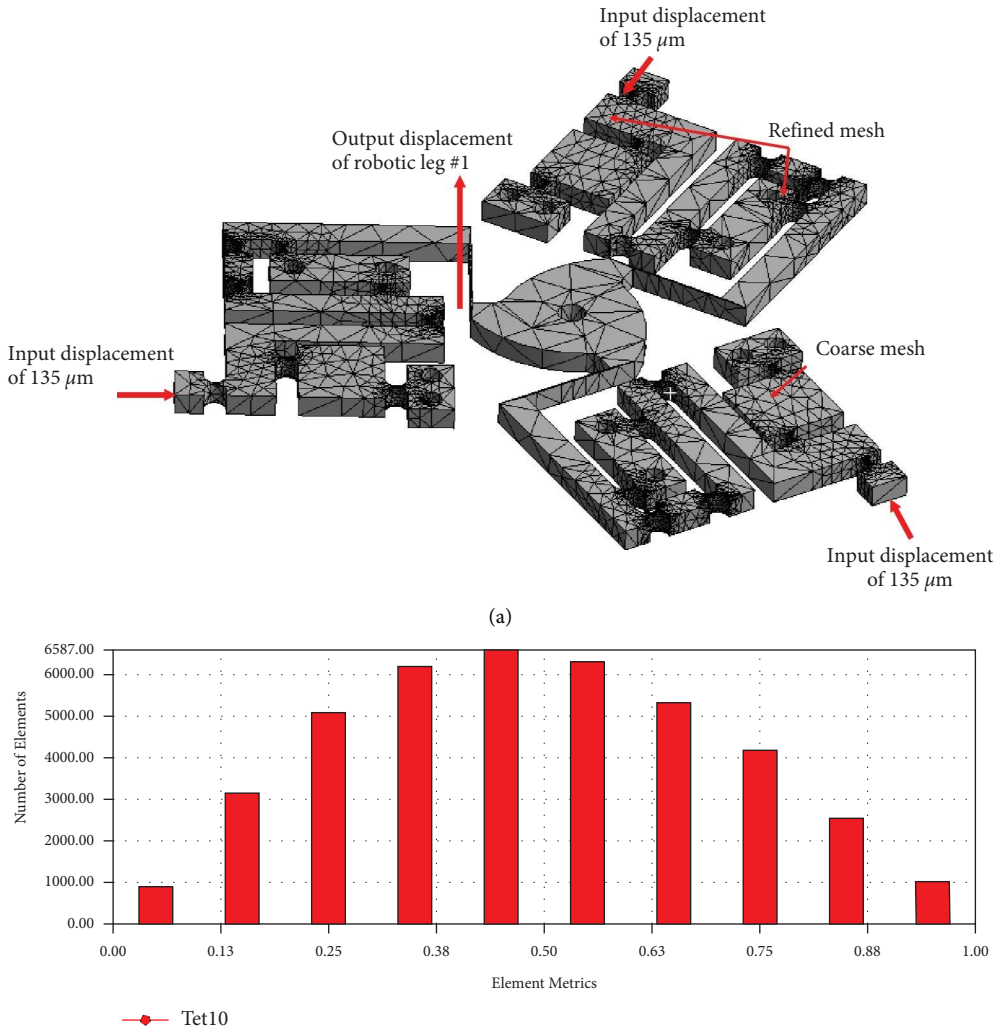


Figure 6: Simulation of the microrobotic platform: (a) boundary conditions, (b) mesh quality.

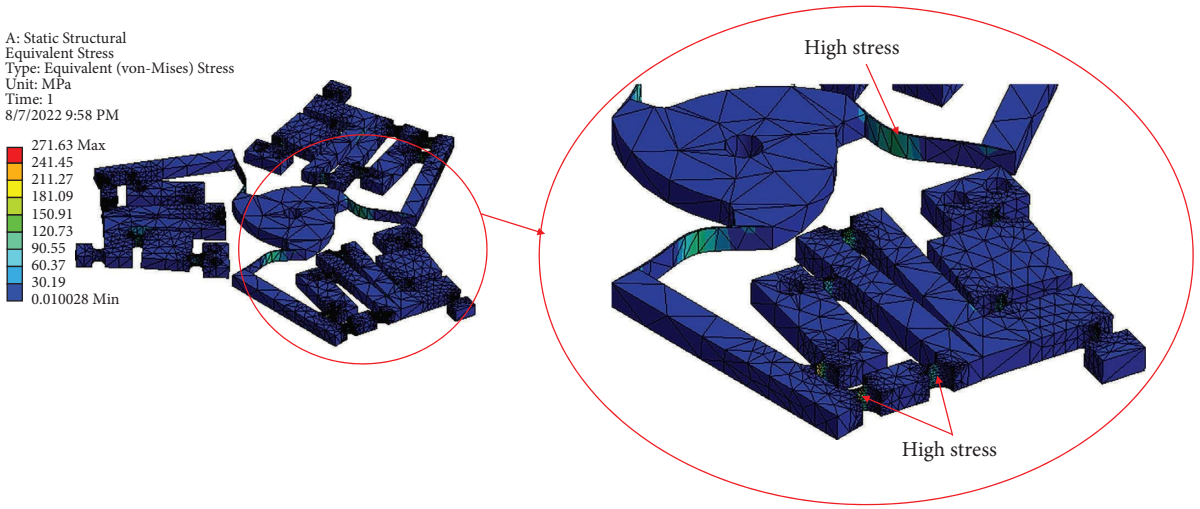


Figure 7: Stress concentration.

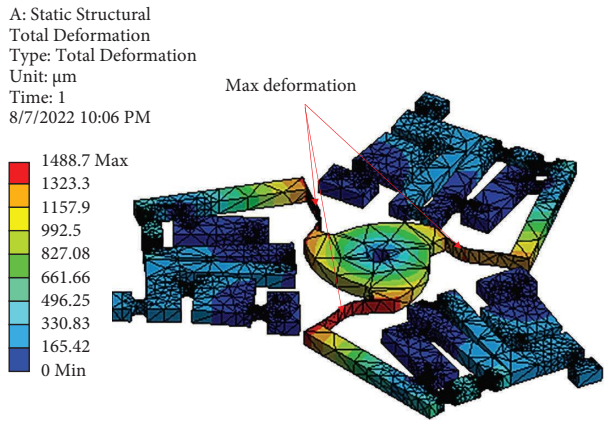


Figure 8: Deformation simulation.

Table 3: Results of static behavior.

Input (μm)	Output (μm)	Amplification ratio	Stress (MPa)	Safety factor
90	886.59	9.85	181.09	2.77
105	1034.4	9.85	211.27	2.38
125	1231.4	9.85	251.51	1.99
135	1329.9	9.85	271.63	1.85
145	1428.4	9.85	291.75	1.72

where, t is the measured target and $\hat{t}$ is the predicted target, and k is the dimension of inputs, so-called the number of data points.

Additionally, the coefficient of determination (R^2) is computed to estimate the regression model:

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{k}(t_i - \bar{t})(\hat{t}_i - \bar{t})}{\sqrt{\sum_{i=1}^{k}(t_i - \bar{t})^2 \sum_{i=1}^{k}(\hat{t}_i - \bar{t})^2}}, \tag{9}$$

where t is the actual target, $\hat{t}$ is the predicted target, and $\bar{t}$ is the average target.

3.3. Optimization of Microrobotic Platform by TLBO Method. According to the TLBO algorithm, a good teacher can train a better learner. The task of teachers in a classroom is critically important [19]. The leaner is a population where a vector of design is a course vector. The two main strategies of the TLBO include teaching and learning.

3.3.1. Teaching Strategy. The teacher strategy proposes some key ideals as follows.

(i) Search the teacher with best solution from the population.

(ii) Determine the mean results of learners ($M_{j,i}$) with respect to a specific subject.

(iii) The teacher's ability affects the quality of students by following equation.

$$Dm_{j,k,i} = r_{j,i}\left(X_{j,kbest,i} - T_F M_{j,i}\right). \tag{10}$$

where, $Dm_{j,k,i}$ is the increased mean value. $X_{j,kbest,i}$ is the best learner (i.e., teacher) in jth subject. T_F is the teaching factor. $r_{j,i}$ is a random value in [0, 1]. The T_F value is either 1 or 2. The T_F value is randomly determined by the following formula:

$$T_F = \mathrm{round}[1 + \mathrm{rand}(0,1)\{2-1\}]. \tag{11}$$

After that, the existing solution is updated by the following equation in the teacher strategy.

$$X'_{j,k,i} = X_{j,k,i} + Dm_{j,k,i}, \tag{12}$$

where, $X'_{j,k,i}$ is the updated value of $X_{j,k,i}$. If the results of this phase are satisfied, and then, they are considered as inputs for the learner strategy.

3.3.2. Learning Strategy. The learners can study somethings from other students in a classroom. At any iteration i, a learner is compared with the other learners. Specifically, U and V are two learners which are compared together ($X'_{U,i} \neq X'_{V,i}$) by following formula.

$$\begin{cases} X''_{j,U,i} = X'_{j,U,i} + r_{j,i}\left(X'_{j,U,i} - X'_{j,V,i}\right), \text{if } f\left(X'_{U,i}\right) < f\left(X'_{V,i}\right) \\ X''_{j,U,i} = X'_{j,U,i} + r_{j,i}\left(X'_{j,V,i} - X'_{j,U,i}\right), \text{if } f\left(X'_{V,i}\right) < f\left(X'_{U,i}\right) \end{cases}. \tag{13}$$

$X''_{j,U,i}$ is accepted when the value of objective function is better. Flowchart of the TLBO method is given in Figure 12.

4. Results and Discussion

In this part, modeling behaviors of the MPM platform is provided. Besides, the optimization problems of the proposed platform are performed. The optimized results are validated.

Table 4: Results of dynamic behavior with input displacement of 135 μm.

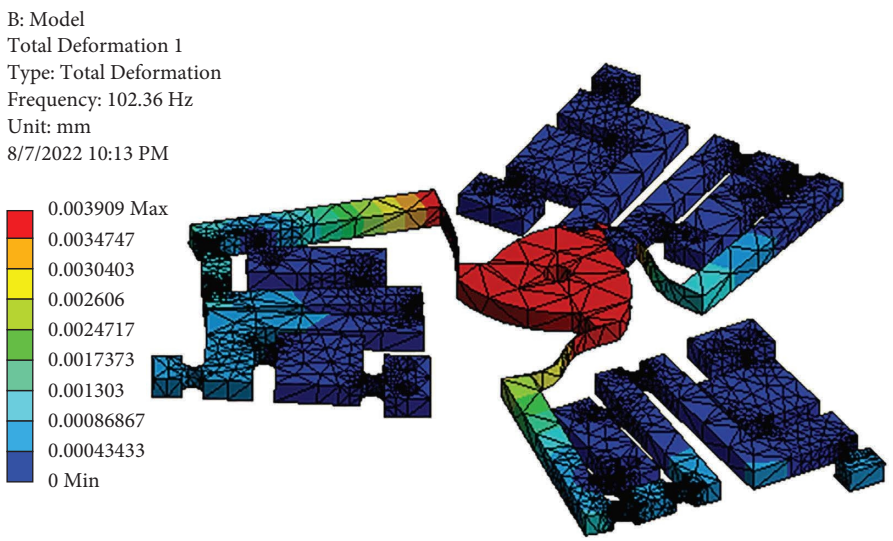
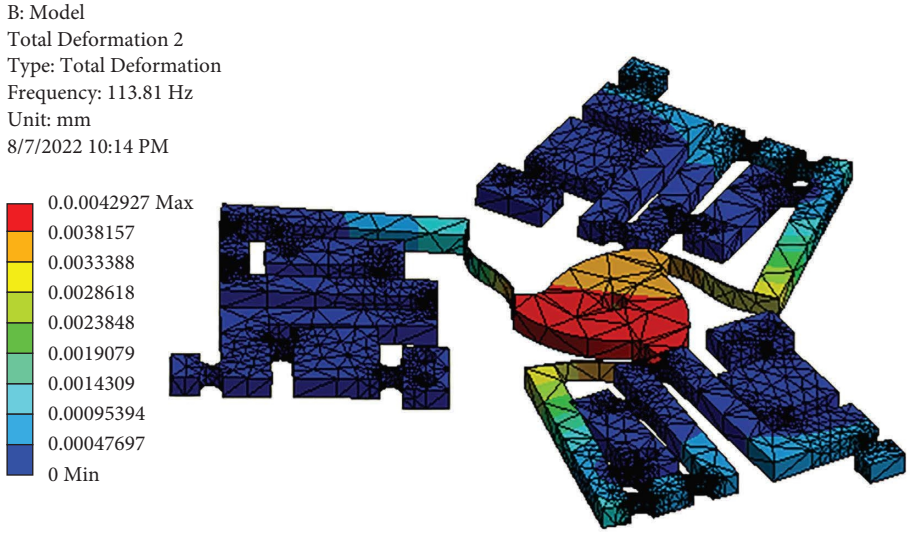
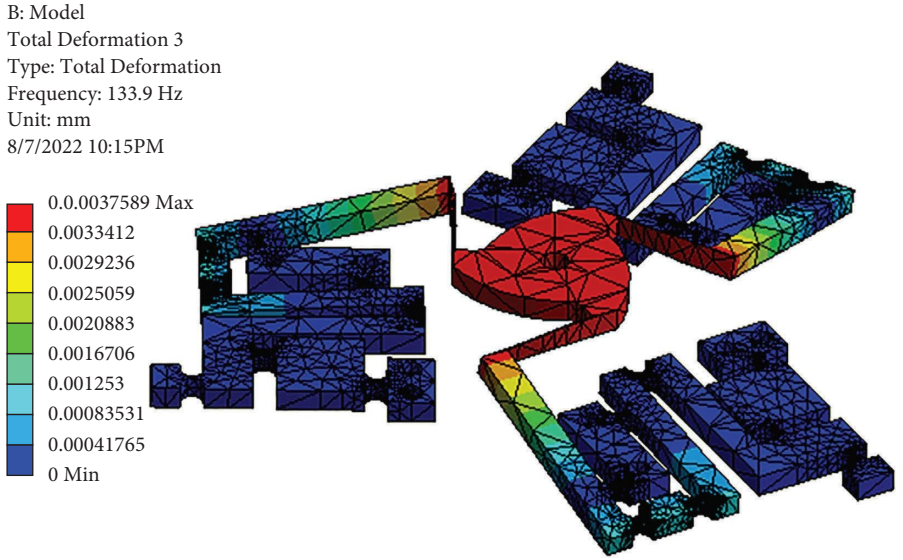
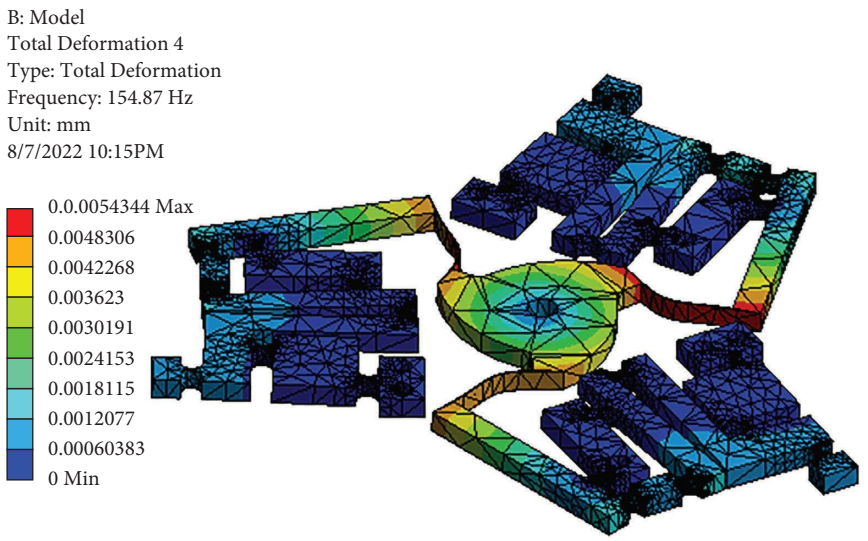
No.	Mode shape	Natural frequency (Hz)
(1) z-axis translation		102.36
(2) x-axis translation		113.81
(3) z-axis translation		113.9

Table 4: Continued.

No.	Mode shape	Natural frequency (Hz)
(4) z-axis rotation		154.84

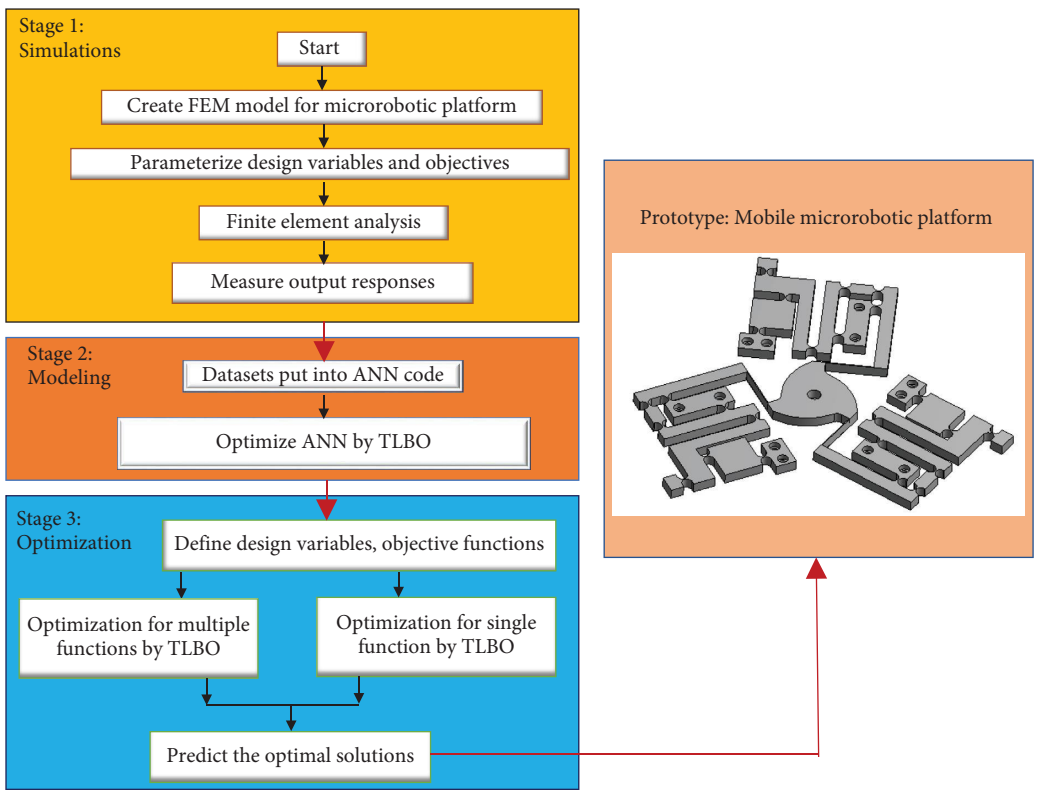


Figure 9: Flowchart of modeling and optimizing method for microrobotic platform.

4.1. Setup of Simulations and Data Collection. From Figure 5, the boundary conditions are seen. Three input displacements from three PZTs are acted simultaneously. The stroke (δ_{y1}) along the y-axis is measured. Besides, the safety factor is calculated. AL 7075-T651 is employed for the platform. The results of 27 experiments are given in Table 5.

4.2. Parametric Evaluation. To assess the associations of the geometrical parameters to the behaviors of the proposed MPM platform, analysis of variance (ANOVA) is adopted to solve this issue. The ANOVA results of stroke are given in Table 6. Moreover, the sensitive plot of whole inputs to the stroke is illustrated in Figure 13. The results indicated that

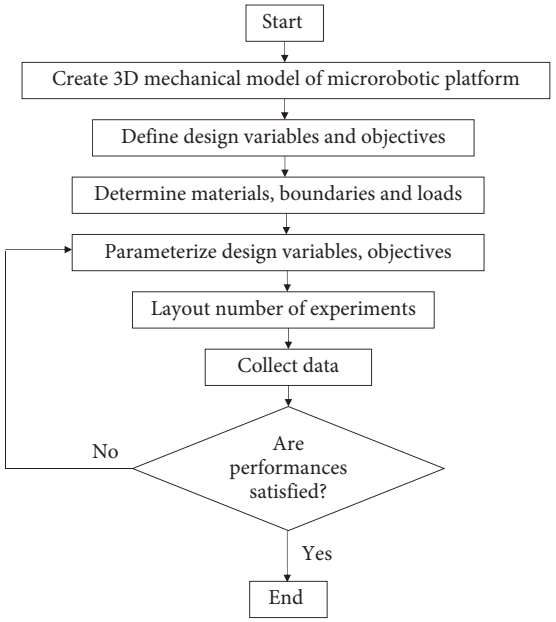


Figure 10: Proposed simulation scheme for microrobotic platform.

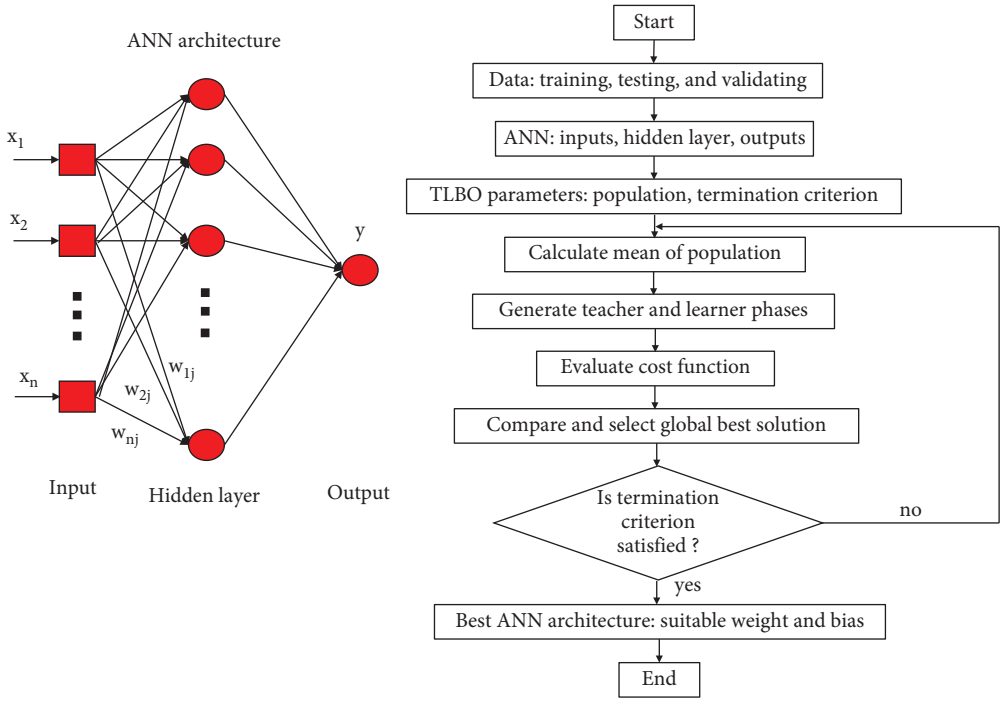


Figure 11: Scheme of optimization of ANN by TLBO.

the contributions of the parameters are listed as follows: C (37.65%), D (12.47%), E (7.20%), B (0.61%), and A (0.47%).

As shown in Table 7, the contributions of the input parameters on the safety factor are ordered as follows. The highest contribution is C (29.35%), A (5.78%), E (1.43%), B (1.98%), and D (0.06%), as provided in Figure 14.

4.3. Modeling Behaviors of Microrobotic Platform by ANN-Based TLBO. Modeling behaviors of the MPM platform is carried out through the ANN. To improve the effectiveness of the ANN technique, the TLBO is embedded into the ANN programming. Firstly, the collected data in Table 5 comprised of training, testing, and validating. The

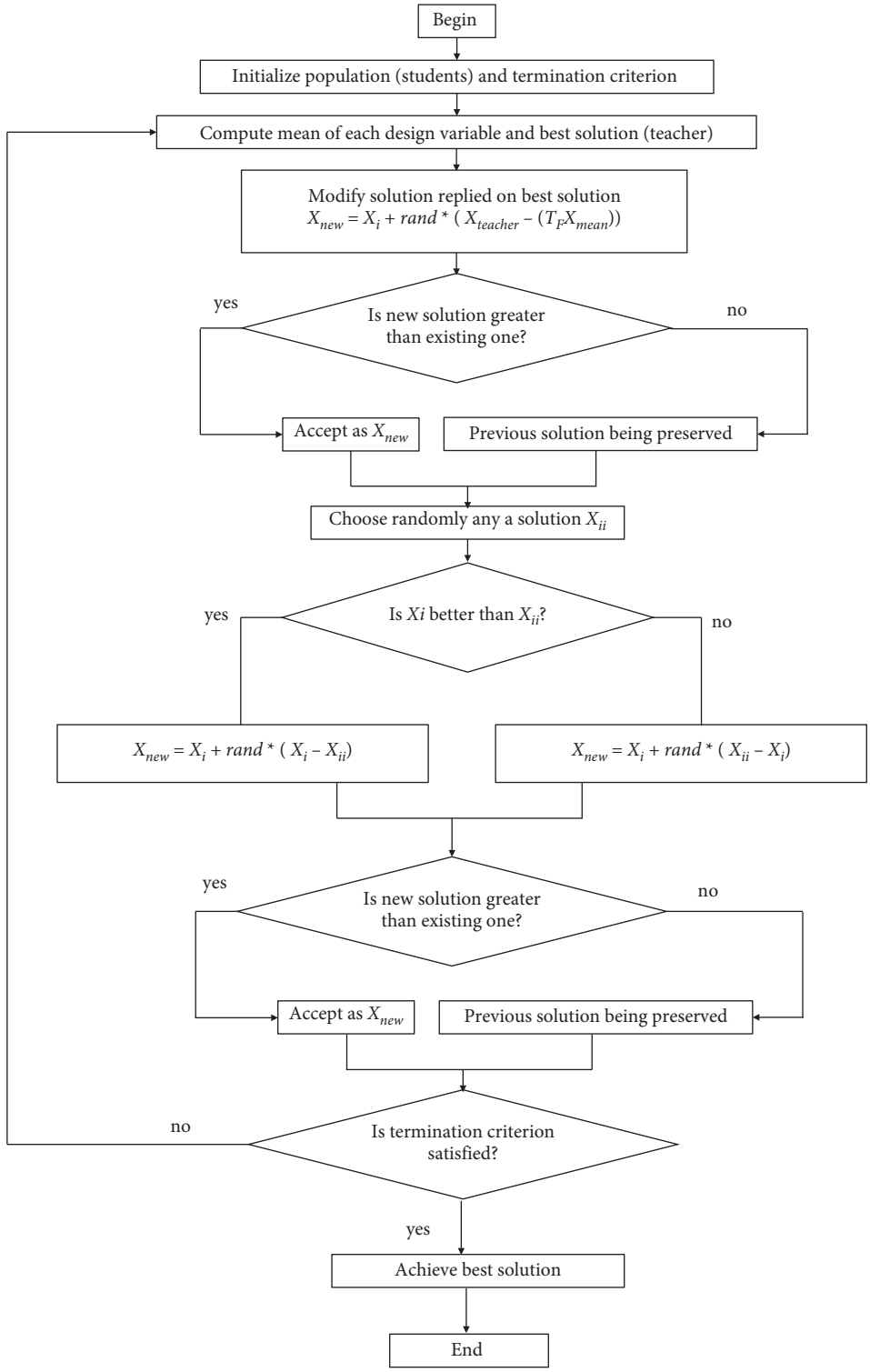


Figure 12: Flowchart of teaching-learning-based optimization method.

optimized ANN architecture can find the best weights and biases. The modeling accuracy of the optimized ANN is assessed by metric computation of the MSE and R^2.

Furthermore, the correlation coefficients (R) are also computed. The modeling results of the stroke and safety factor achieved very well with high R values, as plotted in Figures 15 and 16(a), respectively. The best performance, the prediction error, and the difference among the prediction and numerical values are provided, as seen in Figures 15, 16(c), and 16(d), respectively.

Table 5: Numerical results for the MPM platform.

No.	A (mm)	B (mm)	C (mm)	D (mm)	E (mm)	Stroke (μm)	Safety factor
1	0.85	0.75	0.65	0.6	50	1274.459	1.795
2	0.8	0.75	0.65	0.6	50	1271.977	2.216
3	0.9	0.75	0.65	0.6	50	1247.641	1.819
4	0.85	0.7	0.65	0.6	50	1270.549	2.003
5	0.85	0.8	0.65	0.6	50	1291.601	1.900
6	0.85	0.75	0.6	0.6	50	1316.365	1.747
7	0.85	0.75	0.7	0.6	50	1144.243	2.017
8	0.85	0.75	0.65	0.55	50	1355.784	1.905
9	0.85	0.75	0.65	0.65	50	1219.778	1.951
10	0.85	0.75	0.65	0.6	45	1292.106	1.945
11	0.85	0.75	0.65	0.6	55	1114.789	1.915
12	0.83	0.73	0.63	0.58	51.41	1285.452	1.896
13	0.86	0.73	0.63	0.58	48.58	1339.131	1.989
14	0.83	0.76	0.63	0.58	48.58	1350.635	1.612
15	0.86	0.76	0.63	0.58	51.41	1343.359	1.978
16	0.83	0.73	0.66	0.58	48.58	1245.309	2.223
17	0.86	0.73	0.66	0.58	51.41	1231.288	1.959
18	0.83	0.76	0.66	0.58	51.41	1359.739	2.101
19	0.86	0.76	0.66	0.58	48.58	1320.488	1.967
20	0.83	0.73	0.63	0.61	48.58	1370.22	1.806
21	0.86	0.73	0.63	0.61	51.41	1371.444	1.933
22	0.83	0.76	0.63	0.61	51.41	1362.347	1.929
23	0.86	0.76	0.63	0.61	48.58	1278.259	1.784
24	0.83	0.73	0.66	0.61	51.41	1158.641	2.032
25	0.86	0.73	0.66	0.61	48.58	1231.481	2.034
26	0.83	0.76	0.66	0.61	48.58	1190.894	2.002
27	0.86	0.76	0.66	0.61	51.41	1210.883	2.261

Table 6: Analysis of variance for the stroke.

Source	DF	Seq SS	Contribution (%)	Adj SS	Adj MS	F-value	P value
Model	20	120513	89.49	120513	6025.7	2.56	0.124
Linear	5	78626	58.39	74177	14835.4	6.29	0.022
A	1	629	0.47	402	401.5	0.17	0.694
B	1	816	0.61	538	538.3	0.23	0.650
C	1	50697	37.65	42836	42836.3	18.17	0.005
D	1	16789	12.47	22090	22089.6	9.37	0.022
E	1	9695	7.20	8915	8914.8	3.78	0.100
Square	5	11426	8.49	11673	2334.7	0.99	0.494
A*A	1	3	0.00	543	542.9	0.23	0.648
B*B	1	847	0.63	0	0.1	0.00	0.995
C*C	1	971	0.72	2371	2371.5	1.01	0.355
D*D	1	1855	1.38	125	125.0	0.05	0.826
E*E	1	7750	5.76	7236	7236.3	3.07	0.130
2-Way interaction	10	30461	22.62	30461	3046.1	1.29	0.392
A*B	1	3646	2.71	3897	3897.1	1.65	0.246
A*C	1	1382	1.03	1339	1338.9	0.57	0.480
A*D	1	6	0.00	149	148.8	0.06	0.810
A*E	1	612	0.45	654	654.0	0.28	0.617
B*C	1	5120	3.80	6498	6498.5	2.76	0.148
B*D	1	8707	6.47	7900	7900.3	3.35	0.117
B*E	1	2376	1.76	2570	2570.0	1.09	0.337
C*D	1	7492	5.56	7488	7488.1	3.18	0.125
C*E	1	1115	0.83	1103	1102.8	0.47	0.520
D*E	1	6	0.00	6	5.9	0.00	0.962
Error	6	14147	10.51	14147	2357.9		
Total	26	134661	100.00				

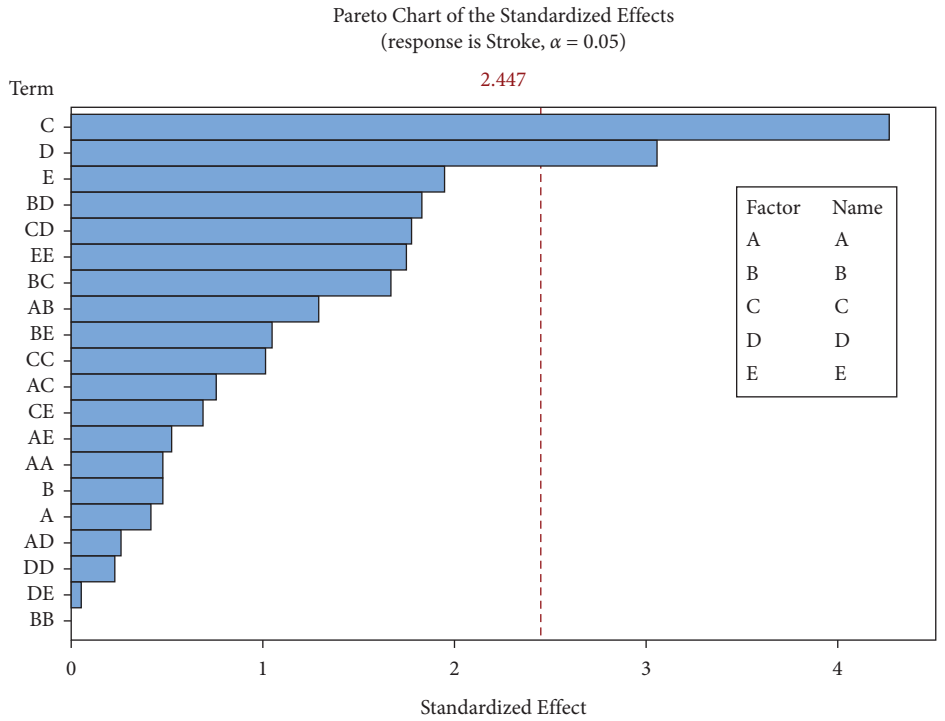


Figure 13: Sensitivity plot of design variables to the stroke.

Table 7: Analysis of variance for safety factor.

Source	DF	Seq SS	Contribution (%)	Adj SS	Adj MS	*F*-value	*P* value
Model	20	0.396108	72.62	0.396108	0.019805	0.80	0.679
Linear	5	0.210502	38.59	0.167944	0.033589	1.35	0.359
A	1	0.031505	5.78	0.041489	0.041489	1.67	0.244
B	1	0.010822	1.98	0.003362	0.003362	0.14	0.726
C	1	0.160082	29.35	0.109272	0.109272	4.39	0.081
D	1	0.000303	0.06	0.001510	0.001510	0.06	0.814
E	1	0.007789	1.43	0.010727	0.010727	0.43	0.536
Square	5	0.020878	3.83	0.020510	0.004102	0.16	0.967
A*A	1	0.006593	1.21	0.001006	0.001006	0.04	0.847
B*B	1	0.000844	0.15	0.000710	0.000710	0.03	0.871
C*C	1	0.010358	1.90	0.012192	0.012192	0.49	0.510
D*D	1	0.000886	0.16	0.002319	0.002319	0.09	0.770
E*E	1	0.002196	0.40	0.002143	0.002143	0.09	0.779
2-Way interaction	10	0.164728	30.20	0.164728	0.016473	0.66	0.731
A*B	1	0.001177	0.22	0.001342	0.001342	0.05	0.824
A*C	1	0.056306	10.32	0.056556	0.056556	2.27	0.182
A*D	1	0.000000	0.00	0.000038	0.000038	0.00	0.970
A*E	1	0.000713	0.13	0.001278	0.001278	0.05	0.828
B*C	1	0.004343	0.80	0.003475	0.003475	0.14	0.721
B*D	1	0.018617	3.41	0.018649	0.018649	0.75	0.420
B*E	1	0.057861	10.61	0.060386	0.060386	2.43	0.170
C*D	1	0.000085	0.02	0.000084	0.000084	0.00	0.955
C*E	1	0.022801	4.18	0.023640	0.023640	0.95	0.367
D*E	1	0.002825	0.52	0.002825	0.002825	0.11	0.748
Error	6	0.149333	27.38	0.149333	0.024889		
Total	26	0.545441	100.00				

As depicted in Figures 15 and 16, the proposed artificial intelligent technique had better performances than those achieved from the linear regression.

4.4. Parameter Optimization. In this part, the TLBO algorithm is initialized with a population of 50 and iterations of 5000. The optimization programming is implemented

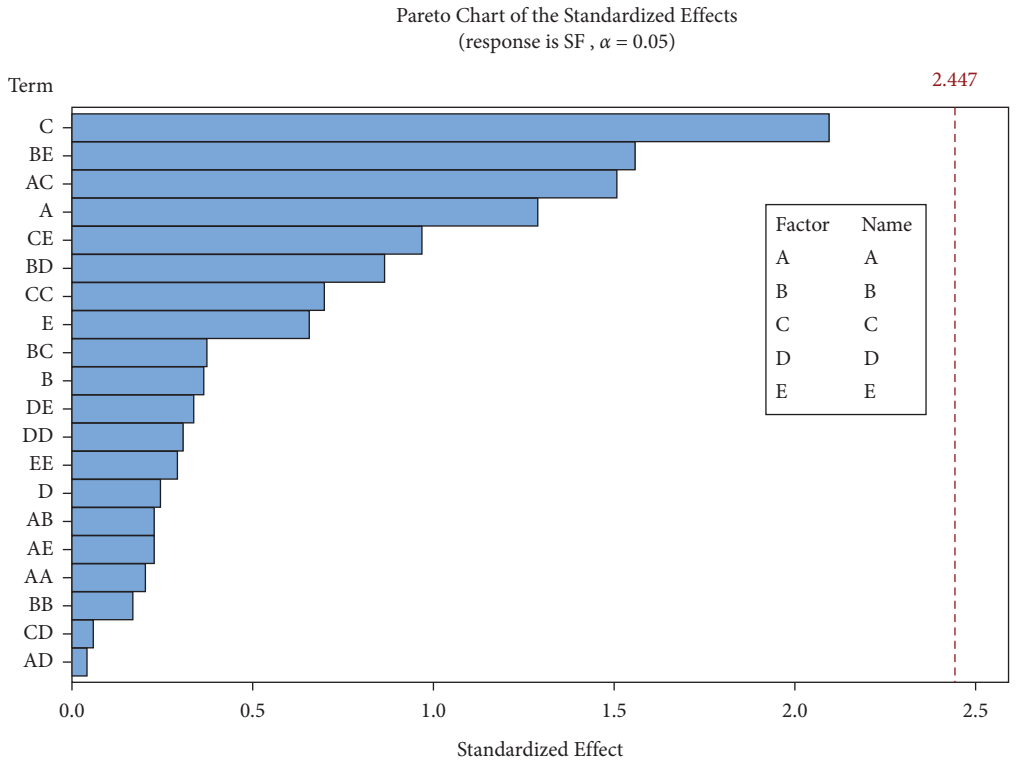


Figure 14: Sensitivity plot of design variables to the safety factor.

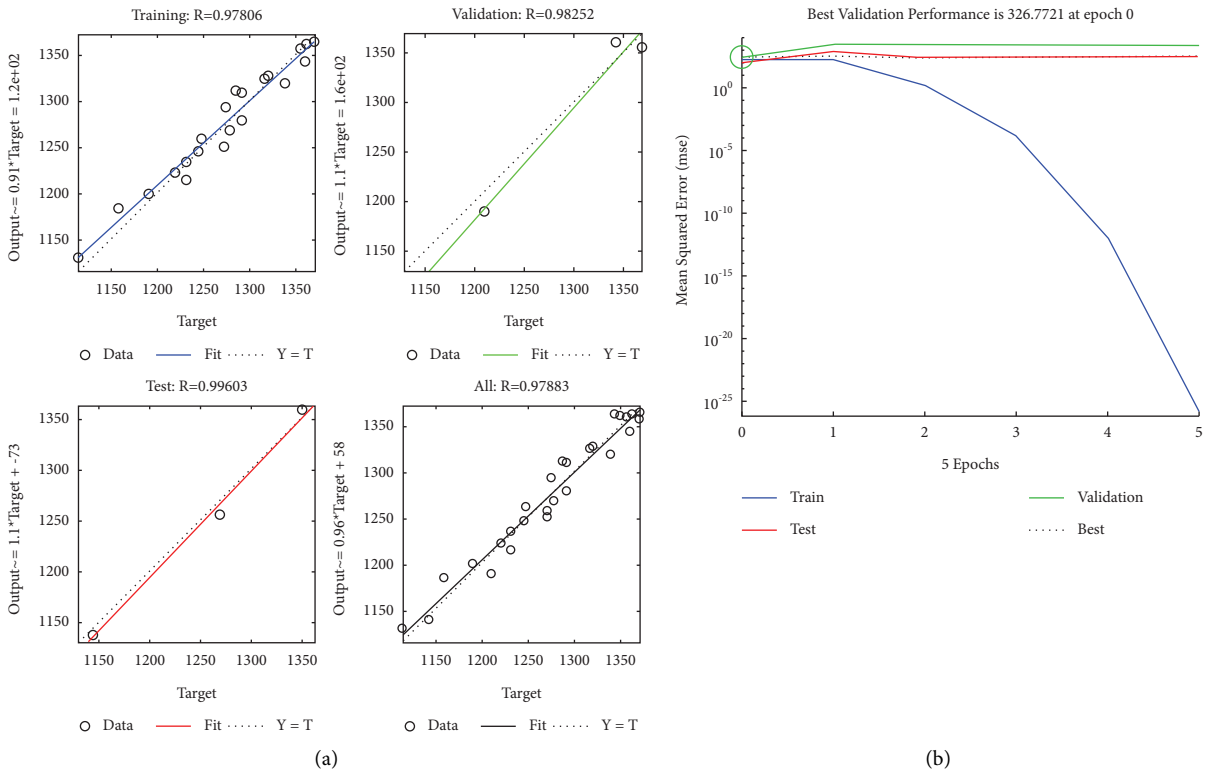
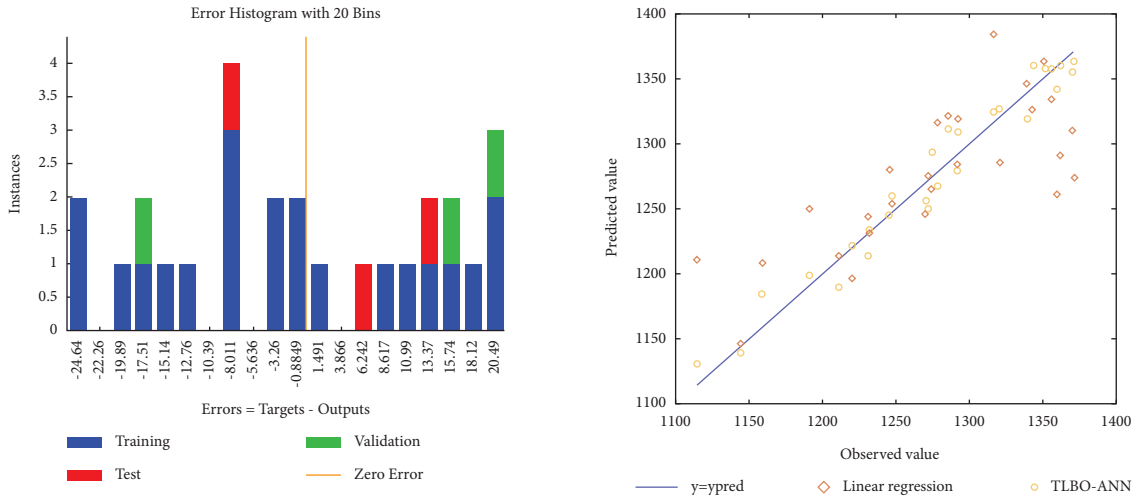
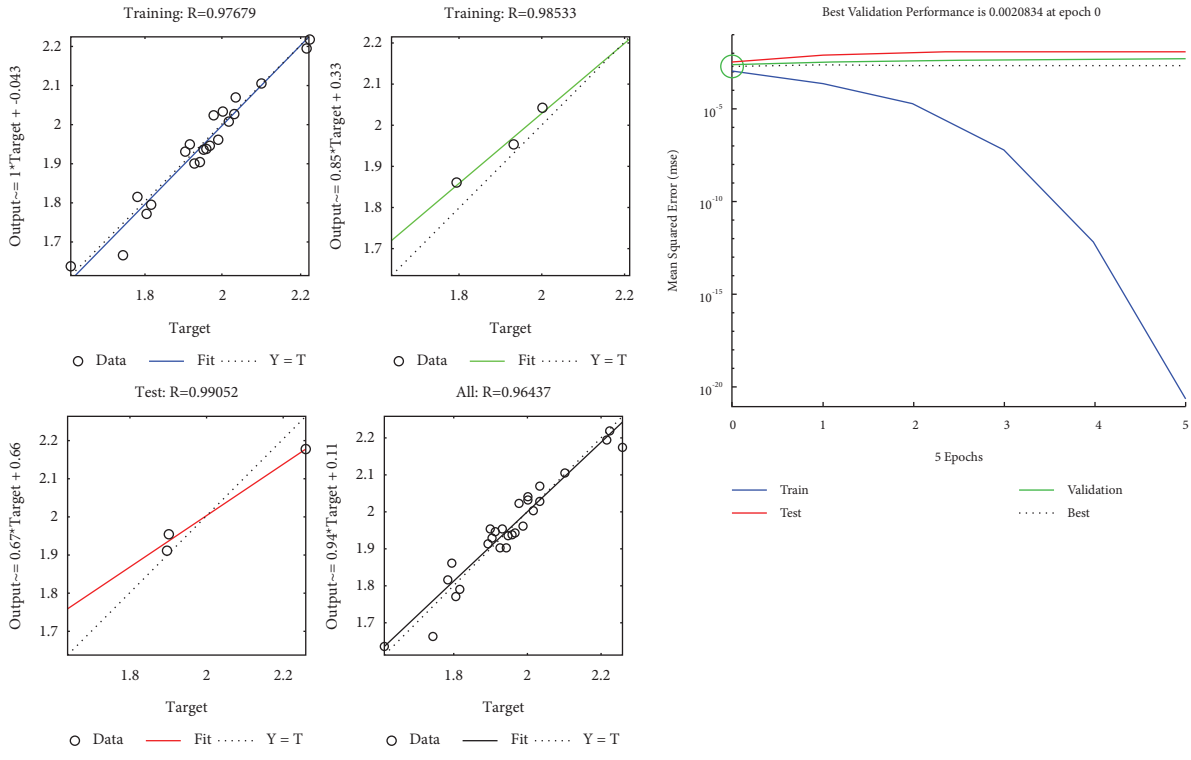


Figure 15: Continued.



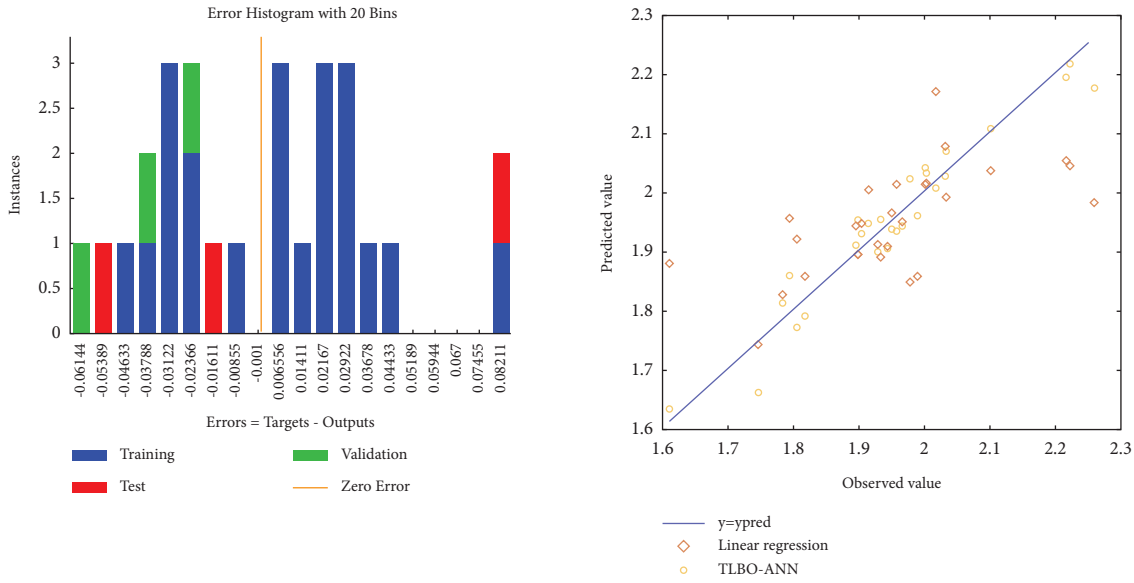
(c) (d)

Figure 15: Modeling for stroke by ANN-combined TLBO method: (a) training, (b) performance, (c) error, (d) predicted vs measured value.



(a) (b)

Figure 16: Continued.



(c) (d)

Figure 16: Modeling for safety factor by ANN-combined TLBO method: (a) training, (b) performance, (c) error, (d) predicted vs measured value.

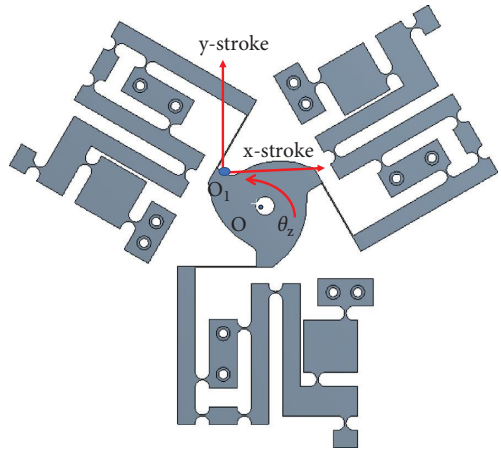


Figure 17: Measurement of rotation angle of the microrobotic platform.

MATLAB R2019 environment. The optimized results for the three case studies are provided in Table 6. From Figure 17, the rotation angle (θ_z) around the O point of the proposed MPM platform is measured by FEA ANSYS software. From Table 8, the y-stroke is the displacement along the y-axis at O_1 point. The y-stroke is the optimized displacement which is predicted from the proposed meta-heuristic-intelligent method (ANN-TLBO). The x-stroke is the displacement along the x-axis at O_1 point. The x-stroke, the stress and the rotation angle are calculated from the FEA ANSYS software.

From the achieved results of Table 8, it revealed that the optimized strokes in the y-axis of the MPM platform can obtain 1555.6763 μm, 1300.6 μm, and 1568 μm for case #1, case #2, and case #3, respectively. Besides, the x-axis strokes of the platform are 266.4 μm, 735.55 μm, and 714 μm for case #1, case #2, and case #3, respectively. The safety factor of the platform is over 1.5. Meanwhile, the stress appeared in three case studies is always lower than the yield stress (503 MPa) of AL 7075-T651. This guarantees a long working strength for the platform. The stress is calculated by the following equation.

$$S = \frac{S_{\text{yield}}}{SF}, \tag{14}$$

where, S represents the stress of the MPM platform. S_{yield} is the yield stress of AL 7075-T651. SF is the safety factor.

Based on the output stroke of the proposed MPM platform, the displacement amplification ratio can be calculated by following formula.

$$A_R = \frac{O_S}{I_S}, \tag{15}$$

Table 8: Optimum results for three case studies.

	Cases	Optimal solutions (mm)	y-stroke (μm)	x-stroke (μm)	Safety factor	Stress (MPa)	Rotation angle (degree)
TLBO for single-objective problems	Case 1	$A = 0.9$, $B = 0.8$, $C = 0.6$, $D = 0.6$, $E = 50$	1558.6763	266.4	1.58	318.35	1.85
	Case 2	$A = 0.87$, $B = 0.7$, $C = 0.6$, $D = 0.55$, $E = 49$	1300.6	735.55	2.33	215.87	1.97
TLBO for multi-objective problems	Cases	Optimal solutions (mm)	Stroke (μm)	x-stroke (μm)	Safety factor	Stress (MPa)	Rotation angle (degree)
	Case 3	$A = 0.89$, $B = 7.97$, $C = 0.6$, $D = 0.55$, $E = 45$	1568.1	714	2.04	246.56	2.26

Table 9: Validation results.

Case study	Method	Performances	
		y-stroke (μm)	Safety factor
Case 1	Proposed method	1558.6763	1.58
	FEA results	1432.2	1.47
Error (%)		8.8	7.48
Case 2	Proposed method	1300.6	2.3
	FEA results	1368.7	2.2
Error (%)		4.97	4.54
Case 3	Proposed method	1568.1	2.04
	FEA results	1689.8	2.17
Error (%)		7.2	5.9

where, A_R is the displacement amplification ratio. The O_S and I_S note the output y-stroke and input stroke.

By using equation (15), the A_R values are about 11.54 for case study #1, 9.63 for case study #2, and 11.61 for case study #3.

4.5. Validations of Optimized Results. By using the optimized design parameters, the prototypes are built in Inventor software, and then, the simulations are performed to verify the optimized results. As given in Table 9, the errors between the proposed method and the simulation method are under 9%. The proposed method is reliable optimization technique in modeling and optimizing the MPM platform.

5. Conclusions

This article has presented an optimized design method for the mobile microrobotic platform. The proposed MPM platform was built via using two combined modules, including the hybrid displacement amplification mechanism and leaf hinges. The developed HDAM was created by combination of Scott–Russell mechanism and two-double lever amplification mechanism. The new proposed HDAM amplifier could allow a large amplification ratio. With such a high amplification value, it ensured a large output stroke for the MPM platform. The developed MPM platform was able to be employed for locating the sample in the polishing robot system. The platform could achieve three motions, including two translations and one rotation.

In modeling the stroke and safety factor of the MPM platform, the ANN was used in combination with the TLBO method. By using the TLBO, the ANN architecture was optimized to a better approximation. And then, three optimized case studies were studied by the TLBO to improve the stroke and safety factor. Moreover, the case studies also demonstrated the effectiveness of the methodology. In this study, the FEM data was combined with ANN, TLBO for modeling process. The results of this paper could be listed as follows.

The modeling results from the TLBO-based ANN were well established. The metrics were relatively good with the values of R and R^2 being near 1 while the values of MSE were very small.

The established intelligent predictors were better than the linear regression. The predicted values from the TLBO-ANN were close to the measured values.

In case study #1, the optimized platform could operate with the y-axis stroke over 1558.6763 μm and a safety factor of 1.58.

In case study #2, the optimized platform could achieve a large y-axis of 1300 μm and a safety factor of 2.3.

In case study #3, the optimized platform could displace a large y-axis of 1568.1 μm and a safety factor of 2.04.

In summary, the proposed MPM platform could achieve a max-y stroke of 1568.1 μm, max-x stroke of 735.55 μm, and max-θ rotation angle of 2.26 degrees.

The stress of three cases were still lower than the yield stress of Al 7075-T651.

The proposed MPM platform could achieve a high displacement amplification ratio at least of 9.

In upcoming study, the real prototypes will be manufactured by WEDM. The physical verifications will be carried out. The polishing experiments will be conducted.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work belongs to the project grant no: T2021-11TĐ and was funded by Ho Chi Minh City University of Technology and Education, Vietnam.

References

[1] G. Chen and F. Ma, "Kinetostatic modeling of fully compliant bistable mechanisms using timoshenko beam constraint model," *Journal of Mechanical Design*, vol. 137, no. 2, pp. 1–10, 2015.

[2] N. Le Chau, N. T. Tran, and T. P. Dao, "A multi-response optimal design of bistable compliant mechanism using efficient approach of desirability, fuzzy logic, ANFIS and LAPO algorithm," *Applied Soft Computing*, vol. 94, Article ID 106486, 2020.

[3] L. Mingming, D. Zhao, J. Lin, X. Zhou, B. Chen, and H. Wang, "Design and analysis of a novel piezoelectrically actuated vibration assisted rotation cutting system," *Smart Materials and Structures*, vol. 27, no. 9, pp. 1–9, Article ID 095020, 2018.

[4] D. N. Nguyen, N. L. Ho, T.-P. Dao, and N. Le Chau, "Multi-objective optimization design for a sand crab-inspired compliant microgripper," *Microsystem Technologies*, vol. 25, no. 10, pp. 3991–4009, 2019.

[5] X. Ma, A. Wilson, C. D. Rahn, and S. Trolier-McKinstry, "Efficient energy harvesting using piezoelectric compliant mechanisms: theory and experiment," *Journal of Vibration and Acoustics*, vol. 138, no. 2, pp. 1–9, 2016.

[6] M. L. Culpepper and G. Anderson, "Design of a low-cost nano-manipulator which utilizes a monolithic, spatial compliant mechanism," *Precision Engineering*, vol. 28, no. 4, pp. 469–482, 2004.

[7] L. U. Odhner and A. M. Dollar, "The smooth curvature model: an efficient representation of Euler-Bernoulli flexures as robot joints," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 761–772, 2012.

[8] R. Wang and X. Zhang, "Preload characteristics identification of the piezoelectric-actuated 1-DOF compliant nano-positioning platform," *Frontiers of Mechanical Engineering*, vol. 10, no. 1, pp. 20–36, 2015.

[9] F. Wang, X. Zhao, Z. Huo et al., "A 2-DOF nano-positioning scanner with novel compound decoupling-guiding mechanism," *Mechanism and Machine Theory*, vol. 155, Article ID 104066, 2021.

[10] L. Clark, B. Shirinzadeh, Y. Tian, and B. Yao, "Development of a passive compliant mechanism for measurement of micro/nanoscale planar 3-DOF motions," *IEEE*, vol. 21, no. 3, pp. 1222–1232, 2016.

[11] S. Iqbal and A. Malik, "A review on MEMS based micro displacement amplification mechanisms," *Sensors and Actuators A: Physical*, vol. 300, Article ID 111666, 2019.

[12] H. Wang and X. Zhang, "Input coupling analysis and optimal design of a 3-DOF compliant micro-positioning stage," *Mechanism and Machine Theory*, vol. 43, no. 4, pp. 400–410, 2008.

[13] M. T. Pham, T. J. Teo, S. H. Yeo, P. Wang, and M. L. S. Nai, "A 3-D printed Ti-6Al-4V 3-DOF compliant parallel mechanism for high precision manipulation," *IEEE*, vol. 22, no. 5, pp. 2359–2368, 2017.

[14] Y. Li and Z. Wu, "Design, analysis and simulation of a novel 3-DOF translational micromanipulator based on the PRB model," *Mechanism and Machine Theory*, vol. 100, pp. 235–258, 2016.

[15] U. Bhagat, B. Shirinzadeh, L. Clark et al., "Design and analysis of a novel flexure-based 3-DOF mechanism," *Mechanism and Machine Theory*, vol. 74, pp. 173–187, 2014.

[16] W. L. Zhu, Z. Zhu, S. To, Q. Liu, B. F. Ju, and X. Zhou, "Redundantly piezo-actuated XYθz compliant mechanism for nano-positioning featuring simple kinematics, bi-directional motion and enlarged workspace," *Smart Materials and Structures*, vol. 25, no. 12, Article ID 125002, 2016.

[17] D. H. Chao, R. Liu, Y. M. Wu, L. Shi, and G. H. Zong, "Manufacturing error analysis of compliant 3-DOF micro-robot," *Frontiers of Mechanical Engineering in China*, vol. 1, no. 3, pp. 299–304, 2006.

[18] G. Villarrubia, F. Juan, D. Paz, P. Chamoso, and F. De la Prieta, "Artificial neural networks used in optimization problems," *Neurocomputing*, vol. 272, pp. 10–16, 2018.

[19] R. V. Rao, V. J. Savsani, and J. Balic, "Teaching-learning-based optimization algorithm for unconstrained and constrained real-parameter optimization problems," *Engineering Optimization*, vol. 44, no. 12, pp. 1447–1462, 2012.

Hindawi
Complexity
Volume 2022, Article ID 4275868, 15 pages
https://doi.org/10.1155/2022/4275868

Research Article

An Improved Sequential Recommendation Algorithm based on Short-Sequence Enhancement and Temporal Self-Attention Mechanism

Jianjun Ni,[1,2] **Guangyi Tang**,[1] **Tong Shen**,[1] **Yu Cai**,[1] **and Weidong Cao**[1,2]

[1]*College of Internet of Things Engineering, Hohai University, Changzhou 213022, China*
[2]*Jiangsu Key Laboratory of Power Transmission & Distribution Equipment Technology, Hohai University, Changzhou 213022, China*

Correspondence should be addressed to Jianjun Ni; njjhhuc@gmail.com

Received 11 February 2022; Accepted 19 April 2022; Published 29 September 2022

Academic Editor: Jesus Vega

Sequential recommendation algorithm can predict the next action of a user by modeling the user's interaction sequence with an item. However, most sequential recommendation models only consider the absolute positions of items in the sequence, ignoring the time interval information between items, and cannot effectively mine user preference changes. In addition, existing models perform poorly on sparse data sets, which make a poor prediction effect for short sequences. To address the above problems, an improved sequential recommendation algorithm based on short-sequence enhancement and temporal self-attention mechanism is proposed in this paper. In the proposed algorithm, a backward prediction model is trained first, to predict the prior items in the user sequence. Then, the reverse prediction model is used to generate a batch of pseudo-historical items before the initial items of the short sequence, to achieve the goal of enhancing the short sequence. Finally, the absolute position information and time interval information of the user sequence are modeled, and a time-aware self-attention model is adopted to predict the user's next action and generate a recommendation list. Various experiments are conducted on two public data sets. The experimental results show that the method proposed in this paper has excellent performance on both dense and sparse data sets, and its effect is better than that of the state of the art.

1. Introduction

With the development of Internet technology, recommender systems have become one of the indispensable tools in people's daily life [1–4]. Compared with traditional methods, the sequential recommendation model performs well on the Top-N recommendation problem [5]. In recent years, with the development of deep learning technology, sequential recommendation models based on deep learning have been widely used, such as e-commerce shopping platforms, medical and health services [6, 7], and audiovisual platforms [8]. The user's interaction behavior with items in such application platforms can be regarded as a sequence of behaviors in chronological order. Based on this, researchers have proposed various sequential recommendation models to mine and analyze user-item interaction information. The purpose of these models is to provide users with a personalized recommendation list containing N items to help users filter out valuable information.

The recommendation model based on the Markov chain (MC) [9] method is one of the early methods of sequential recommendation, which assumes that the user's next action is determined by his historical behavior and transforms the recommendation problem into a sequence prediction problem. In recent years, with the continuous breakthroughs of the deep neural networks (DNN) in the field of artificial intelligence [10–12], researchers have tried to introduce a series of deep neural network models into the field of recommendation and have achieved a series of results [13–15]. For example, Huang et al. [16] combined the traditional MC

method and the recurrent neural network (RNN) to optimize the recommendation model and improve the recommendation accuracy. Based on long-short-term memory (LSTM) network, Xu et al. [17] combined self-attention network to capture users' complex and dynamic behavioral preferences. Inspired by the semantic understanding model, Sun et al. [18] applied the bidirectional attention model to sequential recommendation, combining user context information and making recommendations. The existing sequential recommendation models tend to perform poorly when there are a large number of short-sequence users in the data set [19]. In addition, most of the existing sequential recommendation models only consider the absolute position information of the user sequence and assume that each item has the same time interval, ignoring the impact of the time interval between items on the recommendation results, which cannot capture user preferences effectively [20].

To address these issues introduced above, some improved sequential recommendation models are proposed. For example, Zhao et al. [21] employed a deep bidirectional long-short-term memory network and attention mechanism to capture the changes in user preferences. Liu et al. [22] first used the method of reverse training short sequences to expand short sequences and fine-tuned the model through the enhanced short sequences, which can achieve certain results on sparse data sets. Ahmadian et al. [23] adopted a deep learning based trust- and tag-aware recommender system, to extract potential features through sparse automatic encoder, which can effectively solve the problem of data sparsity. Li et al. [24] adopted a time-aware self-attention mechanism to explore the effect of different time intervals on the prediction results. These methods lay a good foundation for the research of sequential recommendation models, but there are still some problems that are not well solved. For example, the pseudo-historical items generated by direct reverse training are not accurate enough, and the time interval information is not sufficiently mined to capture user preferences well.

Based on previous research, we propose a sequential recommendation model based on short-sequence enhancement and improved time-aware self-attentive mechanism to address the above-mentioned problems. In the proposed model, the data set is first preprocessed to divide users into long-sequence sets and short-sequence sets. Then, by reverse training the long-sequence set, a reverse prediction model is generated. Finally, the model is transferred for the short sequence, and a batch of pseudo-historical items is generated before the initial item of the short sequence, to enhance the short sequence and solve the problem of data sparsity. At the same time, the model adopts an improved time interval self-attention mechanism, which not only considers the influence of absolute location information on the recommendation effect but also considers the influence of the time interval information between any two items on the recommendation result.

The proposed model in this paper can fully reflect the changes of user preferences over time and improve the accuracy of the recommender system. In summary, the main contributions of this paper are as follows: (1) pretrain a reverse prediction model, use the transfer learning method to reverse predict short sequences, and generate a batch of pseudo-historical items before the initial items of the short sequence, so as to achieve the purpose of enhancing short sequences. (2) Combined with the absolute position information and time interval information of the item, an improved time-aware self-attention mechanism is used to give the absolute position weight and time interval weight of different items, fully exploit the change of user behavior preferences, predict the user's next action, and generate a list of recommendations. (3) Extensive experiments on two real data sets are conducted. The results demonstrate the effectiveness of the proposed model, which can outperform existing methods on two different metrics. In addition, the influence of each key component in the proposed model on the recommendation results is discussed through multiple experiments.

This paper is organized as follows. Section 2 introduces the related works. Section 3 gives out the details of the proposed model. Section 4 provides experiments and analysis of results. Section 5 discusses the parameters and important components of the proposed algorithm. Section 6 provides the conclusions.

2. Related Works

2.1. Sequential Recommendation Model. The earliest sequential recommendation models are mainly based on the MC method [25]. These MC-based models have a significant improvement over other types of recommendation algorithms in terms of short-term prediction. However, this type of model cannot capture the long-term behavioral features in the user sequence and has low accuracy and high computational complexity in long-term prediction.

As deep learning technology shines in the fields of machine vision and natural language processing [26, 27], the introduction of deep learning technology into recommender systems has also become the focus of researchers. For example, Zhang et al. [28] designed a new session-based recommendation method based on recurrent neural network, which fuses user's general preference information and dynamic preference information. Sun et al. [29] proposed a method based on temporal context awareness and RNN, which can effectively capture the correlation between items. In addition, long-short-term memory (LSTM) and gated recurrent unit (GRU) (two popular variants of RNNs) have also achieved results in the field of recommendation. For example, Yuan et al. [30] computed the global state transitions of user sequences to model user interest preference changes, based on an improved GRU model. Zhao et al. [31] proposed a content-aware movie recommendation model based on LSTM, which effectively utilizes the long-term and short-term information of the sequence for content perception and movie recommendation. However, most models assume that user behavior sequences are simple time-sequential sequences, without considering the time interval information between items. At the same time, existing models perform poorly on sparse data sets and short-sequence users.

2.2. Transformer-Based Model. Attention mechanism has achieved great results in a large number of works, such as image processing [32, 33] and natural language processing [34]. The essence of the attention mechanism can be understood as selecting some important information from a large amount of information and giving them weights, where the size of the weights represents the importance of the information. In recent years, transformer, a neural network architecture based on pure attention mechanism, has achieved excellent performance and effects in the field of machine translation [35]. Inspired by this, researchers introduced the transformer model into the recommender system [36] and achieved good results. The transformer-based model uses scaled dot-product attention, which is presented as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q, K, V are three matrices representing queries, keys, and values, respectively; $\sqrt{d}$ is the scaling factor, which is used to avoid the inner product value being too large; and Softmax is the normalized function [33].

The attention function can be described as mapping a query and a set of key-value pairs to an output, where the queries, keys, values, and output are all vectors. The output is computed as a weighted sum of the values. The transformer model adopts the multihead attention mechanism, which executes the attention function in parallel, and connects each output value with item linearly again to obtain the final result. The multihead attention mechanism enables the model to focus on the information of different subspaces from different locations at the same time. The architecture of the transformer-based model is shown in Figure 1.

3. Proposed Sequential Recommendation Model

In this paper, a sequential recommendation model (SeTsRec) based on short-sequence enhancement and temporal self-attention mechanism is proposed, which is shown in Figure 2. First, the original data are preprocessed, where users are divided into long-sequence user sets and short-sequence user sets; and then the long-sequence sets are reversely input into the transformer network to train a reverse prediction model. Subsequently, inspired by the transfer learning method [37], this reverse prediction model is transferred to short-sequence users to generate a batch of pseudo-historical items before the initial item of the short-sequence user behavior list. By combining pseudo-historical items and short-sequence user behavior lists, an augmented sequence of short sequences is generated to enhance short sequences. Finally, the long sequences and enhanced short sequences are used as input to train a time-aware self-attention recommendation model and predict the user's next action. The model proposed in this paper will be described in detail as follows.

3.1. Problem Description. In the sequential recommendation problem, we assume that $U = \{u_1, u_2, \ldots, u_n\}$ is the user set of the system, where n is the number of the users in the data set, and $I = \{i_1, i_2, \ldots, i_m\}$ is the item set of the system, where m is the number of the items in the data set. For a certain user u, $S^u = \{i_1, i_2, \ldots, i_w\}$ and $T^u = \{t_{i_1}, t_{i_2}, \ldots, t_{i_w}\}$ are the user behavior sequence and time series, respectively, indicating that the length of the behavior sequence of the user u is w. Each item in the behavior sequence S^u is arranged in the chronological order of the user's interaction with it, and each element in the time sequence T^u represents the actual interaction time between the user u and the item i. At a certain moment, given the user's behavior sequence S^u and time series T^u, the goal of the model is to predict the next item that the user u is most likely to interact with, which is expressed as

$$p(i_{w+1}) = f(i_1, \ldots, i_w, t_{i_1}, \ldots, t_{i_w}), \tag{2}$$

where $p(\cdot)$ is the output probability of a certain item and $f(\cdot)$ is the nonlinear function that needs to be learned.

Recommendation systems usually provide users with multiple recommendation results and finally generate a recommendation list containing N items. Set $Y^u = \{y_1^u, y_2^u, \ldots, y_N^u\}$ as the output possibility of all the candidates, according to the output probability of the candidates, select the previous N items for recommendation, which is the famous Top-N recommendation problem in the recommendation system.

3.2. Short-Sequence Enhancement. The sequential recommendation algorithm is a recommendation method that predicts the user's next action by mining the information contained in the user's behavior sequence. Therefore, the validity of user behavior sequence information is crucial. Existing sequential recommendation methods have achieved good results. However, most of the existing methods do not solve the short-sequence prediction problem well and often perform poorly on sparse data sets. To deal with the limitation problem of the sequential recommendation model on sparse data sets, the proposed method in this paper utilizes the transfer learning to enhance short sequences on the basis of existing research, which will be introduced in detail as follows.

3.2.1. Reverse Prediction Model. Ideally, in the field of machine learning, it is always expected that the data sets used for model training are dense and efficient. However, in the actual research process, the data sets often have a large amount of data sparse phenomenon. In sparse data sets, there are often a large number of missing or zero data, which makes the data availability very poor, and brings many difficulties to establish the recommendation models.

In this paper, the user set U is first divided into a long-sequence user set U_L and a short-sequence user set U_S according to the length of the user sequence. The long-sequence user set U_L is a dense datas et, and the short-sequence user set U_S is a sparse data set. For the long-sequence

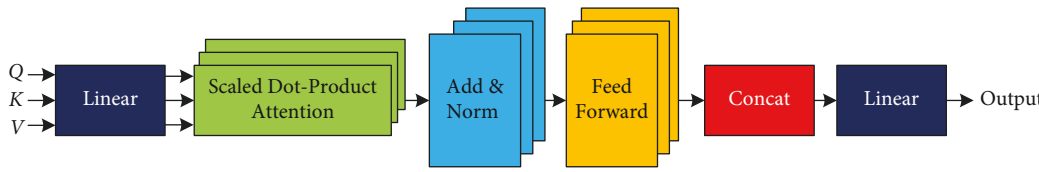


Figure 1: The architecture of the transformer-based model, where Q, K, V are three matrices representing queries, keys, and values, respectively.

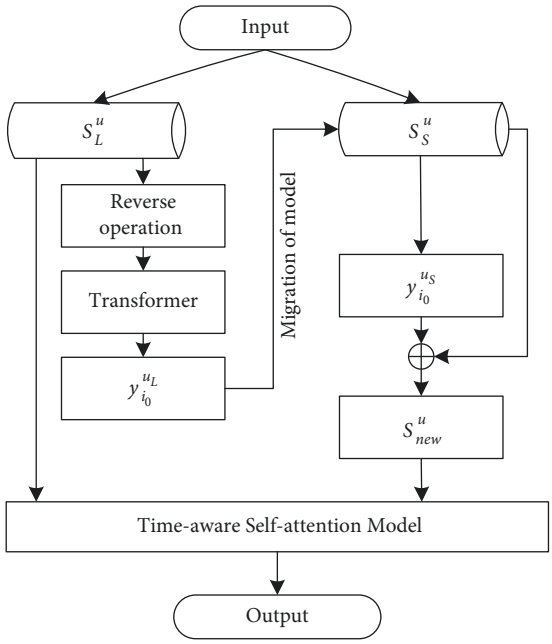


Figure 2: The framework of the proposed model, where S_L^u represents the behavior sequence of the long-sequence user set U_L; S_S^u represents the behavior sequence of the short-sequence user set U_S; $y_{i_0}^{u_L}$ represents the previous item of the item i_1 in the sequence S_L^u; $y_{i_0}^{u_S}$ represents the previous item of the item i_1 in the sequence S_S^u; and S_{new}^u is the generated enhanced short sequence.

user set U_L, the behavior sequence S_L^u can be obtained. In this paper, the long sequence is first reversed to obtain the reverse sequence S_r, and then the reverse sequence S_r is input into the transformer layer for training, to obtain a reverse prediction model. For the user u, the purpose of this reverse prediction model is to predict the previous item of the sequence $S_L^u = \{i_1, i_2, \ldots, i_n\}$:

$$y_{i_0}^{u_L} = f\left(S_L^u : i\right), \tag{3}$$

where $y_{i_0}^{u_L}$ represents the previous item of the item i_1 in the sequence S_L^u. Although the model is reverse-trained, the transformer network is also able to mine interitem correlations, which has been demonstrated in previous work [22].

3.2.2. Pseudo-Historical Item Generation. The existing methods often regard the data set as a whole for recommendation tasks, which ignore the different data quality of different users in the same data set. Specifically, some users have interacted more with the item, and the data of these users is relatively rich and reliable; while some users have little interaction data with the item, so the data of these users are sparse and can be poor usability. In order to solve this problem, this paper uses the transfer learning method to transfer the reverse prediction model of long-sequence users obtained above to short-sequence users [38]. The long-sequence reverse prediction task is taken as the source task, and the short-sequence reverse prediction task is taken as the target task. By fully mining the rich data information provided by the long-sequence users, the data sparsity problem of the short-sequence users is alleviated, which can improve the overall recommendation quality. Taking the short-sequence set as input, the reverse prediction model is used to generate pseudo-history items of short-sequence users, namely:

$$y_{i_0}^{u_S} = f\left(S_S^u : i\right), \tag{4}$$

where S_S^u represents the behavior sequence of the short-sequence user set U_S and $y_{i_0}^{u_S}$ represents the previous item of the item i_1 in the sequence S_S^u.

For a data set, we define the length L to represent the threshold for the short-sequence user set. Namely, if the length of a sequence (denoted by $|S^u|$) is less than L, this sequence is regarded as a short sequence; otherwise, the sequence is regarded as a long sequence.

In this paper, we denote the generated set of pseudo-historical items as $\{i_{-q+1}^u, \ldots, i_{-1}^u, i_0^u\}$ and place this set before the initial items i_1^u of the original short sequence to form an augmented short sequence, where q is the total number of pseudo-historical items generated by short sequences. Figure 3 shows the enhanced short-sequence set, in which the yellow part represents the generated pseudo-historical item, and the green part represents the original short sequence. In Figure 3, it is assumed that $q = 3, L = 4$. The generated enhanced short sequence is denoted by

$$S_{\text{new}}^u = \left\{y_{-q+1}, \ldots, y_{-1}, y_0, i_1, i_2, \ldots, i_n\right\}. \tag{5}$$

3.3. Time-Aware Self-Attention Model. The existing sequential recommendation model simply regards the user's behavior list as a sequential sequence according to the interaction time between the user and the item. In addition, the items are regarded as having the same time interval. Specifically, as shown in Figure 4(a), if the user A and the user B have been exposed to the same item, the traditional method will regard the time interval between the items in the two sequences as a fixed value of N days, which will lead to the same result for the two different users. However, such a result is unreasonable because the actual time that the user A and the user B have access to these items is different.

In the actual application scenarios, the time interval between items will be different even if the user's behavior list

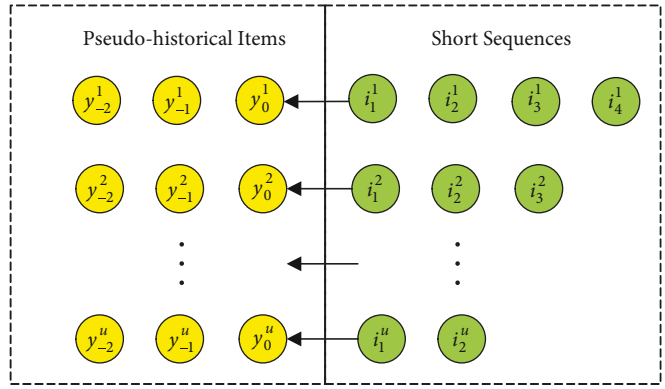


Figure 3: The example of the short-sequence enhancement process.

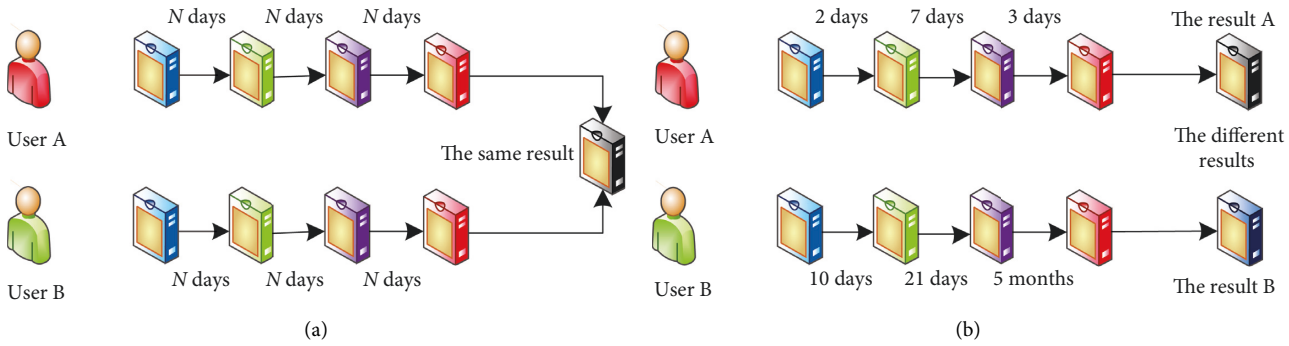


Figure 4: The effect of different time intervals on recommendation results. (a) Traditional sequential recommendation. (b) Our proposed method.

is exactly the same due to the different actual interaction time between users and items. As shown in Figure 4(b), although the user A and the user B have been exposed to the same items, the time interval between items is different. In this case, if the model can combine the different time interval information between the items, it is possible to make more accurate recommendation results. To solve the above problems, this paper adopts an improved time-aware self-attention model. The overall framework of the proposed model is shown in Figure 5, which will be introduced in detail as follows.

3.3.1. Time Interval Matrix. After getting the augmented sequence of user behavior S^u_{new}, it is used as model input together with the long-sequence S^u_L. First, the two different types of behavior sequences are converted into a fixed-length sequence S:

$$S^u_{\text{new}} \cup S^u_L \longrightarrow S = (s_1, s_2, \ldots, s_m), \tag{6}$$

where m represents the maximum sequence length of the input model. If the length of the sequence S^u_{new} or S^u_L is greater than m, only the latest m items are considered; otherwise, padding items are added to the left of the sequence S until its length reaches m.

Similarly, for the time series T^u_{new} and T^u_L, they can be converted to a fixed sequence t:

$$T^u_{\text{new}} \cup T^u_L \longrightarrow t = (t_1, t_2, \ldots, t_m). \tag{7}$$

If the length of the sequence T^u_{new} or T^u_L is greater than m, only the latest m items are considered; otherwise, the time corresponding to the first item t_1 is used on the left side of the sequence t, and padding it until its length reach m. In this study, for the time of the pseudo-historical items generated in Section 3.2.2, the average time interval $t_{\text{avg}} = \sum_m^{i=1} \sum_m^{j=1} r^u_{ij}/2$ is used to define them in turn, which are calculated as follows:

$$\begin{cases} t_{y_0} = t_1 - t_{\text{avg}}, \\ t_{y_{-1}} = t_{y_0} - t_{\text{avg}}, \\ \vdots \\ t_{y_{-q+1}} = t_{y_{-q+2}} - t_{\text{avg}}. \end{cases} \tag{8}$$

After obtaining the user's fixed time series $t = (t_1, t_2, \ldots, t_m)$, define the time interval between any items as $\triangle t = |t_i - t_j|$. Due to the different frequency of interaction between different users and items, this paper adopts the relative length of the time interval between items, which is defined as follows [24]:

$$r^u_{ij} = \left| \frac{\Delta t}{r_{\min^u}} \right|, \qquad r_{\min^u} = \min(\Delta t). \tag{9}$$

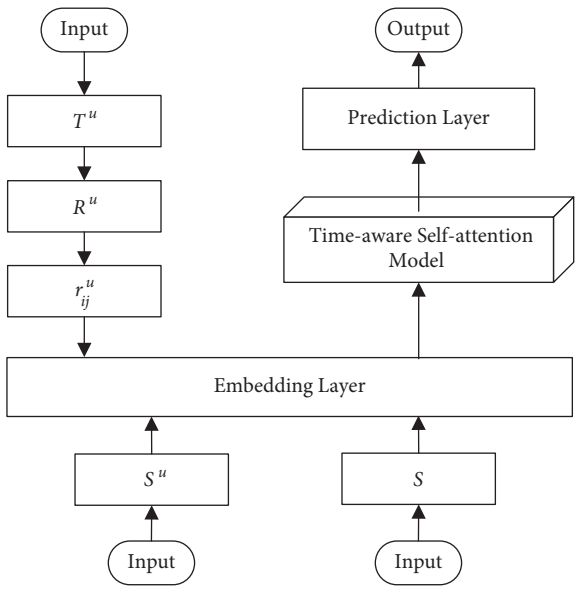


Figure 5: The framework of the proposed model, where T^u represents the time series of user u; R^u is the time interval matrix of user u; r_{ij}^u is the element in the time interval matrix R^u; S^u represents the behavior sequence of user u; and S is the absolute position information sequence of the item.

Finally, the time interval matrix R^u of the user u can be obtained:

$$R^u = \begin{bmatrix} r_{11}^u & r_{12}^u & \cdots & r_{1m}^u \\ r_{21}^u & r_{22}^u & \cdots & r_{2m}^u \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}^u & r_{m2}^u & \cdots & r_{mm}^u \end{bmatrix}. \tag{10}$$

3.3.2. Time-Aware Self-Attention Module. (1) Time-Aware Self-Attention Layer. For each input sequence, an embedding layer is applied to convert the user behavior sequence into item embedding matrix $E^I \in \mathfrak{R}^{m\times d}$, absolute position information into position embedding matrix $E_K^P, E_V^P \in \mathfrak{R}^{m\times d}$, and time interval information into time interval embedding matrix $E_K^{\mathfrak{R}}, E_V^{\mathfrak{R}} \in \mathfrak{R}^{m\times m\times d}$ (d is the latent dimension):

$$\begin{aligned} E^I &= \left[m_{s1}, m_{s2}, \ldots, m_{sm}\right]^T, \\ E_K^P &= \left[p_1^k, p_2^k, \ldots, p_m^k\right]^T, \\ E_V^P &= \left[p_1^v, p_2^v, \ldots, p_m^v\right]^T, \\ E_K^{\mathfrak{R}} &= \begin{pmatrix} r_{11}^k & r_{12}^k & \cdots & r_{1m}^k \\ r_{21}^k & r_{22}^k & \cdots & r_{2m}^k \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}^k & r_{m2}^k & \cdots & r_{mm}^k \end{pmatrix}, \\ E_V^{\mathfrak{R}} &= \begin{pmatrix} r_{11}^v & r_{12}^v & \cdots & r_{1m}^v \\ r_{21}^v & r_{22}^v & \cdots & r_{2m}^v \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}^v & r_{m2}^v & \cdots & r_{mm}^v \end{pmatrix}. \end{aligned} \tag{11}$$

Then a new sequence $Z \in \mathfrak{R}^{m\times d}$ is calculated:

$$Z = \left[z_1, z_2, \ldots, z_m\right]^T, \tag{12}$$

where $z_i \in \mathfrak{R}^d$ is obtained by the input item embedding, absolute position embedding, and time interval embedding, namely

$$\begin{aligned} Z_i &= \sum_m^j \alpha_{ij}\left(m_{sj}W^V + r_{ij}^V + p_j^V + b_i\right), \\ \alpha_{ij} &= \frac{\exp\left(e_{ij}\right)}{\sum_m^{k=1}\exp\left(e_{ik}\right)}, \\ e_{ij} &= \frac{m_{sj}W^Q\left(m_{sj}W^K + r_{ij}^k + p_j^k\right)^T}{\sqrt{d}}, \end{aligned} \tag{13}$$

where W^V, W^Q, $W^K \in \mathfrak{R}^{d\times d}$, respectively, represent the input item matrix of the value, query and key; $\sqrt{d}$ is the scale factor, which is used to prevent the inner product from being too large; and b_i is the bias term.

(2) Point-Wise Feed-Forward Network. The self-attention layer of the model is mainly based on linear combination to realize the combination of absolute position information and relative time interval information of items. In order to make the model have nonlinear characteristics and consider the interaction between different latent dimensions, we apply a point-wise feed-forward network to each output of the self-attention layer:

$$\mathrm{FFN}\left(z_i\right) = G\left(\left(z_iW_1 + b_1\right)W_2\right) + b_2, \tag{14}$$

where $G(\cdot)$ is an activation function, which is ELU in this paper. The main reason of using ELU function is that it can solve the Dead Relu problem, while reducing the influence of the bias term offset, and the learning rate is faster. $W_1, W_2 \in \mathfrak{R}^{d\times d}$ represents the weight matrix and $b_1, b_2 \in \mathfrak{R}^d$ represents the bias term.

(3) Stacking Self-Attention Blocks. With the continuous stacking of self-attention layers and point-wise feed-forward networks, problems such as overfitting and long training time will occur. In order to solve these problems, this paper adopts the residual connection, dropout, and layer normalization processing methods [36]:

$$\begin{aligned} Z_i &= z_i + \mathrm{Dropout}\left(\mathrm{FFN}\left(LN\left(z_i\right)\right)\right), \\ \mathrm{LN}\left(x\right) &= \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}\odot\gamma + \beta, \end{aligned} \tag{15}$$

where $\odot$ is the element-level product; μ, σ represents the mean and variance of x; and γ, β represents the learned scale factors and bias terms.

The specific workflow is as follows: for each self-attention block, layer normalization is first applied to each input z_i, which is beneficial to stabilize and speed up the training process of the neural network. Then, the output of the self-attention layer is applied to the point-wise feed-forward network, to give the model nonlinearization features. Finally, a dropout regularization technique is applied to the

output of the position feed-forward network, to alleviate the overfitting problem that occurs in deep neural networks. The main reason for using the dropout regularization technology is that it can control overfitting by artificially destroying data, which has been proven to be effective in various neural network architectures [39, 40].

3.3.3. Prediction Layer. After the prediction layer obtaining the final representation of the absolute position of the item and the time interval, in order to predict the possible next action of the users, we use a Softmax function to calculate the user's interaction probability with the candidate item $y_{i,t}$, namely:

$$p\left(y_{i,t}\right) = \text{Softmax}\left(Z_t M_i^I\right), \tag{16}$$

where M_i^I represents the embedding vector of items i and Z_t represents the first given sequence $(i_1, i_2, \ldots, i_t)$ containing t items and their time interval $(r_{1(t+1)}, r_{2(t+1)}, \ldots, r_{t(t+1)})$ between the $t+1$th item.

3.3.4. Model Inference. This paper uses user implicit interaction data. In implicit feedback, the interaction between the user and the item can be regarded as a binary classification problem, where 1 means that the user likes an item and 0 means that the user does not like or has not touched the item. Therefore, the items in the user behavior sequence can be regarded as positive samples. At the same time, all the items that the user has not touched are regarded as negative feedback, which is sampled as negative samples. In this paper, the sampling is carried out according to the ratio of 1: 1. When negative sampling is performed for each user, the principle is to select those items with higher popularity, which are more representative. The loss function is as follows [30]:

$$\text{Loss} = \sum_{(u,i,j)\in S} \log\left(1 + e^{-\left(p\left(y_i^u\right) - p\left(y_j^u\right)\right)}\right) + \lambda\|\ominus\|_F^2, \tag{17}$$

where i represents the predicted candidate items; j represents the negative sample; $\ominus = \{E^I, E_K^P, E_V^P, E_K^{\mathfrak{R}}, E_V^{\mathfrak{R}}\}$ is the set, which represents the embedding matrix; and λ is the regularization parameter, which is used to prevent the model from overfitting. In the training process, the Adam optimizer is used to optimize the model, which is a variant of the stochastic gradient descent (SGD) algorithm [41]. As an adaptive learning rate optimization algorithm, the Adam optimizer is usually used for tasks in sparse data scenarios, and its convergence speed is fast [42].

In summary, the process of the proposed algorithm is shown in Algorithm 1.

4. Experiments

4.1. Setting of Experiments

4.1.1. Data set. In order to verify the effectiveness of the proposed algorithm in this paper, experiments were carried out on two public data sets, namely Movielens-1M data set (denoted by ML-1M, see https://files.grouplens.org/datasets/movielens/ml-1m.zip) and Amazon Beauty data set (denoted by AM-BE, see https://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_5.json.gz). Among them, the ML-1M data set is a dense data set, and the AM-BE data set is a sparse data set. The dense data set ML-1M in this paper is used to evaluate the effectiveness of the time self-attention improvement in the proposed model, while the sparse data set AM-BE is used to evaluate the effectiveness of the improved short-sequence enhancement method of the proposed model. The statistics of the two data sets are listed in Table 1, which contains the information such as users, items, and timestamps.

Before the experiments, the two data sets are preprocessed [17]. For all data sets, we treat the rating behaviors as implicit feedback, where "1" means that there is an interaction between the user and the item, on the contrary, "0" means that there is no interaction between the user and the item. Then, the behaviors are sorted according to the chronological order of the actual interaction between users and items to generate the historical behavior sequence of users.

In this paper, the leave-one-out method is used to train and test the model [43]. Namely, the user's last behavior-producing item is taken as the true value, which is used as the test set. The last second-behavior-producing item is taken as the validation set, and all other remaining items are used as the training set. The advantage of the leave-one-out method is that it is not affected by the random sample division method and can use as large a sample as possible for training. It is suitable for sparse data sets.

4.1.2. Evaluation Metrics. This paper adopts two commonly used metrics in Top-N recommendation problem, namely hit rate (HR) and normalized discounted cumulative gain (NDCG) [29] to evaluate the recommendation performance of the model.

Hit rate (HR) is a common indicator for measuring recall rate, which can intuitively measure whether the predicted item exists in the first k items of the real list. The larger the hit rate (HR), the more accurate the recommendation. The calculation of HR is as follows:

$$\text{HR@}k = \frac{\text{Number of Hits @}k}{|GT|}, \tag{18}$$

where $|GT|$ represents all items in the test set, Number of Hits @k represents in the user's recommendation list, and the number of the top k items belonging to the test set.

NDCG is often used to evaluate the accuracy of ranking of recommendation results [44]. NDCG introduces a location influence factor to discount lower ranked recommendations. The calculation of NDCG is as follows:

$$\text{NDCG@}k = z_k \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}, \tag{19}$$

where z_k is the normalization factor, which is used to make the value of NDCG between 0 and 1 and r_i

Input: The behavior sequence S^u of user u
Output: The recommendation list result of user u, denoted as Y^u
(1) **for** u in length $(|U|)$**do**
(2) **if**length $(|S^u|) > L$**then**
(3) $S^u = S_L^u$
(4) **else**
(5) $S^u = S_S^u$ %Date preprocessing
(6) **end if**
(7) **end for**
(8) **for** u in S_L^u**do**
(9) $y_{i_0}^{u_L} = f(S_L^u: i)$ %Reverse prediction model training
(10) **end for**
(11) **for** u in S_S^u**do**
(12) $y_{i_0}^{u_S} = f(S_S^u: i)$ %Short sequence enhancement
(13) **end for**
(14) **for** u in $S_{\text{new}}^u \cup S_L^u$**do**
(15) Generate time interval matrix;
(16) Calculate time-aware self-attention model;
(17) Apply the point-wise feed-forward network and further processed;
(18) Calculate prediction and loss;
(19) **end for**
(20) **return** Y;

Algorithm 1: The sequential recommendation algorithm proposed in this paper.

Table 1: The information of the two data sets.

Data set	Number of			
	Users	Items	Actions	Average actions
ML-1M	6,040	3,900	1,000,209	165.6
AM-BE	51,369	19,369	225,509	4.4

represents the predicted correlation of the item at position i in the sequence. If the item is in the test set, $r_i = 1$ otherwise, $r_i = 0$.

The above two indicators can well reflect the performance of the recommendation list. This paper intercepts the top 10 of the recommendation list, namely $k = 10$, and uses HR@10 and NDCG@10 to evaluate the performance of the recommendation model.

Remark 1. The proposed method in this paper is based on the deep neural network and transfer learning technology, which needs more time in the process of the model training. However, the proposed model is trained offline and the computational time of the prediction is very fast. Thus, the computational complexity is not used as the evaluation metric in this study, which is a common way in the literature about the recommendation problem [19, 44].

4.1.3. Comparison Methods. To show the efficiency of the proposed methods (denoted by SeTsRec), various methods are used for comparison in this paper, including the recommendation method without considering the order, the classic order recommendation method, and the latest order recommendation method. In the experiments, the settings of these comparison methods are made by their optimal parameters according to the respective paper declarations. The comparison methods are listed as follows:

(a) POP [28]: POP is a simple baseline method that generates recommendation lists based on item popularity rankings, namely more popular items rank higher.

(b) BPR [45]: Bayesian personalized ranking method, which is a classic nonsequential recommendation method using matrix factorization.

(c) FPMC [46]: A sequential recommendation method that combines matrix factorization and Markov chains method.

(d) GRU4Rec+ [47]: An RNN-based deep sequential recommendation model for user sessions.

(e) Caser [48]: A CNN-based sequential recommendation method that captures higher order Markov chains by applying a convolution operation to the embedding matrix of the nearest term.

(f) SASRec [36]: One of the state-of-the-art sequential recommendation methods, which is the first method using a self-attention-based sequential recommendation model.

(g) TiSASRec [24]: A state-of-the-art sequential recommendation model that applies a multiorder

Table 2: The parameters of the proposed model and the experimental environment.

Model	Learning rate	0.001
	Momentum	0.9
	Dropout rate	0.2
	Batch size	128
	Maximum iterations	200
	Validation interval	20
	Regularization	0.00005
	Short-sequence threshold	20
	Maximum sequence length for ML-1M	70
	Maximum sequence length for AM-BE	30
	Latent dimension for ML-1M	50
	Latent dimension for AM-BE	20
	Pseudo-historical item for ML-1M	5
	Pseudo-historical item for AM-BE	15
Environment	Programming software	Python3.6
	Deep learning framework	Pytorch
	Computer system	Windows 10
	Cpu	E5-2620 v4
	RAM	32.0 GB
	Gpu	GeForce RTX 2080

attention mechanism to capture personal and item relatedness.

4.1.4. Other Settings. The experiments are conducted on a computer with Windows 10 system and the programming language used in the experiment is Python3.6. In this study, the model uses two self-attention blocks. Because different data sets have different sparsity, some parameters are different, such as the maximum sequence length (m) and the latent dimension (d). The setting of these parameters will be discussed in Section 5 for details. The parameters of the proposed deep network and the experimental environment are listed in Table 2.

4.2. Experimental Results and Analysis. Table 3 shows the experimental results of the proposed algorithm and all baseline methods on two different data sets with different indicators. The results in Table 3 show that the best recommendation effect is achieved by the proposed method in this paper, which prove the superiority of the model in this paper. The results are analyzed in details as follows:

(1) In most cases, the sequential recommendation methods FPMC, GRU4Rec+, and Caser outperform the nonsequential recommendation methods POP and BPR. This indicates the necessity of considering the order of user behavior lists in recommender systems. The user's behavior sequence order information can effectively characterize the user's preference change to a certain extent and can effectively improve the performance of the recommendation system.

(2) Compared with the three classical sequential recommendation methods, the latest attention-based SASRec and TiSASRec methods outperform all other baseline methods on the two different types of data sets, which indicates that the attention mechanism can effectively improve the performance of recommender systems.

Table 3: The experimental results.

Models	ML-1M		AM-BE	
	HR@10	NDCG@10	HR@10	NDCG@10
POP	0.4386	0.2389	0.3215	0.1758
BPR	0.5952	0.3421	0.2554	0.1523
FPMC	0.6182	0.3917	0.3771	0.2477
GRU4Rec+	0.6522	0.4334	0.3949	0.2556
Caser	0.7517	0.5011	0.4064	0.2547
SASRec	0.7929	0.5524	0.4185	0.2722
TiSASRec	0.8038	0.5706	0.4345	0.2818
Ours (SeTsRec)	**0.8127**	**0.5805**	**0.4754**	**0.3036**

(3) The algorithm SeTsRec proposed in this paper is improved on the basis of the existing algorithm. Through short-sequence enhancement and the use of an improved time-aware self-attention mechanism, it not only works well on dense data sets but also has the best results on sparse data sets. On the dense data set ML-1M, the HR@10 and NDCG@10 of the proposed method are improved by 1.1 % and 1.7 %, respectively, compared with the second best method TiSASRec. On the sparse data set AM-BE, the performance of the proposed method is improved by 9.4 % and 7.7 %, respectively, compared with TiSASRec. Compared with the baseline method, our model adopts an improved time-aware self-attention mechanism, which can adaptively adjust the item absolute position information and time-interval information to assign different weights in two different types of data sets.

5. Discussions

The results of the experiments in Section 4 show that the proposed model has better performance than that of the state of the art. The influence of the key parameters is discussed in

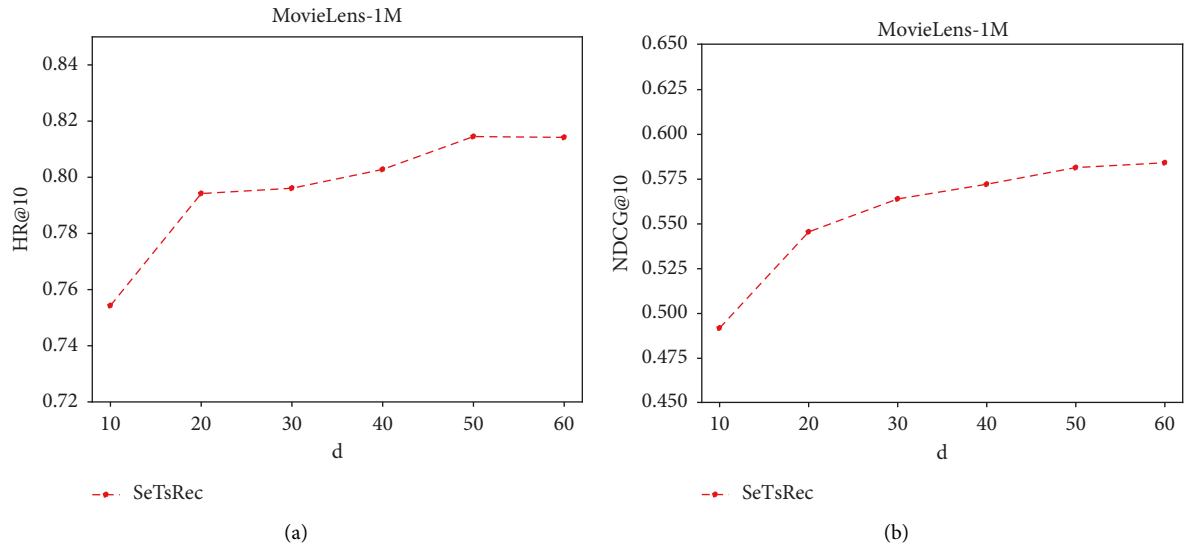


Figure 6: Experiment results of different latent dimension d on ML-1M. (a) Results on HR@10. (b) Results on NDCG@10.

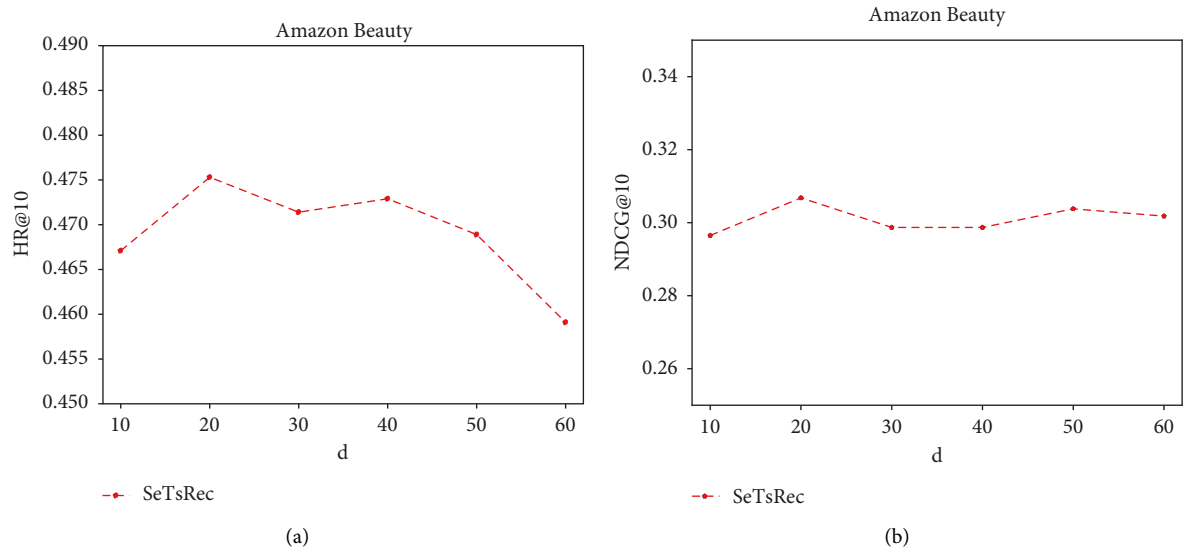


Figure 7: Experiment results of different latent dimension d on AM-BE. (a) Results on HR@10. (b) Results on NDCG@10.

this section. In addition, the ablation experiments are conducted in this section to further discuss the effectiveness of the improvement in the proposed model.

5.1. About the Latent Dimension. First, the influence of latent dimension d on the performance of the recommendation results of our model is discussed, and some experiments are conducted, where other hyperparameters are kept unchanged while the latent dimension d is changed within the range $[10, 60]$. The experimental results are shown in Figures 6 and 7.

It can be observed from Figure 6 (on the dense data set ML-1M) that the overall performance of the model improves with increasing potential dimensionality and tends to converge gradually, as the latent dimension increases. However, on the sparse data set AM-BE, the larger latent dimensions do not lead to better performance. The reason is that too many latent dimensions will lead to overfitting and thus degrade the model performance in a sparse data set. On the ML-1M data set, the algorithm in this paper tends to converge when $d \geq 50$. Considering the performance and time cost of the model, this paper sets the potential dimension $d = 50$ on the ML-1M data set and sets $d = 20$ on the AM-BE data set.

5.2. About the Maximum Sequence Length. Another important parameter of the proposed model is the maximum sequence length m. To discuss the influence of the maximum sequence length m of the input model on the performance of recommendation results, some experiments are conducted,

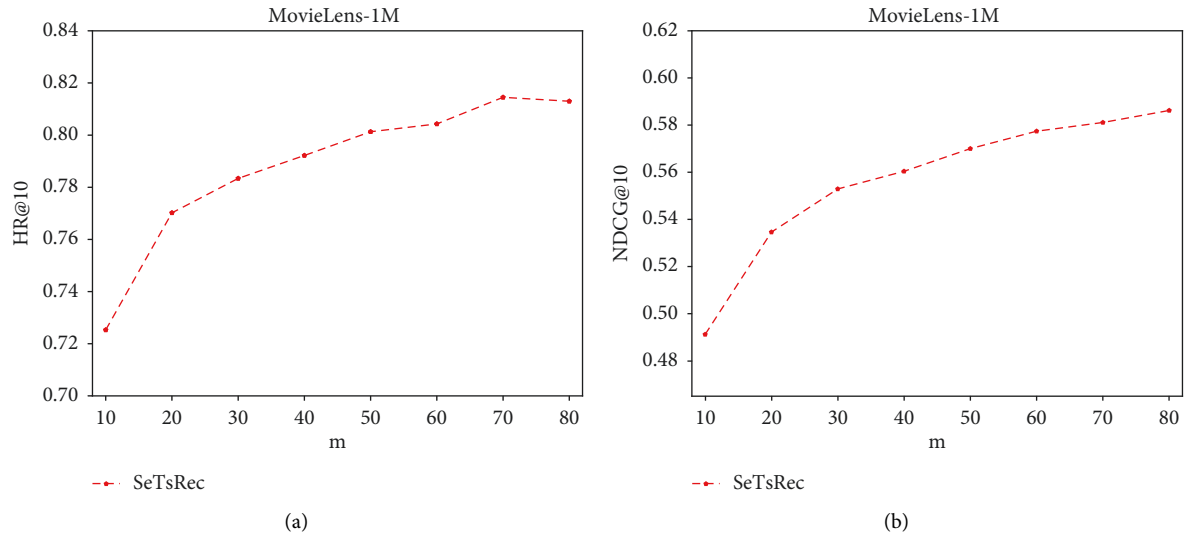


Figure 8: Experiment results of different maximum sequence length m on ML-1M. (a) Results on HR@10. (b) Results on NDCG@10.

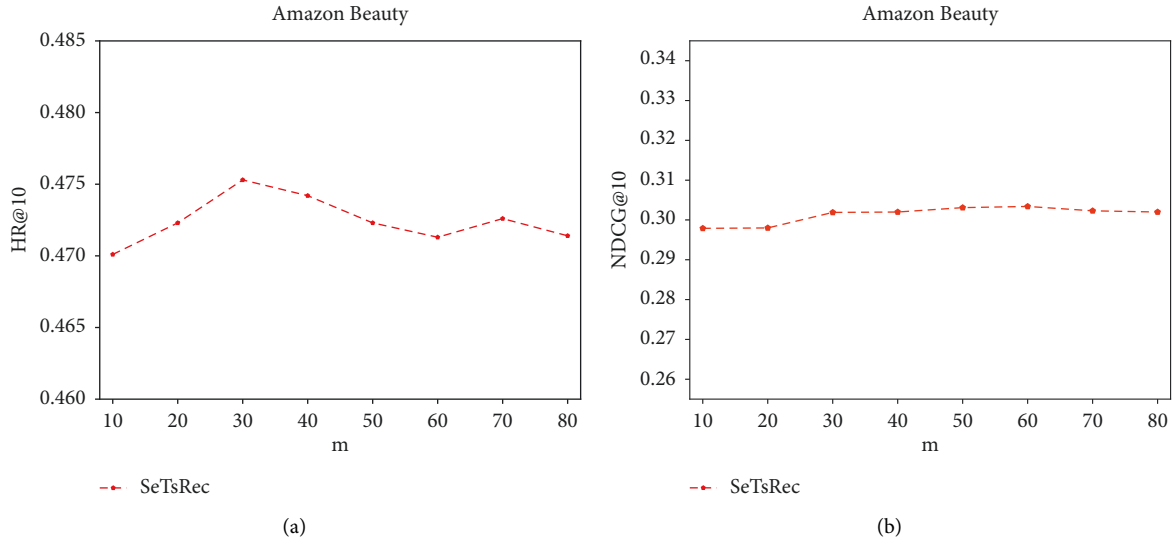


Figure 9: Experiment results of different maximum sequence length m on AM-BE. (a) Results on HR@10. (b) Results on NDCG@10.

where the maximum sequence length m is changed in the range [10,80], while keeping other hyperparameters unchanged. The experimental results are shown in Figures 8 and 9.

It can be observed from Figure 8 (on the dense data set ML-1M) that the model achieves satisfactory performance when the sequence length is $m \geq 70$. Therefore, under the consideration of balancing model performance and time cost, we set the maximum sequence length $m = 70$ on the ML-1M data set. On the sparse data set AM-BE, it can be observed that the model performance does not change much when m changes, this is because the average sequence length of the AM-BE data set is 4.4 (see Table 1), even after a certain degree of short-sequence enhancement, the longer sequence input does not provide more useful information, but will increase the time cost of the model. Therefore, the maximum sequence length is set as $m = 30$ on the data set AM-BE.

5.3. Ablation Experiments. This section discusses the impact of two major improvements in the proposed model, namely the short-sequence enhancement and time-aware self-attention mechanism. In these ablation experiments, the method based only on short-sequence enhancement is defined as SeTsRec-Se, and the method based only on time-aware self-attention is defined as SeTsRec-Ts, and they are compared with the existing SASRec method and our proposed algorithm SeTsRec. The experimental results are shown in Table 4, Figures 10 and 11. The results are analyzed in details as follows.

(1) On the dense data set ML-1M, the SeTsRec-Ts method outperforms the SeTsRec-Se method and the SASRec method. The experimental results show that the improvement of the model by the short-sequence enhancement method is limited. In this case, the

Table 4: Results of ablation experiment.

Methods	ML-1M		AM-BE	
	HR@10	NDCG@10	HR@10	NDCG@10
SASRec	0.7929	0.5524	0.4185	0.2722
SeTsRec-Se	0.7648	0.5297	0.4503	0.2907
SeTsRec-Ts	0.8038	0.5706	0.4345	0.2818
Ours (SeTsRec)	**0.8127**	**0.5805**	**0.4754**	**0.3036**

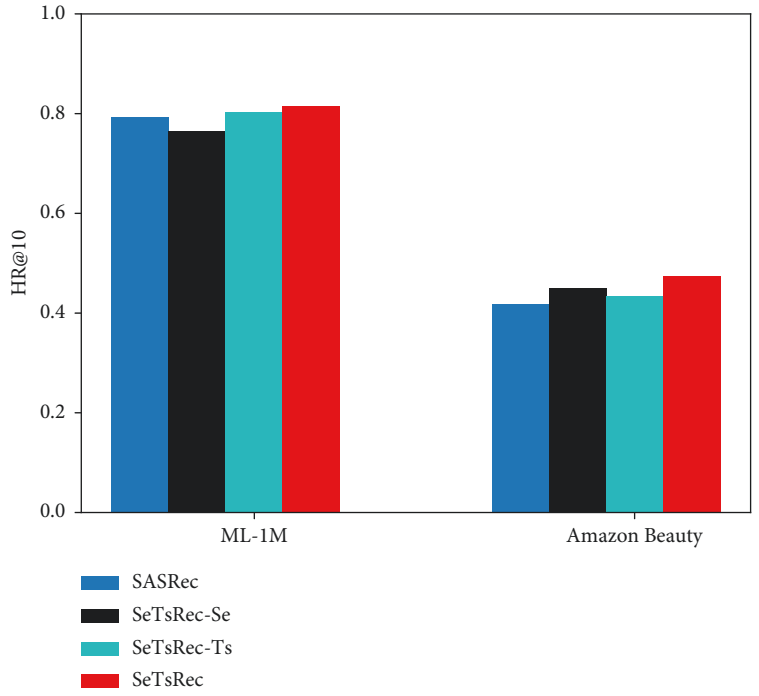


Figure 10: Results of ablation experiment on HR@10.

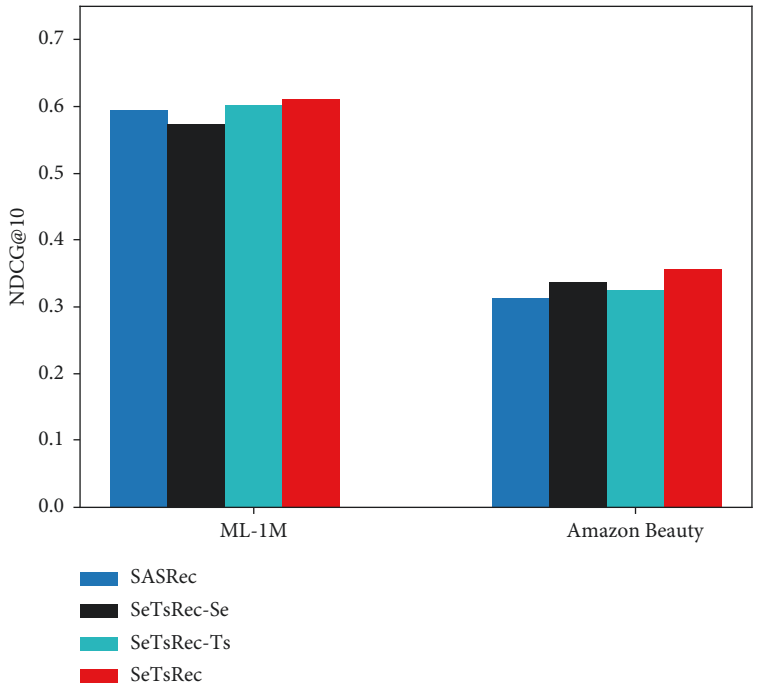


Figure 11: Results of ablation experiment on NDCG@10.

SeTsRec-Ts method can achieve good results, and its performance is improved by 1.4 %and 3.3 %, respectively, on the HR@10 and NDCG@10, compared with SASRec, which is close to the improved algorithm SeTsRec in this paper.

(2) On the sparse data set AM-BE, the SeTsRec-Se method is better than the SeTsRec-Ts method and the SASRec method, and its performance is improved by 3.6 % and 3.2 %, respectively, on the HR@10 and NDCG@10, compared with SeTsRec-Ts. The experimental results show that it is necessary to use the method of enhancing short sequences on sparse data sets. In addition, the model effect would not improve much if only the time-aware self-attentive mechanism approach is used. This is due to the large proportion of short-sequence users in the sparse data set, which limits the overall recommendation effect of the model.

(3) In summary, the proposed algorithm SeTsRec in this paper not only considers short-sequence enhancement to alleviate the problem of data sparsity but also combines the time-aware self-attention mechanism to fully consider the change of user preferences over time. Thus, it outperforms existing methods on both dense and sparse data sets.

6. Conclusions

This paper proposes a sequential recommendation algorithm based on improved short-sequence enhancement and temporal self-attention mechanism. The proposed algorithm first trains a reverse prediction model through the long-sequence users in the data set, to predict the reverse recommendation in the user sequence. Then, the model is transferred to short-sequence users, and pseudo-historical items of short-sequence users are generated to enhance short sequence. After enhancing short sequences, an improved time-aware self-attention model is adopted, which adaptively assigns different weights by combining the time interval information and absolute position information between items. It can deeply mine the changes of user preferences over time. Experimental results show that our method outperforms the existing sequential recommendation methods on different data sets. In the future, it can be considered to generate more accurate pseudo-historical items by improving the reverse prediction model to improve the recommendation effect further.

Data Availability

Publicly available data sets were analyzed in this study. These data can be found at https://files.grouplens.org/datasets/movielens/ml-1m.zip and https://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_5.json.gz.

Conflicts of Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61873086) and the Science and Technology Support Program of Changzhou (CE20215022).

References

[1] S. Ahmadian, N. Joorabloo, M. Jalili, and M. Ahmadian, "Alleviating data sparsity problem in time-aware recommender systems using a reliable rating profile enrichment approach," *Expert Systems with Applications*, vol. 187, 2022.

[2] Z. Sun, Q. Guo, and J. Yang, "Research commentary on recommendations with side information: a survey and research directions," *Electronic Commerce Research and Applications*, vol. 37, 2019.

[3] A. Pujahari and D. Singh Sisodia, "Handling dynamic user preferences using integrated point and distribution estimations in collaborative filtering," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, In press, 2022.

[4] F. Tahmasebi, M. Meghdadi, S. Ahmadian, and K. Valiallahi, "A hybrid recommendation system based on profile expansion technique to alleviate cold start problem," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2339–2354, 2021.

[5] Xu Chen, H. Xu, and Y. Zhang, "Sequential recommendation with user memory networks," in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining, WSDM 2018*, pp. 108–116, Marina Del Rey, CA, USA, February 2018.

[6] E. Mutabazi, J. Ni, G. Tang, and W. Cao, "A review on medical textual question answering systems based on deep learning approaches," *Applied Sciences*, vol. 11, no. 12, p. 5456, 2021.

[7] L. Quijano-Sánchez, I. Cantador, M. E. Cortés-Cediel, and O. Gil, "Recommender systems for smart cities," *Information Systems*, vol. 92, Article ID 101545, 2020.

[8] J. Ni, Yu Cai, G. Tang, and Y. Xie, "Collaborative filtering recommendation algorithm based on TF-IDF and user characteristics," *Applied Sciences*, vol. 11, no. 20, p. 9554, 2021.

[9] H. Jia and J. Yang, "Research on joint ranking recommendation model based on Markov chain," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 6, Article ID e5191, 2020.

[10] H. Kang, S. Yang, J. Huang, and J. Oh, "Time series prediction of wastewater flow rate by bidirectional lstm deep learning," *International Journal of Control, Automation and Systems*, vol. 18, no. 12, pp. 3023–3030, 2020.

[11] J. Park and D.-J. Jung, "Deep convolutional neural network architectures for tonal frequency identification in a lofargram," *International Journal of Control, Automation and Systems*, vol. 19, no. 2, pp. 1103–1112, 2021.

[12] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[13] A. Yengikand, M. Meghdadi, and S. Ahmadian, "Deep representation learning using multilayer perceptron and stacked autoencoder for recommendation systems," in *Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2485–2491, Melbourne, Australia, October 2021.

[14] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: a survey and new perspectives," *ACM Computing Surveys*, vol. 52, no. 1, 38 pages, 2020.

[15] M. Ahmadian, M. Ahmadi, and S. Ahmadian, "Integration of deep sparse autoencoder and particle swarm optimization to develop a recommender system," in *Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2021*, pp. 2524–2530, Melbourne, Australia, October 2021.

[16] L. Huang, M. Fu, F. Li, H. Qu, Y. Liu, and W. Chen, "A deep reinforcement learning based long-term recommender system," *Knowledge-Based Systems*, vol. 213, Article ID 106706, 2021.

[17] C. Xu, J. Feng, P. Zhao, F. Zhuang, D. Wang, and Y.V. Liu, "Long-and short-term self-attention network for sequential recommendation," *Neurocomputing*, vol. 423, pp. 580–589, 2021.

[18] F. Sun, J. Liu, and J. Wu, "Bert4rec: sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, Beijing, China, November 2019.

[19] X. Wang, Y. Sheng, H. Deng, and Z. Zhao, "Top-n-targets-balanced recommendation based on attentional sequence-to-sequence learning," *IEEE Access*, vol. 7, pp. 120262–120272, 2019.

[20] Q. Tan and F. Liu, "Recommendation based on users' long-term and short-term interests with attention," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–13, 2019.

[21] C. Zhao, J. You, X. Wen, and X. Li, "Deep bi-lstm networks for sequential recommendation," *Entropy*, vol. 22, no. 8, p. 870, 2020.

[22] Z. Liu, Z. Fan, Yu Wang, and P. S. Yu, "Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1608–1612, Canada, July 2021.

[23] S. Ahmadian, M. Ahmadian, and M. Jalili, "A deep learning based trust- and tag-aware recommender system," *Neurocomputing*, vol. 488, 2021.

[24] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the 13th International Conference On Web Search And Data Mining*, pp. 322–330, Houston, TX, USA, February 2020.

[25] R. He and J. McAuley, "Fusing similarity models with Markov chains for sparse sequential recommendation," in *Proceedings of the 2016 IEEE 16th International Conference On Data Mining (ICDM)*, pp. 191–200, IEEE, Barcelona, Spain, December 2016.

[26] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, "A survey on theories and applications for self-driving cars based on deep learning methods," *Applied Sciences*, vol. 10, no. 8, p. 2749, 2020.

[27] S. J. Lee, H. Choi, and S. S. Hwang, "Real-time depth estimation using recurrent cnn with sparse depth cues for slam system," *International Journal of Control, Automation and Systems*, vol. 18, no. 1, pp. 206–216, 2020.

[28] J. Zhang, C. Ma, X. Mu, P. Zhao, C. Zhong, and A. Ruhan, "Recurrent convolutional neural network for session-based recommendation," *Neurocomputing*, vol. 437, pp. 157–167, 2021.

[29] Ke Sun, T. Qian, Xu Chen, and M. Zhong, "Context-aware seq2seq translation model for sequential recommendation," *Information Sciences*, vol. 581, pp. 60–72, 2021.

[30] W. Yuan, H. Wang, X. Yu, N. Liu, and Z. Li, "Attention-based context-aware sequential recommendation model," *Information Sciences*, vol. 510, pp. 122–134, 2020.

[31] W. Zhao, B. Wang, M. Yang et al., "Leveraging long and short-term information in content-aware movie recommendation via adversarial training," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 4680–4693, 2020.

[32] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "Attentionfgan: infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1383–1396, 2021.

[33] P. Shi, X. Fang, J. Ni, and J. Zhu, "An improved attention-based integrated deep neural network for pm2.5 concentration prediction," *Applied Sciences*, vol. 11, no. 9, p. 4001, 2021.

[34] I. Lopez-Gazpio, M. Maritxalar, M. Lapata, and E. Agirre, "Word n-gram attention models for sentence similarity and inference," *Expert Systems with Applications*, vol. 132, pp. 1–11, 2019.

[35] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[36] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proceedings of the 2018 IEEE International Conference On Data Mining (ICDM)*, pp. 197–206, IEEE, Singapore, November 2018.

[37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[38] K.-H. Ahn and J.-B. Song, "Image preprocessing-based generalization and transfer of learning for grasping in cluttered environments," *International Journal of Control, Automation and Systems*, vol. 18, no. 9, pp. 2306–2314, 2020.

[39] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 3947–3986, 2020.

[40] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems, NIPS 2013*, Lake Tahoe, NV, USA, December 2013.

[41] J. Mills, J. Hu, and G. Min, "Multi-task federated learning for personalised deep neural networks in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 630–641, 2022.

[42] J. David Camacho, C. Villaseñor, Y. Alma, C. Lopez-Franco, and N. Arana-Daniel, "Kadam: using the kalman filter to improve Adam algorithm," in *Proceedings of the 24th Iberoamerican Congress on Pattern Recognition, CIARP 2019*, pp. 429–438, Havana, Cuba, October 2019.

[43] C. Lonjarret, R. Auburtin, C. . Robardet, and M. Plantevit, "Sequential recommendation with metric models based on frequent sequences," *Data Mining and Knowledge Discovery*, vol. 35, no. 3, pp. 1087–1133, 2021.

[44] X. Wang, Q. Tan, and L. Zhang, "A deep neural network of multi-form alliances for personalized recommendations," *Information Sciences*, vol. 531, pp. 68–86, 2020.

[45] Y.-C. Lee, T. Kim, J. Choi, X. He, and S. W. Kim, "M-bpr: a novel approach to improving bpr for recommendation with multi-type pair-wise preferences," *Information Sciences*, vol. 547, pp. 255–270, 2021.

[46] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized Markov chains for next-basket recommendation," in *Proceedings of the 19th international*

conference on World Wide Web, pp. 811–820, Raleigh, NC, USA, April 2010.

[47] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 843–852, Torino, Italy, October 2018.

[48] D. Yang, J. Zhang, S. Wang, and X. D. Zhang, "A time-aware cnn-based personalized recommender system," *Complexity*, vol. 2019, Article ID 9476981, 11 pages, 2019.

Hindawi
Complexity
Volume 2022, Article ID 8765949, 9 pages
https://doi.org/10.1155/2022/8765949

Research Article

Relationship between Urban Innovation Capability and Energy Utilization Efficiency: An Empirical Study of 281 Prefecture-Level Cities in China

Wanshu Wu [1] **and Kai Zhao** [2]

[1]*Qingdao University of Technology, Qingdao, China*
[2]*Qingdao University, Qingdao, China*

Correspondence should be addressed to Kai Zhao; kzhao_kai@126.com

Received 12 July 2022; Accepted 9 September 2022; Published 27 September 2022

Academic Editor: Andrea Murari

Following a dynamic nonlinear perspective, this study explores the relationship between urban innovation capability and energy utilization efficiency by employing the Panel Vector Autoregression (PVAR) and Dynamic Panel Threshold Regression (DPTR) methods. Using the 2003–2020 panel data of 281 prefecture-level cities in China, this study confirms that energy utilization efficiency improves owing to the improvement of urban innovation capability. Depending on the characteristics of the city, such as population density, industrial structure, and environmental pollution, high energy utilization efficiency in the early stages of city development may help or hinder the improvement of energy utilization efficiency in the later stages. The enhancement in urban innovation capability has failed to improve energy utilization efficiency and has adversely affected cities with a low population density or weak secondary industrial foundation. However, in cities with a high population density or proportion of secondary industry, the improvement in innovation capability significantly increases the efficiency of energy utilization. In addition, the positive effect that urban innovation capability has on energy utilization efficiency is higher in low-pollution cities than in high-pollution cities.

1. Introduction

Energy consumption is an important factor in the economic development and social progress of China. Given the increasing total economic scale, the demand for and dependence on energy in China are rising [1]. The latest data from the BP World Energy Statistics Yearbook highlights that in 2018, the total primary energy consumption in China is equivalent to 3273.5 million tons of oil, the highest in the world. Moreover, according to the "China Energy Supply and Demand Report," the total energy consumption of China amounts to 4.64 billion tons of standard coal, accounting for 23.6% of the total global primary energy consumption, and has ranked first worldwide for 10 consecutive years. The environmental deterioration in China owing to excessive energy consumption coexists with the energy tension caused by economic development. In addition, the increasingly severe energy situation entails a greater need for energy utilization efficiency, and improving the efficiency of energy utilization has become the focus of economic development in China at this stage [2]. However, compared with the top countries regarding economic aggregate, energy consumption per unit of the Gross Domestic Product (GDP) is 2.14 times in the United States, 2.63 in Japan, 2.97 in Germany, 3.53 in the United Kingdom, and 2.75 in France. This implies that the economy in China is still supported by a large amount of energy consumption, and there is still a large gap between China and the developed countries regarding energy utilization efficiency [3].

The exponential growth of the economy and the limited development of resources have elevated the transformation of the "factor-driven" to the "innovation-driven". Thus, technological innovation has become a vital means for countries and cities to solve economic problems and occupy

development opportunities under the wave of the new technological revolution [4, 5]. Recent findings have confirmed that the city serves as the main location for scientific and technological innovation activities, and the increase in innovation capability is helpful in improving energy efficiency [6]. Improving energy utilization efficiency can also improve urban innovation capabilities [7]. However, does this conclusion apply to Chinese cities? Does energy utilization efficiency affect urban innovation capability in China? Does urban innovation capability affect energy utilization efficiency? Or do they interact? Is the relationship between the two forced or driven? Will this relationship change with the changes in urban population density, industrial structure, environmental pollution, and other factors? There are numerous questions that are not yet settled. Against this background, clarifying the dynamic relationship and mechanism between urban innovation capability and energy utilization efficiency in China is not only beneficial to ensuring national energy security and transforming the mode of economic growth, but also conducive to the sustainable and coordinated development of scientific and technological innovation and new urbanization.

As an important issue in the field of energy economics, energy utilization efficiency has been widely concerned by numerous scholars [8]. The connotation of energy utilization efficiency gradually extends, from the initial single-factor energy utilization efficiency to the total-factor energy utilization efficiency based on the traditional DEA model [9], from the static energy utilization efficiency to the dynamic total-factor energy utilization efficiency based on the Malmquist index model [10], and from only focusing on economic development to considering environmental pollution [11] and energy utilization efficiency at the enterprise level [12]. Similarly, urban innovation capability, as an important issue in regional economics, also attracts attention. Previous studies have discussed the definition and related concepts of urban innovation capability from the perspectives of innovation environment and resource integration [13, 14]. Moreover, the measurement standards and evaluation systems of urban innovation capability are extensively and fully discussed [15, 16], which triggers a dispute between a single indicator and an indicator system. However, Huang et al. [17] put regional innovation capability and energy utilization efficiency in China into a research framework and examined the coupling relationship between them from the perspective of spatial and temporal coordination. However, following the extant literature, most discussions on energy utilization efficiency and urban innovation capability exist independently, and few studies have investigated the relationship between the two, especially the dynamic nonlinear relationship.

The main contribution of this study is reflected in the following three aspects: first, from the perspective of dynamic nonlinearity, the dynamic correlation and mechanism between urban innovation capability and energy utilization efficiency are discussed. Second, the combined method of the Panel Vector Autoregression (PVAR) and the Dynamic Panel Threshold Regression (DPTR) is helpful in accurately identifying the dynamic causal relationship between urban innovation capability and energy utilization efficiency and clarifying the mechanism of action, as well as examining the dynamic nonlinear relationship between urban innovation capability and energy utilization efficiency under different constraints. Finally, this study uses nighttime lighting data, which have been widely used in the field of economic research recently; it measures the energy consumption of various prefecture-level cities following the idea that the brighter the night light is, the greater the total energy consumption is, solving the shortcomings of existing research in time span and urban measurement.

The remainder of the paper is structured as follows: Section 2 explains the research design and method; Section 3 introduces the data source and variable definition; Sections 4 and 5 discuss the PVAR system and DPTR analyses, respectively; and Section 6 concludes the study.

2. Methodology

2.1. PVAR System. PVAR can treat all variables as endogenous systems and examine the lagged terms of each variable, reflecting the interaction between variables. This method can capture individual differences and common shocks to different cross-sections by introducing individual effect and time-point effect variables, respectively, adding to the advantages of Vector Autoregression (VAR) models and panel data models. It can not only solve the problem of endogeneity but also effectively characterize the shock response and variance decomposition among system variables. We can explore the dynamic relationship between urban innovation capability and energy utilization efficiency as well as the direct, strengthening, feedback, and other dynamic interaction effects by constructing the PVAR system.

The PVAR system for analysis comprises the following main steps: (1) construct a Generalized Method of Moments (GMM) estimation to obtain the regression relationship between variables; (2) determine the influence of orthogonalization on other variables in the system by analyzing the impulse-response function; and (3) obtain the variance decomposition results in the prediction period and measure the contribution of each variable using the variance analysis. Because the estimation of the PVAR system is based on the fixed-effect dynamic panel model, the intragroup mean difference method should be used before the GMM estimator to eliminate the time effect. Subsequently, to eliminate the individual effect, the onward mean difference method should be employed. The PVAR system is expressed as follows:

$$\begin{aligned} Y_{it} = {} & Y_{it-1}A_1 + Y_{it-2}A_2 + \cdots + Y_{it-p+1}A_{p-1} + Y_{it-p}A_p \\ & + X_{it}B + f_i + \mu_t + \varepsilon_{it}, \end{aligned} \tag{1}$$

where $i \in \{1, 2, \ldots, N\}$ represents the prefecture-level cities in China; $t \in \{1, 2, \ldots, T\}$ indicates the year; Y_{it} is a $(1 \times k)$ vector of dependent variables; X_{it} is a $(1 \times l)$ vector of exogenous covariates (control variables); f_i represents an unobservable intercept effect, and this fixed effect can be eliminated using the forward difference Helmert transformation method (the forward difference Helmert

transformation method avoids the orthogonality between the lag regression and difference terms of the instrumental variable by removing the forward mean, so that the measurement test results can be more accurate); μ_t denotes the time effect; and ε_{it} is the random error term, which has the following characteristics: $E(\varepsilon_{it}) = 0$ and $E(\varepsilon_{it}'\varepsilon_{it}) = \Sigma$, and $E(\varepsilon_{im}'\varepsilon_{in}) = 0$.

2.2. DPTR. Traditional panel threshold regression focuses on static effects and requires strong exogenous control variables [18]. However, strong exogenous conditions are often difficult to meet in the real world. Therefore, Seo and Shin [19] extended the traditional panel threshold model to the dynamic model, and the First Difference Generalized Method of Moments (FD-GMM) is employed to estimate it in solving the endogenous problem in the DPTR model. The specific form of the DPTR model is as follows:

$$y_{it} = (1, x_{it}')\phi_1 \cdot I\{q_{it} \le \gamma\} + (1, x_{it}')\phi_2 \cdot I\{q_{it} > \gamma\} + \varepsilon_{it}. \quad (2)$$

The first-order difference form of (2) can be expressed as follows:

$$\Delta y_{it} = \beta' \Delta x_{it} + \delta' X_{it} 1_{it}(\gamma) + \Delta\varepsilon_{it}, \quad (3)$$

where $\underset{k_1\times 1}{\beta} = (\phi_{12}, \ldots, \phi_{1,k_1+1})\prime$, $\underset{(k_1+1)\times 1}{\delta} = \phi_2 - \phi_1$, $\underset{2\times(1+k_1)}{X_{it}} = \begin{pmatrix} (1, x_{it}') \\ (1, x_{i,t-1}') \end{pmatrix}$, and $\underset{2\times 1}{1_{it}(\gamma)} = \begin{pmatrix} 1\{q_{it} > \gamma\} \\ -1\{q_{i,t-1} > \gamma\} \end{pmatrix}$. Making $\theta = (\beta', \delta', \gamma')$, and supposing θ is a compact set, $\Theta = \Phi\times\Gamma \subset \mathbb{R}^k$, where $k = 2k_1 + 2$. Making $\Gamma = [\underline{\gamma}, \overline{\gamma}]$, $\underline{\gamma}$ and $\overline{\gamma}$ represent two percentiles of the threshold variables, respectively. Owing to the correlation between the regression element and individual effect, the parameter estimation obtained using the ordinary least squares regression directly on (3) is biased. Therefore, we need to find a $l \times 1$ dimensional tool variable $(\underset{zit_0}{'}, \ldots, z_{iT}')\prime$ that satisfies $E(\underset{zit_0}{'}\Delta\varepsilon_{it_0}, \ldots, z_{iT}'\Delta\varepsilon_{iT})\prime = 0$ for any $2 < t_0 \le T$ and $l \ge k$.

Because the model allows the endogeneity of threshold variable q_{it}, it is $E(q_{it}\Delta\varepsilon_{it}) \ne 0$. Therefore, q_{it} does not belong to the set of instrumental variables $\{z_{it}\}_{t=t_0}^{T}$, and the sample moment conditions of the following one-dimensional column vectors are considered:

$$\begin{aligned} \overline{g}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} g_i(\theta), \underset{l\times 1}{g_i(\theta)} \\ &= \begin{pmatrix} z_{it_0}\left(\Delta y_{it_0} - \beta'\Delta x_{it_0} - \delta'\,\underset{Xit_0}{'} 1_{it_0}(\gamma)\right) \\ \vdots \\ z_{iT}\left(\Delta y_{iT} - \beta'\Delta x_{iT} - \delta' X_{iT}' 1_{iT}(\gamma)\right) \end{pmatrix}. \end{aligned} \quad (4)$$

Suppose that if and only if $\theta = \theta_0$, $E(g_i(\theta)) = 0$. Thus, making $g_i = g_i(\theta_0) = (\underset{zit_0}{'}\Delta\varepsilon_{it_0}, \ldots, z_{iT}'\Delta\varepsilon_{iT})\prime$ and $\Omega = E(g_i g_i')$, where Ω is assumed to be a positive definite. For a positive definite matrix W_n and $W_n \longrightarrow^{p} \Omega^{-1}$, making $\overline{J}_n(\theta) = \overline{g}_n(\theta)\prime W_n \overline{g}_n(\theta)$, where

$$W_n = \begin{pmatrix} \frac{2}{n}\sum_{i=1}^{n} z_{it_0}\,\underset{zit_0}{'} & \frac{-1}{n}\sum_{i=1}^{n} z_{it_0}\,\underset{zi,t_0+1}{'} & 0 & \cdots \\ \frac{-1}{n}\sum_{i=1}^{n} z_{i,t_0+1}\,\underset{zit_0}{'} & \frac{2}{n}\sum_{i=1}^{n} z_{i,t_0+1}\,\underset{zi,t_0+1}{'} & \ddots & \ddots \\ 0 & \ddots & \ddots & \frac{-1}{n}\sum_{i=1}^{n} z_{i,T-1} z_{iT}' \\ \vdots & \ddots & \frac{-1}{n}\sum_{i=1}^{n} z_{iT} z_{i,T-1}\,\underset{}{'} & \frac{2}{n}\sum_{i=1}^{n} z_{iT} z_{iT}' \end{pmatrix}, \quad (5)$$

θ estimates can be derived from $\hat{\theta} = \arg\min_{\theta\in\Theta} \overline{J}_n(\theta)$. For fixed γ, let $\overline{g}_{1n} = 1/n\sum_{i=1}^{n} g_{1i}$, $\overline{g}_{2n}(\gamma) = 1/n\sum_{i=1}^{n} g_{2i}(\gamma)$, where $\underset{l\times 1}{g_{1i}} = \begin{pmatrix} z_{it_0}\Delta y_{it_0} \\ \vdots \\ z_{iT}\Delta y_{iT} \end{pmatrix}$, $\underset{l\times(k-1)}{g_{2i}(\gamma)} = \begin{pmatrix} z_{it_0}(\Delta x_{it_0}, 1_{it_0}(\gamma)\prime X_{it_0}) \\ \vdots \\ z_{iT}(\Delta x_{iT}, 1_{iT}(\gamma)\prime X_{iT}) \end{pmatrix}$, then for given γ, β, and δ, the estimators are expressed as the following equation:

$$\begin{aligned} \left(\hat{\beta}(\gamma)', \hat{\delta}(\gamma)'\right)' &= \left(\overline{g}_{2n}(\gamma)' W_n \overline{g}_{2n}(\gamma)\right)^{-1} \overline{g}_{2n}(\gamma)' W_n \overline{g}_{1n} \\ W_n &= \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}\,\underset{i}{'} - \frac{1}{n^2}\sum_{i=1}^{n} \hat{g}_i \sum_{i=1}^{n} \hat{g}\,\underset{i}{'}\right)^{-1} \\ \hat{g}_i &= \left(\Delta\hat{\varepsilon}_{it_0}\,\underset{zit_0}{'}, \ldots, \Delta\hat{\varepsilon}_{iT} z_{iT}'\right)\prime. \end{aligned} \quad (6)$$

Returning $\hat{\beta}(\gamma)$ and $\hat{\delta}(\gamma)$ to the objective function yields an estimate of θ: $\hat{\gamma} = \arg\min_{\gamma\in\Gamma} J_n(\gamma)$, $(\hat{\beta}, \hat{\delta})' = (\hat{\beta}(\gamma)', \hat{\delta}(\gamma)')'$.

3. Data

This study uses panel data from 281 prefecture-level cities in China from 2003 to 2020. The relevant data on the regional economy, industrial structure, and urban environmental pollution in various prefecture-level cities stem from the annual "China Statistical Yearbook" and "China Urban Statistical Yearbook." The data on the invention patent authorization in various prefecture-level cities are obtained from the official websites of the State Intellectual Property Office. The energy consumption of prefecture-level cities is calculated based on the nighttime light data that have been widely used in recent economic research [20–22]. The idea is that the brighter the night light is, the greater the total energy consumption. The nighttime lighting data are obtained from the "Global Night-time Light Database." This database was developed based on the Defense Meteorological Satellite Program (the DMSP global nighttime lighting data are available at "https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html"). The nighttime light data include cloudless observation frequency, average light image, and stable light image. Because the stable lighting image data contain relatively stable lighting in cities and towns, this study selects the stable lighting image data as the basic data night-light image data and the Visible Infrared Imaging Radiometer Suite (VIIRS night lighting data are available at "https://ncc.nesdis.noaa.gov/VIIRS/"). night light image data of the National Oceanic and Atmospheric Administration of the United States. These data reflect the nighttime lighting data of the cities and counties in China (The National Geophysical Data Center (NGDC) of the United States conducts a series of noise processing on the basic data, such as eliminating the influence of nighttime clouds, short-term fires, aurora, and lightning, so the processed data can truly reflect the energy consumption of human beings). We average the nighttime light data for each year in the research window period to ensure that nighttime light data cover all prefecture-level cities in China from the time and space dimensions. In addition, we convert the brightness of the light into a digital number (DN). The DN value range of each raster is 0–63 (63 is the saturation value of the data). The spatial dimension covers the longitude from 135°degrees east to 73°degrees west and the latitude from 3°degrees north to 54°degrees north.

The core variable energy utilization efficiency (energy) is measured by the logarithm of the per capita GDP of a prefecture-level city divided by the total energy consumption of the prefecture-level city (i.e., the reciprocal of energy consumption per unit GDP). The higher the value is, the higher the energy utilization efficiency is. The main variable, urban innovation capability (*inno*), is measured by the total number of invention patents in the prefecture-level cities. Moreover, the urban population density (density) is obtained by dividing the population of the prefecture-level cities by administrative area, thereby characterizing the differential impact of the scale of urban human activities. The industrial structure (struc) is measured by the proportion of the added value of the secondary industry in the regional GDP, thereby characterizing the overall industrial structure of the city. The degree of urban environmental pollution (pollu) is measured by the sulfur dioxide emissions of the prefecture-level cities. The descriptive statistics of the aforementioned variables are presented in Table 1.

4. PVAR Analysis

4.1. Model Estimation. The nonstationary problem of the variables often leads to the phenomenon of "pseudoregression" in the analysis, making the regression results deviate or even invalid. Therefore, we use Levin–Lin–Chu (LLC), Harris–Tzavalis (HT), and Fisher-ADF methods to examine whether the core variables have panel unit roots to ensure the robustness of the test results. Table 2 reports that the test results of the three methods reject the hypothesis that the variables are nonstationary, and it can be considered that the two core variables of energy utilization efficiency and urban innovation capability are stationary, which is suitable for the PVAR system analysis.

The orthogonal transformation between variables and lagged regression coefficients with the help of the Helmert method and the optimal lag order of the PVAR system is selected according to the information criteria, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the quasi-information criterion (QIC). When the lag term is 1, the BIC reaches the minimum, and when the order of the lag term is 2, the AIC and QIC reach the minimum (Table 3). Following the principle of "minority obeys majority," a PVAR system with lag order 2 is constructed.

In Table 3, the energy equation estimation results (Column 1) suggest that the early energy utilization efficiency significantly affects the later energy utilization efficiency, and the early urban innovation capability is also conducive to improving the later energy utilization efficiency. However, the estimation results of inno equation (Column 2) reveal that the estimation coefficient of energy utilization efficiency lagging one period is negative and does not exhibit aboriginality, indicating that the urban energy utilization efficiency of the previous period cannot significantly improve the urban innovation capability of the latter period and may even inhibit the urban innovation capability. The early urban innovation capability will be beneficial to the later innovation capability, which has certain "inertia" characteristics.

4.2. Impulse Response and Variance. The stability of the PVAR (2) model is first tested before analyzing the impulse response function and variance decomposition. Table 4 and Figure 1 demonstrate that the absolute values of the real and imaginary parts of the eigenvalues are all within the range of [0, 1]. Therefore, the PVAR model is considered stable.

Table 1: Descriptive statistics of variables.

Name	Symbol	Mean	SD	Min	Max	Obs
Energy utilization efficiency	energy	0.1259	0.8483	−2.1271	4.1374	5058
Innovation capability	inno	3.7893	1.9326	0	10.7377	5058
Population density	density	572.1839	313.0527	5.2016	2666.9483	5058
Industrial structure	struc	0.4850	0.1099	0.0900	0.9097	5058
Environmental pollution	pollu	56458.8437	58015.8401	1.9756	683170.7138	5058

Table 2: Unit root test for core variables.

Variable	Method			Conclusion		
	LLC	HT	ADF	LLC	HT	ADF
energy	−9.1795***	0.5182**	1140.7838***	Steady	Steady	Steady
inno	−9.3768***	0.7563***	684.5540***	Steady	Steady	Steady

Note. ***, **, and * represent the significance levels at 1%, 5%, and 10%, respectively.

Table 3: Estimated results of the PVAR system.

	(1) energy	(2) inno
Coefficients		
L.energy	0.5594*** (0.1523)	−0.4483 (0.5282)
*L*2.energy	0.3021*** (0.1031)	0.2933 (0.2772)
L.inno	0.0010*** (0.0002)	0.7452*** (0.0273)
*L*2.inno	0.0036* (0.0020)	0.0409*** (0.0154)
Control variables	Yes	Yes

Lag order	AIC	BIC	QIC
1	19.2748	−101.4501	−18.1671
2	16.5196	−80.0603	−24.0836
3	19.3278	−53.1073	−6.6874

Note. ***, **, and * represent the significance levels at 1%, 5%, and 10%, respectively; "L" and "L2" represent lag order 1 and lag order 2, respectively; standard error is presented in parentheses.

Table 4: Stability test of the PVAR (2) model.

Eigenvalue		Module
Real	Imaginary	
0.6744	0	0.6744
0.2816	0.4673	0.5456
0.0.2816	−0.4673	0.5456
0.0670	0	0.0670

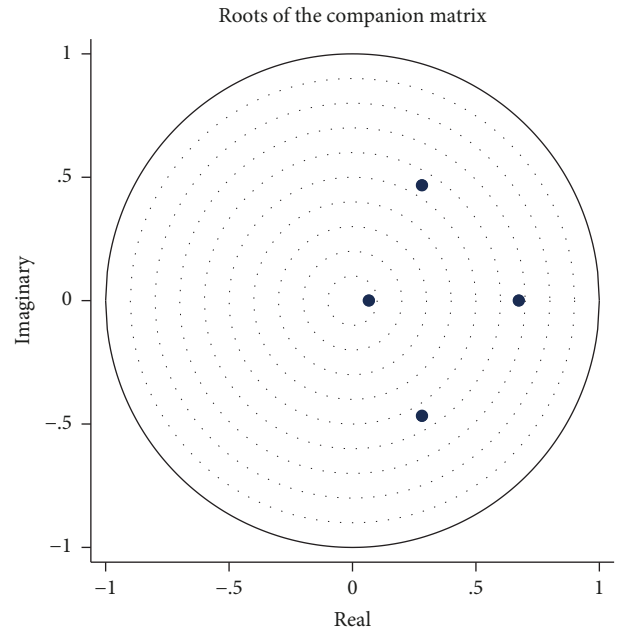


Figure 1: Roots of the companion matrix.

The impulse response function describes the response of an endogenous variable to an error; that is, the trajectory of the impact of a standard deviation of the random disturbance term on the current and future values of other variables. It can intuitively describe the dynamic interaction between energy utilization efficiency and urban innovation capability and determine the time lag relationship between variables. To intuitively describe the dynamic delay relationship between the variables in the system, we give each variable a standard deviation of the impact and use the Monte Carlo method to simulate 300 times, obtaining the impact of each variable on the 0–20 periods after each variable. The curve of the impulse response function of two variables is illustrated in Figure 2. The horizontal axis represents the response period of the shock response, and the maximum lag period is 20. The vertical axis represents the corresponding degree of the variable to the shock. The shadow part represents the 95% confidence interval, and the middle real line represents the size of the shock response in each period.

There are three kinds of dynamic interaction effects in the PVAR system: direct, reinforcement, and feedback effects. First, the direct effect, which is the lag term of urban innovation capability variables on energy efficiency, can be concerned with the first line and the second column of the impulse response in Figure 2. In the face of an orthogonal impact of urban innovation capability (inno), the overall response of energy utilization efficiency shows an inverted "U-shaped" trend. In the first three periods, improving urban innovation capability can quickly improve energy utilization efficiency, whereas, from the fourth period, the positive effect gradually decreases and approaches 0. This implies that urban innovation capability has a positive effect on energy utilization efficiency, and it will significantly improve energy utilization efficiency in the early stages. However, its effect will gradually weaken with the continuous renewal of urban development and technological innovation. Second, the strengthening effect is the

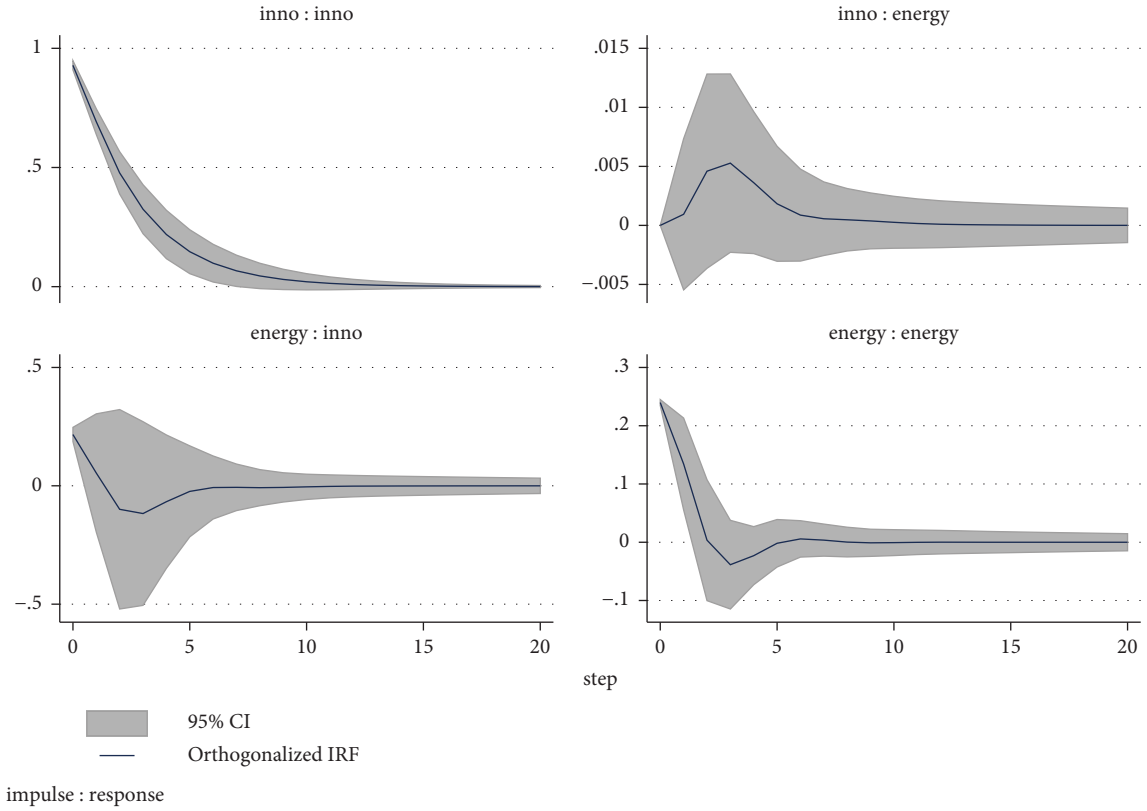


Figure 2: Impulse response. Note: the transverse axis represents the lag period of the impact; the middle curve is the impulse response function curve; and the shadow part is the 95% confidence interval.

lag effect of two variables on the current period. Although the strengthening effect of energy utilization efficiency displays a "U-shaped" trend of "positive first and then negative" and gradually converges to zero, the impulse response diagram on the diagonal can be observed. Finally, the feedback effect is the lag of energy utilization efficiency on urban innovation capability. The impulse response in Figure 2 (Row 2 and Column 1) describes the response of the urban innovation capability to energy utilization efficiency's orthogonal impact. Given an orthogonal impact on energy utilization efficiency, urban innovation capability presents a "U-shaped" change of "positive first and then negative" and converges to zero in the 10^{th} phase.

Variance decomposition means the decomposition of the prediction mean square error of any endogenous variable into the contribution made by random shocks to each variable in the system. It calculates the percentage size of the contribution made by shocks to each variable shock, evaluating the impact of one variable on another. On the basis of the analysis of impulse response (Figure 2), we use variance decomposition to further examine the degree of interaction between urban innovation capability and energy utilization efficiency and obtain the contribution of the impact response of each equation to the fluctuation of each variable in the PVAR (2) system. The error variance decomposition results of the two core variables of energy utilization efficiency and urban innovation capability in the 1^{st}–20^{th} forecast periods are reported in Table 5. The test results prove that the variance decomposition of the 8th period is basically stable, and the conclusion is meaningful.

Moreover, it can be inferred that the variance of the prediction error of energy utilization efficiency comes from itself in the first period, which is unrelated to urban innovation capability (Table 5). However, the contribution rate of urban innovation capability to the change in energy use efficiency has increased over time and finally been maintained at approximately 9.09%, whereas the contribution rate of energy use efficiency to the change in urban innovation capability remains at approximately 4.28%. Compared with the contribution rate of energy utilization efficiency to the change of urban innovation capability, the latter has a greater explanation than the former.

4.3. Granger Causality Analysis. A Granger causality test is conducted on the two core variables in the PVAR system to examine whether there is an obvious causal relationship between urban innovation capability and energy utilization efficiency. The results are reported in Table 6.

Combining the Granger causality analysis results in Table 6 and the variance decomposition results in Table 5, it can be observed that the improvement of urban innovation capability is the reason for the improvement of energy utilization efficiency. The increase in energy utilization efficiency is not the reason for the increase in urban

Table 5: Variance decomposition of the prediction error of core variables.

	Variance decomposition			
	energy		inno	
	energy	inno	energy	inno
1st	100%	0%	5.17%	94.83%
2nd	95.99%	4.01%	3.59%	96.41%
3rd	93.27%	6.73%	3.67%	96.33%
4th	91.56%	8.44%	4.20%	95.80%
5th	90.92%	9.08%	4.33%	95.67%
6th	90.91%	9.09%	4.31%	95.69%
7th	90.90%	9.10%	4.29%	95.71%
8th	90.91%	9.09%	4.28%	95.72%
9th	90.91%	9.09%	4.28%	95.72%
10th	90.91%	9.09%	4.28%	95.72%
11th	90.91%	9.09%	4.28%	95.72%
12th	90.91%	9.09%	4.28%	95.72%
13th	90.91%	9.09%	4.28%	95.72%
14th	90.91%	9.09%	4.28%	95.72%
15th	90.91%	9.09%	4.28%	95.72%
16th	90.91%	9.09%	4.28%	95.72%
17th	90.91%	9.09%	4.28%	95.72%
18th	90.91%	9.09%	4.28%	95.72%
19th	90.91%	9.09%	4.28%	95.72%
20th	90.91%	9.09%	4.28%	95.72%

Table 6: Granger causality test.

Variable	Granger test (null hypothesis)	χ^2 value	Degree of freedom	P value
Energy	The increase in urban innovation capability is not the reason for the increase in energy utilization efficiency.	14.131	2	0.001
Inno	The increase in energy utilization efficiency is not the reason for the increase in urban innovation capability.	1.167	2	0.558

innovation capability, and whether it is energy utilization efficiency or urban innovation capability, the fluctuation of its prediction error is mainly due to itself. This conclusion provides a basis for using the dynamic threshold regression model to test the nonlinear effect of urban innovation capability on energy utilization efficiency.

5. DPTR Analysis

The threshold variables are set as the population density, industrial structure, and environmental pollution of the prefecture-level cities, and the DPTR model is established in this section to analyze the differences in the impact of urban innovation capability on energy utilization efficiency under different population density, industrial structure, and environmental pollution levels. The specific forms can be expressed as follows:

$$\begin{aligned}\text{energy}_{it} = c_0 &+ \left(\phi_1 \text{energy}_{i,t-1} + \theta_1 \text{inno}_{it}\right) \mathrm{I}\{q_{it} \le \gamma\} \\ &+ \left(\phi_2 \text{energy}_{i,t-1} + \theta_2 \text{inno}_{it}\right) \mathrm{I}\{q_{it} > \gamma\} + \alpha_i + \upsilon_{it},\end{aligned} \tag{7}$$

where energy_{it} is a time-varying dependent variable; inno_{it} and lag-dependent variable $\text{energy}_{i,t-1}$ are explanatory variables; $\mathrm{I}\{\cdot\}$ represents an indicator function, which is equal to 1 when the conditions in brackets are satisfied, otherwise 0; q_{it} denotes the three threshold variables that describe the urban population density, industrial structure, and environmental pollution; γ represents the threshold value; ϕ_1, ϕ_2, θ_1, and θ_2 represent the relevant slope parameters corresponding to the different intervals. Because the explanatory and threshold variables in the model may have endogenous problems, the error term of the model is set to $\varepsilon_{it} = \alpha_i + \upsilon_{it}$, which is composed of two parts by Seo and Shin [19]; α_i is an unobservable individual fixed effect; and υ_{it} is a zero mean heterogeneous random disturbance term (υ_{it} is assumed to be a martingale difference sequence, namely, $\mathrm{E}(\upsilon_{it}|\chi_{t-1}) = 0$, where χ_{t-1} is the natural filtering in period t, and it is not assumed that inno_{it} or q_{it} is measurable relative to χ_{t-1}, namely, $\mathrm{E}(\upsilon_{it}\text{inno}_{it}) \neq 0$ or $\mathrm{E}(\upsilon_{it}q_{it}) \neq 0$. This setting allows the endogeneity of the explanatory variable inno_{it} and the threshold variable q_{it} in the model). The estimation results of the impact of urban innovation capability on energy utilization efficiency based on DPTR are summarized in Table 7. Population density, industrial structure, and environmental pollution level are used as threshold variables to represent the population, industry, and environmental constraints of the city to a certain extent.

We use the bootstrap method proposed by Hansen [23] to simulate the asymptotic distribution and p value of the statistics to test the validity of the estimation results of the DPTR model shown in Table 7. The nonlinear test results show that p values are close to zero and the model does have a nonlinear relationship (Table 7). Consequently, a dynamic threshold model with population density, industrial structure, and environmental pollution level as threshold variables can be established. First, from the parameter estimation results with population density as the threshold variable, the threshold value is 263.9851, which divides the sample into two intervals of low population density ($q_{\text{pop}} \le 263.9851$) and high population density ($q_{\text{pop}} > 263.9851$), and the coefficients of variables in these two intervals are significantly different. When the urban population density is lower than approximately 264 people/km^2, the estimated value of the coefficient passes the 1% aboriginality test and demonstrates a positive "inertia" effect. This indicates that early energy utilization efficiency has a positive role in promoting later energy utilization efficiency under this threshold. The estimated value of the coefficient θ_1 is significantly negative, which indicates that the improvement of the innovation capability of cities with a low population density cannot improve their energy utilization efficiency but will inhibit it. However, in the urban population, the density is higher than 264 people/km^2, and the result is exactly the opposite. The energy utilization efficiency in the early stage is not conducive to improving energy utilization efficiency in the later stage, and improving

Table 7: Estimated results based on DPTR.

			Explained variable: energy		
			Model 1	Model 2	Model 3
Threshold variable			Population density (density)	Industrial structure (struc)	Environmental pollution (pollu)
Threshold value	γ		263.9851*** (13.3842)	0.4026*** (0.0004)	36285.2104*** (2190.5583)
	Explanatory variable	Coefficient			
Low					
	$energy_{i,t-1}$	ϕ_1	0.9086*** (0.0019)	0.9665*** (0.0045)	0.9794*** (0.0046)
	$inno_{it}$	θ_1	−0.0528* (0.0281)	−0.0029*** (0.0007)	−0.0095*** (0.0022)
High					
	$energy_{i,t-1}$	ϕ_2	−0.2038*** (0.0201)	−0.0241*** (0.0032)	−0.0408*** (0.0068)
	$inno_{it}$	θ_2	0.0275*** (0.0032)	0.0044*** (0.0007)	0.0005 (0.0024)
	Constant c_0		0.5829*** (0.0554)	−0.1812*** (0.0069)	−0.1540*** (0.0108)
Nonlinear test (p value)			0.00	0.00	0.00
Percentage of samples in high interval (%)			65.27%	59.84%	56.71%

Note. ***, **, and * represent the significance levels at 1%, 5%, and 10%, respectively; standard error is presented in parentheses.

urban innovation capability will significantly promote the improvement of urban energy utilization efficiency. Second, from the parameter estimation results with industrial structure as the threshold variable, the threshold value is 0.4026 and is significantly indigenous at the level of 1%, which indicates that when the proportion of the added value of the secondary industry in the GDP of a prefecture-level city is higher than this threshold, the improvement of urban innovation capability is conducive to the improvement of its energy utilization efficiency. On the contrary, it will damage the improvement of energy utilization efficiency. Finally, from the results of parameter estimation with environmental pollution as the threshold variable, the threshold value is 36285.2104 and shows aboriginality at 1% level. The threshold value divides the samples into high-pollution (pollu > 36285.2104) and low-pollution (pollu ≤ 36285.2104) cities. However, the improvement of urban innovation capability is beneficial to the improvement of energy utilization efficiency for high- and low-pollution cities. Notably, compared with high-pollution cities, the improvement of innovation capability in low-pollution cities will have a stronger effect on improving energy utilization efficiency.

6. Conclusion

From the dynamic nonlinear perspective, this study discusses the relationship between urban innovation capability and energy utilization efficiency by using the PVAR and DPTR methods. Using the 2003–2020 panel data samples of 281 prefecture-level cities in China, we discussed the dynamic correlation and mechanism of energy utilization efficiency and urban innovation capability. The results reveal that the improvement in urban innovation capability is the reason behind the improvement in urban energy utilization efficiency, and the improvement in energy utilization efficiency is not the reason behind the improvement in urban innovation capability. The level of energy utilization efficiency in the early stages of the city may be both a boost and an obstacle to the improvement of energy utilization efficiency in the later stages, depending on the situation of the city in terms of population density, industrial structure, and environmental pollution. For cities with low levels of population density, industrial structure, and environmental pollution, energy utilization efficiency has certain "inertia" characteristics. By contrast, for cities with high levels of population density, industrial structure level, and environmental pollution, the high efficiency of early energy utilization will hinder the improvement in energy utilization efficiency in the later period. From the perspective of urban innovation capability, enhancing urban innovation capability can not only improve energy utilization efficiency but also adversely affect cities with a low population density or weak secondary industrial base. Whereas for cities with a high population density or proportion of secondary industry, improving innovation capability will significantly improve urban energy utilization efficiency. Furthermore, the promoting effect of urban innovation capability on energy utilization efficiency in low-pollution cities is significantly stronger than that in high-pollution cities.

Some shortcomings remain in this study, which is unavoidable. First, the measurement of urban innovation capability is rather rough without considering the differences in patents (for example, patents for invention, patents for utility models, and patents for industrial design). The follow-up research can make a more detailed division of innovation capability according to Chinese patent classification standards so as to reflect the difference in quantity and quality of urban innovation capability. Second, this paper only considers the influence of urban population density, industrial structure, and environmental pollution on the relationship between urban innovation ability and energy utilization efficiency. A future study can further investigate the possible nonlinear relationship between urban innovation ability and energy utilization efficiency caused by economic development, urban infrastructure, policy implementation efficiency, etc.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant no. 51908229.

References

[1] M. T. Xu and C. Bao, "Quantifying the spatiotemporal characteristics of China's energy efficiency and its driving factors: a Super-RSBM and Geodetector analysis," *Journal of Cleaner Production*, vol. 356, Article ID 131867, 2022.

[2] C. L. Miao, D. B. Fang, L. Y. Sun, Q. L. Luo, and Q. Yu, "Driving effect of technology innovation on energy utilization efficiency in strategic emerging industries," *Journal of Cleaner Production*, vol. 170, no. 1, pp. 1177–1184, 2018.

[3] P. Sun, L. L. Liu, and M. Qayyum, "Energy efficiency comparison amongst service industry in Chinese provinces from the perspective of heterogeneous resource endowment: analysis using undesirable super efficiency SBM-ML model," *Journal of Cleaner Production*, vol. 328, Article ID 129535, 2021.

[4] A. T. Xu, K. Y. Qiu, C. Y. Jin, C. J. Cheng, and Y. H. Zhu, "Regional innovation ability and its inequality: measurements and dynamic decomposition," *Technological Forecasting and Social Change*, vol. 180, Article ID 121713, 2022.

[5] X. J. Che, P. Zhou, and K. H. Chai, "Regional policy effect on photovoltaic (PV) technology innovation: findings from 260 cities in China," *Energy Policy*, vol. 162, Article ID 112807, 2022.

[6] J. Y. Yang, G. Q. Xiong, and D. Q. Shi, "Innovation and sustainable: can innovative city improve energy efficiency?" *Sustainable Cities and Society*, vol. 80, Article ID 103761, 2022.

[7] W. Lv, X. Hong, and K. Fang, "Chinese regional energy efficiency change and its determinants analysis: Malmquist index and Tobit model," *Annals of Operations Research*, vol. 228, no. 1, pp. 9–22, 2015.

[8] S. Nizetic, N. Djilali, A. Papadopoulos, and J. J. P. C. Rodrigues, "Smart technologies for promotion of energy efficiency, utilization of sustainable resources and waste management," *Journal of Cleaner Production*, vol. 231, pp. 565–591, 2019.

[9] R. Wang, Q. Z. Wang, and S. L. Yao, "Evaluation and difference analysis of regional energy efficiency in China under the carbon neutrality targets: insights from DEA and Theil models," *Journal of Environmental Management*, vol. 293, Article ID 112958, 2021.

[10] Z. L. Zheng, "Energy efficiency evaluation model based on DEA-SBM-Malmquist index," *Energy Reports*, vol. 7, pp. 397–409, 2021.

[11] Q. Y. Zhu, X. C. Li, F. Li, J. Wu, and D. Q. Zhou, "Energy and environmental efficiency of China's transportation sectors under the constraints of energy consumption and environmental pollution," *Energy Economics*, vol. 89, Article ID 104817, 2020.

[12] M. Incekara, "Determinants of process reengineering and waste management as resource efficiency practices and their impact on production cost performance of Small and Medium Enterprises in the manufacturing sector," *Journal of Cleaner Production*, vol. 356, Article ID 131712, 2022.

[13] H. S. Ai, M. Y. Wang, Y. J. Zhang, and T. T. Zhu, "How does air pollution affect urban innovation capability? Evidence from 281 cities in China," *Structural Change and Economic Dynamics*, vol. 61, pp. 166–178, 2022.

[14] J. Cheng, J. M. Zhao, D. L. Zhu, X. Jiang, H. Zhang, and Y. J. Zhang, "Land marketization and urban innovation capability: evidence from China," *Habitat International*, vol. 122, Article ID 102540, 2022.

[15] C. Zou, Y. C. Huang, S. S. Wu, and S. L. Hu, "Does "low-carbon city" accelerate urban innovation? Evidence from China," *Sustainable Cities and Society*, 2022.

[16] Z. J. Feng, H. C. Cai, Z. N. Chen, and W. Zhou, "Influence of an interurban innovation network on the innovation capacity of China: a multiplex network perspective," *Technological Forecasting and Social Change*, vol. 180, Article ID 121651, 2022.

[17] L. Huang, S. Q. Lin, and J. Chen, "The spatial-temporal coupling analysis of China's regional innovation capability and energy utilization efficiency," *World Regional Studies*, vol. 29, no. 6, pp. 1161–1171, 2020.

[18] B. E. Hansen, "Threshold effects in non-dynamic panels: estimation, testing, and inference," *Journal of Econometrics*, vol. 93, no. 2, pp. 345–368, 1999.

[19] M. H. Seo and Y. Shin, "Dynamic panels with threshold effect and endogeneity," *Journal of Econometrics*, vol. 195, no. 2, pp. 169–186, 2016.

[20] J. Li, S. L. He, J. L. Wang, W. F. Ma, and H. Ye, "Investigating the spatiotemporal changes and driving factors of nighttime light patterns in RCEP Countries based on remote sensed satellite images," *Journal of Cleaner Production*, vol. 359, Article ID 131944, 2022.

[21] Y. M. Zheng, L. N. Tang, and H. W. Wang, "An improved approach for monitoring urban built-up areas by combining NPP-VIIRS nighttime light, NDVI, NDWI, and NDBI," *Journal of Cleaner Production*, vol. 328, 2021.

[22] Z. W. Huang, S. Y. Li, F. Gao, F. Wang, J. Y. Lin, and Z. L. Tan, "Evaluating the performance of LBSM data to estimate the gross domestic product of China at multiple scales: a comparison with NPP-VIIRS nighttime light data," *Journal of Cleaner Production*, vol. 328, Article ID 129558, 2021.

[23] B. E. Hansen, "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, vol. 64, no. 2, pp. 413–430, 1996.

Hindawi
Complexity
Volume 2022, Article ID 8607185, 18 pages
https://doi.org/10.1155/2022/8607185

Research Article

Research on Influencing Factors of Knowledge Hiding Behavior in Socialized Q&A Communities: Taking Zhihu as an Example

Wen-Zhu Li,[1] **Jiang-Fei Chen**,[2,3,4] **Xin Feng**,[3,5] **and Qiang Yan**[1]

[1]*School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*School of Arts & Design, Yanshan University, Qinhuangdao 066004, China*
[3]*Information Center for Military and Civilian Collaboration of Beijing-Tianjin-Hebei, Yanshan University, Qinhuangdao 066004, China*
[4]*Hebei Design Innovation and Industrial Development Research Center, Yanshan University, Qinhuangdao 066004, China*
[5]*School of Economics and Management, Yanshan University, Qinhuangdao 066004, China*

Correspondence should be addressed to Xin Feng; 149987543@qq.com

Received 15 July 2022; Revised 12 August 2022; Accepted 3 September 2022; Published 27 September 2022

Academic Editor: Andrea Murari

With the normalization of epidemic prevention and control, the expression of the public's demand for health information on online platforms continues to increase, while knowledge hiding behavior has seriously hindered the communication and dissemination of epidemic prevention knowledge and has a negative impact on public communication and access to health information in the socialized Q&A communities. Therefore, further stimulating diving users' activity and reducing their knowledge hiding behavior have become the key to the sustainable development of epidemic prevention and control and communities. Based on the social cognition theory, from the perspective of individual cognition and external environment, this study constructs a theoretical model of the influencing factors of users' knowledge hiding behavior in the socialized Q&A communities in the post-epidemic era and puts forward relevant assumptions. 151 effective questionnaires are collected and an empirical analysis is carried out by using the structural equation model. The results show that outcome expectation, community atmosphere, and requesting negatively affect knowledge hiding behavior; self-efficacy, outcome expectation, and community atmosphere negatively affect the three different types of knowledge hiding behavior, which are evasive hiding, playing dumb, and rationalized hiding; community atmosphere positively affects outcome expectation, which plays a significant intermediary effect between community atmosphere and knowledge hiding behavior. The research content and relevant conclusions of this study deepen and expand the connotation and extension of knowledge hiding behavior in the negative performance of Q&A communities. From the perspective of practical application, it can also effectively reduce knowledge hiding behavior, grasp the development direction of public health needs, and strengthen the dissemination of epidemic prevention and control knowledge.

1. Introduction

At the beginning of 2020, the COVID-19 spread rapidly around the world, bringing a crisis and inconvenience to every individual in the society, causing public panic and anxiety, and greatly stimulating the public's demand for health information [1], thus triggering an upsurge of information search on socialized Q&A communities. Socialized Q&A communities are based on communities, user relations, and content operation, emphasizing users' social relations and self-generated content [2]. With the rapid development of network technology, as the emerging interactive platform with the characteristics of communication convenience, communication flexibility, variety, and timeliness of knowledge, socialized Q&A communities, has ushered in its era of rapid development and is gradually becoming an important place for people to acquire daily knowledge and share opinions. In the socialized Q&A communities represented by "Zhihu," users can share and obtain the information they need by searching, browsing, asking questions, commenting, or answering relevant questions; among them, we searched the keyword "epidemic

prevention and control" under the topic column and found that the problem data have accumulated more than 50,000. Therefore, user interaction is the basis of information or knowledge exchange in socialized Q&A communities, and stable user interaction is the guarantee of the normal operation and sustainable development of socialized Q&A communities [3].

The increased public demand for online platforms for health information generated by this outbreak has become a challenge to successfully retain users and motivate them to contribute effective information to Q&A communities. Even for a successful Q&A community, most people only pay attention to personal preference information and rarely focus on other people's questions, which plays a relatively passive role [4]. Under the topic of epidemic prevention and control on Zhihu, although there are a lot of questions, the discussion and interaction on the topic are relatively small. Some users choose to hide or ignore the questions even though they know the answers, which leads to the emergence of knowledge-hiding behavior. Knowledge hiding is a conscious and purposeful behavior of hiding information, retaining knowledge answers, or refusing to answer directly [5], which will lead users to reduce their efforts and scientific level in knowledge sharing, hinder the transmission of new knowledge and the development of new ideas [6], hinder the knowledge mobility among users and sustainable healthy development of socialized Q&A communities, and affect the public's access to health information in the post-epidemic era. Knowledge hiding behavior has serious harm, but it is not an optimal solution to shut down the community just because of the knowledge hiding behavior of some users, which can bring trouble and unfairness to users who are still contributing high-quality answers to the community. Therefore, this study takes finding the influencing factors of knowledge hiding behavior as the research motivation and provides an effective reference for the optimization of socialized Q&A communities to further understand users' preferences and intentions [7], which is conducive to further enhancing the competitiveness of the communities, effectively grasping the laws and characteristics of public health information demand in the post-epidemic era, and providing a reference for relevant government departments and socialized Q&A communities to better serve the public.

In recent years, more and more scholars have studied the negative behaviors in socialized Q&A communities. As one of the users' negative behaviors, knowledge hiding behavior has attracted more attention than before. However, few studies focus on the combination of public health needs and knowledge hiding behavior in the post-epidemic era. In terms of research status, scholars' research balance on knowledge hiding and knowledge sharing is still inclined toward knowledge sharing [8], for example, Shi et al. [9] and Zhao et al. [10] have conducted studies on this. The research on knowledge hiding is mainly concentrated in organizations, companies, and employees, and the knowledge hiding survey from the perspective of *Q*&A communities is relatively missing, which makes the theories in the research field lack corresponding theoretical support in the actual research situation. Therefore, to explore the influencing factors of users' knowledge hiding behavior in the Q&A communities in the post-epidemic era, this study takes "Zhihu" as the research platform, which is highly active in China's socialized Q&A communities and has representative users, which takes epidemic prevention and control as a research topic, applies social cognitive theory, and empirical analysis methods to explore the synergistic effects of self-efficacy, community atmosphere, outcome expectation, and request on knowledge hiding behavior, among them, self-efficacy and outcome expectation, as a form of self-assessment, belong to individual subjective judgment; community atmosphere and request as the external characteristics, refer to the environment gradually formed in the community, which can affect users. Thus, this study selects these factors as research variables to reveal the key influencing factors of users' knowledge hiding behavior and then puts forward suggestions to reduce users' knowledge hiding behavior. The research of this study is conducive to improving users' participation in the socialized Q&A communities, thus promoting and maintaining a virtuous circle of the community. It is conducive to creating a fair and open community environment, promoting mutually beneficial information exchange behavior of users, and guiding users to share knowledge. It is of great significance to enrich the user behavior theory of socialized Q&A communities and strengthen the connection between the community and users.

2. Related Research Studies

2.1. Knowledge Hiding Behavior. Connelly et al. [11] sorted out the concept of knowledge hiding behavior by studying the reasons behind the unsuccessful knowledge sharing and clearly defined knowledge hiding behavior as the intentional act of retaining, hiding, or refusing to give knowledge, the structure of knowledge hiding behavior is explored through empirical study, and believes that there are three dimensions: evasive hiding, playing dumb, and rationalized hiding. Among them, evasive hiding means that the knowledge concealer provides irrelevant information to the inquirer to replace the information which is needed, which delays time and gives misleading promises but does not provide help; playing dumb refers to the behavior of the knowledge concealer pretending not to understand the knowledge they ask; and rationalized hiding means that the person who conceals knowledge gives a valid reason as to why he cannot provide the knowledge. This kind of concealment behavior is not deceptive. According to Nielsen [12], the "90-9-1" rule of unequal participation prevails in Q&A communities, 90% of users are divers who never express their opinions, 9% are occasional contributors, and 1% are experts who provide most of the community's content. This indicates that most users in Q&A communities only browse knowledge and do not actively participate in knowledge sharing, showing a tendency for knowledge hiding.

It is very important to distinguish between knowledge hiding and knowledge sharing. Knowledge hiding is not only a lack of sharing but its motivation may have many different reasons, and the lack of knowledge sharing may only be due

to the lack of knowledge itself. Therefore, the two are not opposed to each other but are two conceptually different structures [11]. Therefore, knowledge hiding behavior has gradually been a concern for scholars who have carried out targeted research with rich research results. For example, Nguyen et al. [13] developed a conceptual framework based on resource conservation to study knowledge hiding behavior and its consequences; its purpose was to deal with the problem that employees may engage in knowledge hiding to maintain their resources and competitive advantage due to organizational crisis under COVID-19. Fauzi [14] used the systematic literature review method to conduct quantitative research on employees' knowledge hiding behavior and regarded it as immoral and antisocial behavior, which is considered to be detrimental to team development. Jafari-Sadeghi et al. [15] applied the DEMATEL to sort out the causal relationships between knowledge hiding components and provided a conceptual framework. Huang [16] explored the influence of overqualification on knowledge hiding by constructing an intermediary model, which showed that employees' sense of excess qualification negatively affects knowledge hiding. This study enriches the mechanism and boundary conditions of excess qualification and knowledge hiding. Li and Ke [17] conducted an empirical study on the influencing factors of users' knowledge hiding behavior in Q&A communities, from three aspects of personal characteristics, situational atmosphere, and knowledge characteristics to analyze the seven factors that affect knowledge hiding behavior, and put forward valuable suggestions for improving the degree of knowledge exchange in virtual communities. Lu et al. [18] conducted a study on the grouping of knowledge hiding behaviors in socialized Q&A communities based on FsQCA and explored the reasons for users' knowledge hiding behaviors; this research is of great significance to enrich the relevant theories of user behavior in socialized Q&A communities.

Moreover, in terms of the reasons for knowledge hiding, Hamza et al. [19] focused on the mediating role of team member exchange (TMX) and examined the influence of personality traits and individual ethnicity on knowledge hiding behavior. The study found that openness, conscientiousness, neuroticism, and ethnicity are positively correlated with knowledge hiding, while TMX as a mediator transforms this positive correlation into a negative one. Anand et al. [20] thought personal beliefs or situational constraints cause knowledge hiding, and identified the driving factors that lead to knowledge hiding. Among them, situational driving explained the reasons as to why performance and competition lead to unconscious hiding: psychological ownership driving leads to controlled hiding, hostility and abuse driving from employees or managers lead to victimization hiding, and identity and norms driving lead to preference hiding. Alam et al. [21] believed that negative emotions are a major cause of knowledge hiding, in which relationship conflict positively affect knowledge hiding, and frustration regulates the relationship between relationship conflict and knowledge hiding to a certain extent.

By analyzing the existing literature on knowledge hiding behavior, it is found that as important emerging platforms for information sharing and acquisition, socialized Q&A communities have few research achievements on it. Users, as the main producers of information and content in socialized Q&A communities, are key factors to promote the sustainable and healthy development of the communities, so the prevalence of knowledge hiding in the communities will inevitably have a negative impact on the development of the communities. On the one hand, knowledge hiding seriously undermines the knowledge creativity of the virtual academic community, reduces the influence of the community [22], and breaks the good academic atmosphere in the community; on the other hand, knowledge hiding reduces users' own knowledge creativity and also affects the willingness of other users in the community to share knowledge, which eventually leads to the vicious development of the communities.

In summary, exploring the influencing factors of users' knowledge hiding behavior not only enriches the user behavior theory of socialized Q&A communities but also contributes to the healthy development of communities. Due to the imperfect community standard system in the socialized Q&A communities and less offline communication of users, the behavior of users in the communities is more affected by personal factors, and there are also environmental factors affecting the behavior of users in the communities. In addition, Meng et al. [1] took "Zhihu" as an example, used the LDA theme model to build a coding system for users' health information needs, and revealed the characteristics and evolution rules of users' health information needs from the dimensions of time and demand theme. The research study found that the health information needs of Internet users in the post-epidemic period mainly focused on the knowledge related to COVID-19, epidemic prevention and control, and social impact, among which the core demand of users was "epidemic prevention and control." Therefore, this study takes the post epidemic era as background, takes the topic of epidemic prevention and control on "Zhihu" as a theme, and explores the users' knowledge hiding behavior in socialized Q&A communities from the perspective of individuals, combined with external environmental factors.

2.2. Social Cognitive Theory. Social cognitive theory (SCT) is derived from Bandura's social cognitive learning theory, which holds that individual behavior is affected by its factors and external environmental factors such as environment and atmosphere. Therefore, the dynamic interaction relationship between behavior, individual, and environment constitutes its core view, that is, the "Triadic Theory" model [23]. The "Triadic Theory" model holds that the interaction of two factors will affect people's motivation, emotion, attitude, and behavior. Flavell proposed that the object of social cognition is human events, which are the cognition of people and their behavior. Fang [24] proposed that social cognition is people's understanding of themselves and others. Shi [25] believed that social cognition is a process in which individuals speculate and judge the psychological state, behavioral motivation, and intention of others. At present, the research

study based on the perspective of social cognitive theory mainly focuses on knowledge-sharing behavior in enterprises, group organizations, and virtual communities. For example, Hsu et al. [26] proposed a knowledge-sharing model based on social cognitive theory, which includes self-efficacy, outcome expectation of personal impact, and multidimensional trust of environmental impact. Cai and Shi [27] discussed the potential mechanisms of community atmosphere on willingness to share knowledge in virtual communities based on social cognitive theory. This study is the first to explore whether and how the community atmosphere affects knowledge sharing and provides practical insights on how to use the community atmosphere to promote knowledge sharing in the Q&A communities. Zhan and Xiong [28] discussed the intermediary role of moral evasion in the spectator response of uncivilized behavior based on social cognitive theory, which is conducive to deepening the understanding of uncivilized behavior.

As a mature theory, the social cognitive theory has been widely used to understand and predict individual and group behavior characteristics. Based on interactive determinism, social cognitive theory proves that there is a causal relationship between individual, environment, and behavior, in which individual and environment can jointly affect or even determine the occurrence of individual behavior. Through the framework of social cognitive theory, this study divides the influencing factors of knowledge hiding behavior in socialized Q&A communities into environmental factors and individual factors. Therefore, users' personal behavior motivation can be well explained by social cognitive theory. Thus, because of the knowledge hiding behavior which is prevalent in the socialized Q&A communities and based on the perspective of social cognitive theory this study explores and effectively extracts the influencing factors of knowledge hiding behavior from this theory.

3. Research Hypothesis and Model Construction

Based on social cognitive theory, this study explores the influencing factors of knowledge hiding behavior by combining individual cognitive and external environmental factors. According to SCT, self-efficacy and outcome expectation belong to individual cognitive factors (subjective feelings and expectation judgments), which influence the individual behavior. In addition, external environmental factors and requests, as indispensable factors in socialized Q&A communities, also have some influence on user behavior.

3.1. Research Hypothesis

3.1.1. Self-Efficacy. Self-efficacy refers to people's beliefs about what they need to do to complete a task or achieve a goal, that is, the degree of confidence that an individual can use the skills they possess to achieve the desired behavior [29]. Self-efficacy has an impact on what decisions individuals make and what behaviors they adopt; thus, self-efficacy is an important factor that potentially influences knowledge hiding behavior. As Nielsen [12] said, the Q&A community has 1% of experts who provide most of the high-quality content. When users have high self-efficacy, it means that such users have relatively high response ability and knowledge reserve levels. They feel confident about the content they contribute and do not shy away from hiding their knowledge, which makes such users become this kind of "1%." In recent years, the research methods of self-efficacy are mainly empirical research, and the research content mainly focuses on three themes: education, organization, and knowledge behavior [30]. Among them, in the existing studies that take self-efficacy as a factor to discuss its impact on users' knowledge behavior, Lee et al. [31] tested multiple mediating effects of self-efficacy between knowledge sharing and sustainable well-being, and the results showed self-efficacy positively mediated the relationship between knowledge sharing and sustained well-being. Yang and Li [32] believed that the higher the self-efficacy, the more confident is one of their own ability and valuable contribution to the virtual communities, and the more willing they are to promote knowledge sharing. Zhao and Li [33] tested the hypothesis that self-efficacy has a direct positive impact on knowledge-sharing behavior. According to SCT, individuals are more motivated to do things they know with full confidence and ability and pay less effort for things they are not sure about. Liu and An [6] showed that employees with high self-efficacy positively affect knowledge hiding behavior. Similarly, this shows that if a user's self-efficacy is low and does not believe that he or she is competent enough to answer questions related to epidemic prevention and control in the communities, then the user will not show positive knowledge-sharing behaviors and instead will present knowledge hiding behaviors, which would mean deliberately ignoring or avoiding the questions asked by others. Accordingly, the following hypothesis is formulated:

H1: Self-efficacy negatively affects knowledge hiding behavior

H1a: Self-efficacy negatively affects evasive hiding

H1b: Self-efficacy negatively affects playing dumb

H1c: Self-efficacy negatively affects rationalized hiding

3.1.2. Outcome Expectation. Outcome expectation refers to an individual's beliefs about the consequences of behavior he or she will take. As mentioned above, Hsu et al. [26] incorporated the outcome expectation into the knowledge-sharing model of the social cognitive theory. Because knowledge sharing and knowledge hiding are almost similar, this study takes the outcome expectation as one of the research variables to discuss the hypothesis between it and knowledge hiding. Based on SCT, positive expectations are regarded as incentives, because individuals often act according to the standard of self-interest. Therefore, we realize that individuals in the Q&A communities will implement knowledge sharing only when their expectations are met. At present, in the existing research on the impact of the outcome expectation on users' knowledge behavior,

Constant et al. [34] found that knowledge sharing behavior occurs when the benefits of the user's knowledge sharing behavior are comparable or exceed the initially expected reward. Bock et al. [35] confirmed that sharing behavior can be promoted when individuals receive benefits. The research conclusions of two scholars confirm that there is a positive impact between outcome expectation and knowledge sharing. In addition, Tang and Mao [36] found that individuals actively share their knowledge to gain respect from others by using the hierarchical regression method but when they feel isolated and threatened, they tend to reduce their knowledge-sharing behavior. This coincides with the research conclusion of Zhang et al. [37] and Dai [38], that is, the outcome expectation negatively affects knowledge hiding behavior. Similarly, in the socialized Q&A communities, users will tend to adopt a positive attitude towards knowledge-sharing behavior if they feel recognition and respect from other users; if users feel their knowledge-sharing behavior brings less expected reward or harms their interests, then users will tend to reduce or stop this behavior, that is, they tend to exhibit knowledge hiding behavior when they perceive bad outcome expectation. Accordingly, the following hypothesis is formulated:

H2: Outcome expectation negatively affects knowledge hiding behavior

H2a: Outcome expectation negatively affects evasive hiding

H2b: Outcome expectation negatively affects playing dumb

H2c: Outcome expectation negatively affects rationalized hiding

3.1.3. Community Atmosphere. The community atmosphere is a relatively enduring characteristic associated with the community environment [39]. Many studies have explored knowledge-sharing behavior based on the perspective of organizational atmosphere, for example, Fu et al. [40] used the grounded theory to conduct exploratory research on the connotation structure and antecedents of the community atmosphere and pointed out that the individual's perception of the atmosphere is highly related to the individual's output. Kim and Park [41] concluded that organizational atmosphere directly affects knowledge sharing. Similarly, Li and Ke [17] and Lu et al. [18] have also studied the impact of external environmental characteristics on knowledge hiding. In the socialized Q&A communities, the community atmosphere is mainly manifested in the user's cognition of the importance of self-identity and the sense of community belonging. The better the community atmosphere, the more responsible and interested users feel in contributing knowledge [42]. This suggests that a good community atmosphere is conducive to a positive community state, strengthening trust and connection among members, which in turn promotes members' knowledge-sharing activities. This study focuses on three dimensions of community atmosphere: reciprocity, trust, and fairness.

Reciprocity refers to users contributing their knowledge to learn and use new knowledge returned by other users in the future [43]. Based on SCT, reciprocity indicates that there is no unremunerative altruistic behavior between individuals. For the common survival and development of the group, individuals will form a variety of interest relationships with each other [44]. Currently, there is a large body of research studies that argue for a relationship between reciprocity and knowledge sharing. For example, Lin [45] investigated the role of extrinsic (expected organizational rewards and reciprocity) and intrinsic (self-efficacy) motivators in knowledge-sharing intention, which found reciprocal benefits significantly affect employees' attitudes and willingness to engage in knowledge-sharing behaviors. According to SCT, since there is a causal relationship between individuals and the environment, stable communication between users is often based on reciprocal exchange behaviors. When users are full of continuous reciprocal behaviors, they can maintain their trust and dependence on each other, thus producing positive knowledge contribution behaviors. As mentioned above, a good reciprocal atmosphere can reduce the occurrence of knowledge-hiding behaviors in the communities.

Trust, as one of the basic elements of socialization, is a manifestation of users' willingness and beliefs, including their perceptions of sincerity, reliability, kindness, and justice [46]. Trust is the basis for user communication and cooperation within socialized Q&A communities. The establishment of trust can deepen the sense of identity and coordinate conflicts among users, thus promoting the sharing of information within communities [47]. Looking at the literature on trust and knowledge sharing, it can be found that the research is mainly divided into three categories [48]: the first category focuses on empirical research and uses survey data to analyze the impact of trust on knowledge sharing. For example, Chi et al. [49] divided trust into member trust and community trust, constructed a theoretical model of the impact of virtual community governance mechanism with trust as an intermediary variable on knowledge sharing behavior, and found that member trust and community trust play a significant mediating role, respectively; the second category focuses on theoretical research and theoretically analyzes the impact of trust on knowledge sharing. For example, Lin et al. [50] pointed out that trust in both the goodwill dimension and capability dimension strongly affect knowledge sharing; the third category is game research. For example, Zhang et al. [51] integrated trust and knowledge-sharing evolutionary game into the same framework and pointed out that cognitive trust plays an important role in knowledge contribution. In socialized Q&A communities, users will gradually develop trust as they interact with each other's information and thus believe that someone will lend a helping hand in a time when they need help. In addition, trust as a complex and multidimensional concept has led scholars to classify trust factors into different dimensions for analysis. Among them, Hsu et al. [26] believed that trust belongs to external environmental factors, which is the user's subjective feeling toward the community. This study agrees with Hsu's points

of view and based on this division of trust, the trust dimension in the community atmosphere is understood as users' trust in sharing information and the spirit of unity and fraternity in the socialized Q&A communities. In summary, a good trust atmosphere will motivate users to actively participate in knowledge sharing and thus reduce the occurrence of knowledge hiding behaviors.

Fairness refers to users' perception that the community treats themselves and others equally and without prejudice. Users often have a psychological perception of organizational fairness through the reasonable distribution of material resources or remuneration with other members [52]. Hao [53] took enterprise employees as the research subject and found that the advancement of knowledge-sharing behavior is constrained by inequities within the organization. According to SCT, the unfair environment reduces the user's identification and emotional attachment to the community, hinders the communication between users, reduces the probability of the occurrence of reciprocal behaviors, and thus leads to the increase of knowledge hiding behaviors. Accordingly, the following hypothesis is formulated:

H3: Community atmosphere negatively affects knowledge hiding behavior

H3a: Community atmosphere negatively affects evasive hiding

H3b: Community atmosphere negatively affects playing dumb

H3c: Community atmosphere negatively affects rationalized hiding

As the community atmosphere is a reflective structure based on the low-level structure of reciprocity, trust, and fairness, so the item of community atmosphere is composed of these three factors. According to Becker et al. [54], the research model based on the concept of more than second order can be selected by two methods, the repeated index method and the two-stage method. The repeated index method is used in this study, and it cites that all the indicators of each LOC belonging to the HOC are designated as the reflective measurement indicators of HOC when constructing the model. As an HOC, the community atmosphere forms a second order with LOC of reciprocity, trust, and fairness, while they form the first order with their nine projects (shown in Table 1). Therefore, the community atmosphere will be set as a measurement index with nine reflections.

Meanwhile, a good atmosphere of reciprocity, trust, and fairness enables users to perceive the degree of value to be obtained in the future, thus enabling them to accurately judge the outcome expectation. Therefore, a good community atmosphere can promote the outcome expectation of users' knowledge sharing. Accordingly, the following hypothesis is formulated:

H4: Community atmosphere positively affects outcome expectation

Among them, outcome expectation, as an intermediary variable, indirectly transmits the influence of community atmosphere to knowledge hiding behavior, playing a transmission role. A good community atmosphere enables users to perceive the value obtained in the future and then enables users to accurately judge the outcome expectation. With accurate outcome expectations, users will tend to strengthen the trust and contact among members, thus, promoting knowledge sharing. In other words, the community atmosphere positively affects outcome expectation, and outcome expectation negatively affects knowledge hiding behavior, that is, the community atmosphere indirectly has a negative impact on knowledge hiding behavior by influencing outcome expectation. Accordingly, the following hypothesis is formulated:

H5: Outcome expectation plays a mediating role between community atmosphere and knowledge hiding behavior

3.1.4. Request. In the post-epidemic era, the requests for knowledge and socialized Q&A community services have changed greatly. In the socialized Q&A communities, the most original and basic needs should be the users' requests for knowledge, which is also the most basic purpose for users to enter the online knowledge community. Requests are the premise and foundation for carrying out knowledge service activities [61], and insight into users' knowledge requirements in the current context is fundamental. Zhang et al. [62] analyzed the causes, levels, and characteristics of requests and found that in the socialized Q&A communities, users' requests for knowledge are the premise to promote the occurrence of knowledge behaviors such as questioning, querying, and acquiring, which determine the content, mode, and future development direction of the knowledge service. Based on the openness of the online community, users can publish their knowledge requests in the community anytime and anywhere. However, openness also brings the problem of a lack of unified planning and management, resulting in an unbalanced and inadequate knowledge supply and knowledge requirements satisfaction in Q&A communities [63]. When the standards required by users for knowledge cannot be met, users tend to hide their efforts when participating in knowledge activities, making their knowledge-sharing efforts lower than the level they can fully share, forming knowledge hiding behavior. Accordingly, the following hypothesis is formulated:

H6: Request negatively affects knowledge hiding behavior

3.2. Research Model Construction. Based on the abovementioned research hypotheses, this study proposes a research model on the influencing factors of knowledge hiding behavior in socialized Q&A communities based on social cognitive theory, as shown in Figure 1. Taking self-efficacy and outcome expectation as individual perception variables, combined with community atmosphere and request, this research work studies the influencing factors of knowledge hiding behavior in socialized Q&A communities and

Table 1: List of questionnaire measurement items.

Variables	Title item	References
Self-efficacy	I think I can bring valuable content related to epidemic prevention and control to other users I think I have a lot to say about relevant knowledge of epidemic prevention and control It makes no difference to me whether other users respond to what I post or not I prefer to answer questions with a high degree of certainty I get a sense of satisfaction and achievement when the knowledge I share is widely recognized	[55]
Outcome expectation	When the knowledge I shared receives many likes, the number of quality answers I contribute will increase When I am praised for the knowledge I share, I feel that I am recognized and respected I am more effective at contributing quality responses when the act of sharing brings the expectation of interpersonal benefit The level of responses I contribute will increase when the act I share brings the expectation of financial gain	[35]
Community atmosphere (reciprocity)	When I answer others' questions in the community, I want others to answer mine too When I share knowledge about epidemic prevention and control, I want to be able to take knowledge from the community as well I think Zhihu is a platform where interests are exchanged	[45]
Community atmosphere (trust)	I believe there is a spirit of camaraderie and mutual support in the community I think you can find a sense of belonging in a community I believe that the knowledge about epidemic prevention and control I have gained from the community is reliable	[56]
Community atmosphere (fairness)	I think what I get is fair compared to how motivated I am to answer questions I think what I get is fair compared to the amount of time I contribute to the community I think it is fair to say that the quality answers I provide are the same as the quality answers I get	[57]
Request	When the requests for knowledge meet physiological and safety needs, I will improve the knowledge sharing When the requests for knowledge meet social needs, I will improve the knowledge sharing When the requests for knowledge meet self-fulfilling needs, I will improve the knowledge sharing	[58]
Knowledge hiding	In the community, when another user asks for knowledge about epidemic prevention and control, I give a response that may not be the information he needs In the community, when other users ask for knowledge about epidemic prevention and control, I refuse to help even if I know the answer In the community, when other users ask for knowledge about epidemic prevention and control, I reply "I will help later" but in reality, it is "I will delay as long as I can." In the community, I prefer to hoard knowledge rather than share it	[59]
Knowledge hiding (evasive hiding)	When someone asks for information about epidemic prevention and control, I will verbally promise to help him but I do not really intend to help him When someone asks for information about epidemic prevention and control, I will agree to help him but will give him a different message When someone asks for information about epidemic prevention and control, I will tell him I will help him in the future but actually, I put it off for as long as possible	
Knowledge hiding (playing dumb)	When I exchange information with other users, although I know some information, I say I do not know When I exchange information with other users, I pretend not to know what they are talking about	[60]
Knowledge hiding (rationalized hiding)	When other users ask for information about epidemic prevention and control, I explain that the information is confidential and is only visible to specific people When other users asked for information about epidemic prevention and control, I say I do not know much about this topic When other users asked for information about epidemic prevention and control, I will tell them I want to tell them, but I cannot	

discusses the impact of three factors in SCT on three different types of knowledge hiding behavior.

3.3. Questionnaire Design. This study uses users of Zhihu as the research target. The samples were collected by randomly issuing questionnaires on the Internet. The questionnaire consists of three parts: (1) a basic description of the questionnaire, which explains the purpose of the research; (2) basic information about the respondents and their basic use of the Zhihu platform; and (3) measurement questions on the research variables, that is, respondents answer questions based on their personal experiences and feelings.

The measurement indicators for this study were taken from existing literature and have been adapted according to the current use of socialized Q&A communities in China

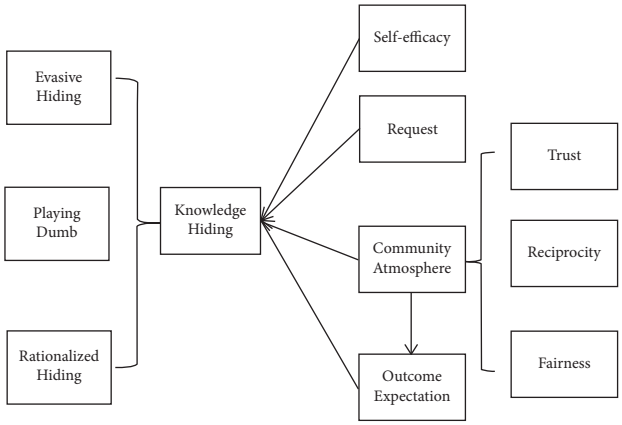


Figure 1: research model of influencing factors of knowledge hiding behavior.

and the research content of this study. The questionnaire used a five-level Likert scale to measure variables, corresponding to the level of strongly disapprove, disapprove, neutral, approve, and strongly approve as shown in Table 1.

4. Data Collection and Analysis

4.1. Sample Collection and Descriptive Analysis. Considering time cost, economic cost, and other factors, this questionnaire survey was conducted from June 7, 2022, to June 27, 2022. The questionnaire was designed and distributed according to 3.3, after comprehensive consideration of the filling time, filling profile, and filling IP, a total of 151 valid questionnaires were collected within 20 days. The sample size calculation and the information of the target respondents are shown in Table 2.

Descriptive statistical analysis is carried out on the abovementioned survey results to find out the internal rules of these data samples, and further understand the characteristics of the audience groups of this survey through the scientific description and to prepare for the next analysis. The research objects have the following characteristics:

(1) *Gender characteristics*: in the results of this research, the sample size of males is 81, accounting for 53.6% of the total sample; the sample size of females is 70, accounting for 46.4% of the total sample, relatively speaking, males account for a larger proportion.

(2) *Age characteristics*: users aged 19–30 account for the highest proportion; such users are generally college students or young people who have just started working. They have more free time and no life pressure, so they may increase investment in entertainment. This is followed by users over 40 years old, whose personal ability tends to be saturated and they can spend more time on the network. Users aged 30–40 account for 12% of the total, such users are generally already working. Because of various factors such as personal thirst for knowledge and social needs, they choose to use online knowledge communities to enrich their personal experience after their daily work. Users under the age of 18 accounts for the smallest proportion, because these users are minors, and the use of the Internet will be controlled by the family, society, software, and other aspects. These age characteristics reflect the diversity, youth, and inclusiveness of the community.

Table 2: Descriptive statistics of knowledge hiding behavior survey.

Item	Category	Frequency ($N = 151$)	Percentage (%)
Gender	Female	70	46.4
	Male	81	53.6
Age	18 years and under	17	11.3
	19–30 years	91	60.3
	30–40 years	18	12.0
	40 years and over	25	16.4
Educational level	Middle school	3	2.0
	High school	7	4.6
	Technical secondary school or junior college	12	8.0
	Undergraduate	123	81.4
	Postgraduate	5	3.3
	Ph.D. and above	1	0.7
Occupation	Student	108	71.5
	Business or self-employed	31	20.5
	Administrative agency or institution	9	6.0
	Other	3	2.0
Years of using Zhihu	1 year and under	9	6.0
	1-2 years	57	37.7
	3-4 years	58	38.4
	4-5 years	25	16.6
	5 years and over	2	1.3
Average number of posts posted per week	0	93	61.6
	1–5	28	18.5
	6–10	18	11.9
	11–20	12	8.0
	More than 20	0	0
Total		151	100.0

Table 3: Overall reliability analysis data.

Cronbach's α coefficient	Standardized Cronbach's α coefficient	Number of items	Number of samples
0.805	0.805	33	151

Table 4: KMO test and Bartlett's test.

KMO value		0.905
Bartlett's spherical test	Approximate chi-square	4206.587
	Freedom	595
	Significance	≤0.001

Table 5: Exploratory factor analysis matrix.

Measurement item	Composition									
	1	2	3	4	5	6	7	8	9	10
OE1	0.852									
OE4	0.824									
OE5	0.808									
OE2	0.807									
OE3	0.778									
KH1		0.760								
KH3		0.753								
KH2		0.752								
T3		0.725								
SE3			0.840							
SE2			0.802							
SE4			0.746							
SE1			0.738							
T1				0.843						
T2				0.827						
T3				0.823						
RE3					0.870					
RE2					0.862					
RE1					0.859					
F2						0.872				
F1						0.841				
F3						0.820				
R3							0.830			
R2							0.824			
R1							0.791			
EH1								0.801		
EH3								0.793		
EH2								0.775		
RH1									0.773	
RH2									0.761	
RH3									0.696	
PD1										0.890
PD2										0.821

(3) *Educational level*: about 85% of the respondents have a bachelor's degree or above, indicating that the users of online knowledge communities are generally well-educated and have a good knowledge reserve.

(4) *Occupational characteristics*: students account for the highest proportion, followed by people who are engaged in business or are self-employed, among which 71.5% are students and 20.5% are businessmen or self-employed. According to the age structure, the number of users under the age of 18 is small but the overall proportion of users is relatively high among students, indicating that the users who use online knowledge communities are mostly college students, masters, and Ph.D. business employees who have a large base in the social structure, so their proportion is also high.

(5) *Years of using Zhihu*: 6% of the respondents have used Zhihu for 1 year or less, 37.7% for 1–2 years, 38.4% for 3-4 years, 16.6% for 4-5 years, and 1.3% for more than 5 years, which ensures that the respondents are all users who have used Zhihu. It also shows that the user viscosity of Zhihu is very good, and the average years can reach more than 3 years.

(6) *The number of posts per week*: most of the Zhihu users in the sample are few replies and posts, and the sample size of users who do not post at all is the largest, accounting for 61.6%. The number of users who post more than 20 posts a week is zero. These data show that the actual activity of users is not high. There are many divers in Zhihu, and only a few users are willing to post and interact in the community, highlighting the potential phenomenon of knowledge-hiding behavior.

4.2. Reliability and Validity Analysis of the Questionnaire. In this study, the reliability analysis of the questionnaire was conducted by SPSS 25.0 to check the stability of the questionnaire; the validity analysis of the questionnaire was conducted by SPSS and AMOS software to verify the reasonableness of the quantitative data; finally, the goodness of fit of the model was verified by the structural equation model.

4.2.1. Reliability Analysis. In general, indicators with good reliability can be repeated under the same or similar

Table 6: Evaluation of the model AVE and CR indicators.

Factor	Mean-variance extraction AVE value	Combined reliability CR value
Factor1 (SE)	0.696	0.899
Factor2 (OE)	0.739	0.933
Factor3 (CA)	0.443	0.888
Factor4 (RE)	0.837	0.938
Factor5 (KH)	0.642	0.877
Factor6 (EH)	0.802	0.923
Factor7 (PD)	0.774	0.871
Factor8 (RH)	0.646	0.845

Table 7: Pearson correlation and AVE root value.

	SE	OE	CA	RE	KH	EH	PD	RH
SE	0.834							
OE	0.367	0.86						
CA	0.568	0.452	0.666					
RE	0.392	0.352	0.484	0.915				
KH	-0.415	-0.509	-0.541	-0.477	0.801			
EH	-0.505	-0.531	-0.544	-0.387	0.512	0.896		
PD	-0.374	-0.395	-0.429	-0.321	0.352	0.348	0.88	
	SE	OE	CA	RE	KH	EH	PD	RH
RH	-0.474	-0.546	-0.562	-0.383	0.434	0.463	0.349	0.804

The diagonal numbers are the root values of this factor AVE.

conditions to obtain consistent results. When reliability tests are conducted among different respondents and scorers, the higher the consistency of the results obtained, the higher the reliability of the questionnaire. This study adopts Cronbach's Alpha reliability measurement method with high recognition, and according to Kaiser's stipulation of Cronbach's α [64], α between 0.5 and 0.6 is not credible. As can be seen in Table 3, the overall reliability of the questionnaire variables is 0.805, which shows that the questionnaire of this study has good reliability, and the scale used has good internal consistency and is relatively reasonable in design.

4.2.2. Validity Analysis. The validity analysis includes content validity and construction validity: for content validity, this study refers to published articles and their designed questionnaire items [35, 45, 55–57, 59] and makes some modifications; for construction validity, the degree of interpretation of the actual test results on the measured indicators this study conducts an exploratory factor analysis to questionnaire scales for construction validity testing. Cerny and Kaiser [65] showed that when the KMO value is between 0.6 and 1, and the validity is appropriate and suitable for factor analysis. In this study, the exploratory factor analysis was adopted to test the validity of the measurement model and the scale, and the KMO test and Bartlett's test table were obtained. According to Table 4, the significance level of Bartlett's spherical test chi-square value is ≤0.001 and the KMO value is 0.905, which indicates that the scale has good validity and is suitable for factor analysis.

In this study, the factors were extracted based on principal component analysis, and the rotation method adopts the Kaiser normalization maximum variance method and sets the absolute value to 0.5 to estimate the factor load as shown in Table 5. Self-efficacy, outcome expectation, community atmosphere, trust, reciprocity, fairness, request, knowledge hiding, evasive hiding, playing dumb, and rationalized hiding are expressed as SE, OE, CA, T, R, F, RE, KH, EH, PD, and RH, respectively. Generally, the absolute value of factor loadings above 0.4 is considered a significant variable, and above 0.5 is considered a very important variable. As can be seen from Table 5, the factor loadings are all greater than 0.5, indicating that the ten factors extracted are well represented and the factors converge well.

In addition, to further confirm the convergent validity of variables within the factors of this model and to identify validity information, the Fornell-Larcker criteria are used to confirm the results of the model AVE and CR indicators. In general, AVE above 0.5 or CR above 0.7

Table 8: Normality test of observed variables.

Observed variables	*Skewness*		*Kurtosis*	
	Statistics	Standard error	Statistics	Standard error
SE1	−0.230	0.197	−0.999	0.392
SE2	0.077	0.197	−1.656	0.392
SE3	−0.207	0.197	−1.139	0.392
SE4	0.076	0.197	−1.214	0.392
OE1	−0.050	0.197	−1.497	0.392
OE2	−0.082	0.197	−1.530	0.392
OE3	−0.219	0.197	−0.886	0.392
OE4	0.026	0.197	−1.496	0.392
OE5	−0.148	0.197	−1.354	0.392
R1	−0.367	0.197	−1.277	0.392
R2	−0.235	0.197	−1.488	0.392
R3	−0.366	0.197	−1.123	0.392
T1	0.017	0.197	−1.581	0.392
T2	0.032	0.197	−1.602	0.392
T3	0.043	0.197	−1.630	0.392
F1	−0.223	0.197	−1.656	0.392
F2	−0.323	0.197	−1.332	0.392
F3	−0.217	0.197	−1.400	0.392
RE1	−0.155	0.197	−1.679	0.392
RE2	−0.178	0.197	−1.656	0.392
RE3	−0.225	0.197	−1.304	0.392
KH1	0.085	0.197	−1.392	0.392
KH2	0.213	0.197	−1.353	0.392
KH3	0.263	0.197	−0.935	0.392
KH4	0.470	0.197	−1.112	0.392
EH1	0.197	0.197	−1.579	0.392
EH2	0.283	0.197	−1.208	0.392
EH3	0.178	0.197	−1.681	0.392
PD1	−0.340	0.197	−1.291	0.392
PD2	−0.432	0.197	−1.249	0.392
RH1	0.120	0.197	−1.221	0.392
RH2	−0.155	0.197	−1.396	0.392
RH3	−0.135	0.197	−1.498	0.392

Table 9: Multicollinearity test results.

Elements		*Collinearity statistics*	
		Tolerance	VIF
SE	SE1	0.438	2.281
	SE2	0.291	3.433
	SE3	0.362	2.763
	SE4	0.490	2.040
OE	OE1	0.279	3.585
	OE2	0.308	3.244
	OE3	0.422	2.369
	OE4	0.283	3.532
	OE5	0.343	2.916
CA	R1	0.445	2.247
	R2	0.306	3.267
	R3	0.401	2.496
	T1	0.281	3.565
	T2	0.211	4.740
	T3	0.228	4.385
RE	F1	0.230	4.357
	F2	0.273	3.663
	F3	0.340	2.940
	RE1	0.213	4.693
	RE2	0.178	5.629
	RE3	0.313	3.195
KH	KH1	0.445	2.249
	KH2	0.428	2.338
	KH3	0.532	1.881
	KH4	0.440	2.271
EH	EH1	0.278	3.591
	EH2	0.351	2.852
	EH3	0.224	4.466
PD	PD1	0.422	2.370
	PD2	0.398	2.515
RH	RH1	0.475	2.105
	RH2	0.456	2.191
	RH3	0.457	2.188

indicates high convergent validity and good construct reliability. The Fornell-Larcker criteria require that the square root of the average variance extracted for a variable should be greater than its highest correlation with any other variable. According to Tables 6 and 7, it can be seen that the degree of extraction of the measures within the factors is excellent and the first-order variables in the model meet this requirement.

4.3. Model Fit Analysis and Hypothesized Results

4.3.1. Normality Test. Normal distributions are used in many scenarios. In general, the study of normality test methods can be based on the following directions: normality tests based on statistical plots, normality tests based on empirical distribution functions, and normality tests based on skewness and kurtosis [66]. In this study, a normality test based on skewness and kurtosis was chosen to verify the normality of the multivariate data and to determine whether AMOS analysis could be performed.

When the skewness coefficient is less than 3 and the kurtosis coefficient is less than 8, the data follow a normal distribution, otherwise, it does not obey the standard normal distribution [67]. As can be seen from Table 8, the skewness coefficients for all variables in this model are less than 3 and the kurtosis coefficients are less than 8, so these data are normally distributed and can be used for the AMOS analysis.

4.3.2. Multicollinearity Diagnosis. To test the multicollinearity problem of the current model, this study performs an analysis of multiple linear regression on the variables. In the regression model, the variance inflation factor (VIF) provides a measure of collinearity. If VIF <5, there is essentially no collinearity; if VIF exceeds 10, multicollinearity exists. According to Table 9, one of the VIF values is greater than 5, which is because the survey object is only for Zhihu users and there are certain restrictions on the basis and scope of the sample, and the basis of data collection is not wide enough. However, on the whole, the values of the independent variable multicollinearity test index VIF of this model are far below 10, which indicates that there is no multicollinearity between the independent variables, and the

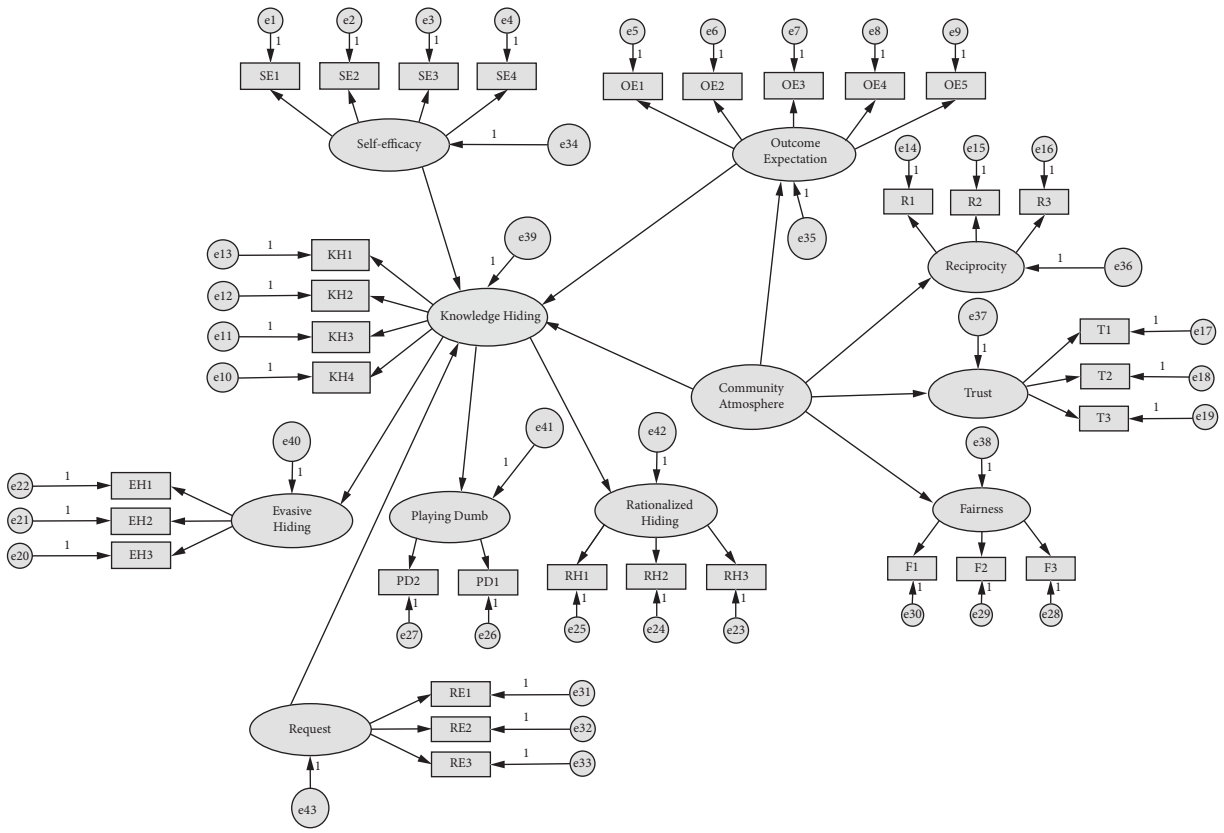


Figure 2: The structural equation model.

Table 10: Model fitting indicators.

	Statistical test volume	Name	Adaptation criteria	Test result data	The model fits or not
Absolute fit index	χ^2/df	Chi-square degrees of freedom ratio	<3	1.995	Fits
	GFI	The goodness of fit index	>0.9	0.977	Fits
	RMSEA	Root mean square error of approximation	<0.1	0.081	Fits
	RMR	Root mean square error	<0.1	0.066	Fits
Value-added adaptation index	CFI	Comparative fit index	>0.9	0.868	Not fits
	NFI	Normed fit index	>0.9	0.927	Fits
	NNFI	Non-normed fit index	>0.9	0.853	Not fits

degree of interaction between them does not affect the accuracy of the analysis of their respective effects, which meets the requirements of the model test criteria.

4.3.3. Simulation Fit Analysis. In this study, 151 valid questionnaires are collected and the Amos 24 is imported for the structural equation model analysis. The measurement model of the structural equation model in this study includes 10 latent variables, 33 observation variables, and 43 residual terms. The final model diagram is shown in Figure 2:

As shown in Table 10, this study analyzes the overall fitting index of the model from two aspects: absolute adaptation index and value-added adaptation index through verification factor analysis. From the data in the table, it can be seen that the index of most aspects of the model is suitable for the evaluation criteria, and the overall goodness of fit is good. A model is acceptable on the premise that multiple of these criteria fit well and cannot be too far from the cut-off values. Therefore, this shows the research model in Figure 2 can evaluate the research question of influencing factors of knowledge hiding behavior in socialized Q&A communities, and it can be considered that the fit of this model is acceptable.

4.3.4. Hypothesis Test Results. In this study, the path analysis of the structural equation model is used to

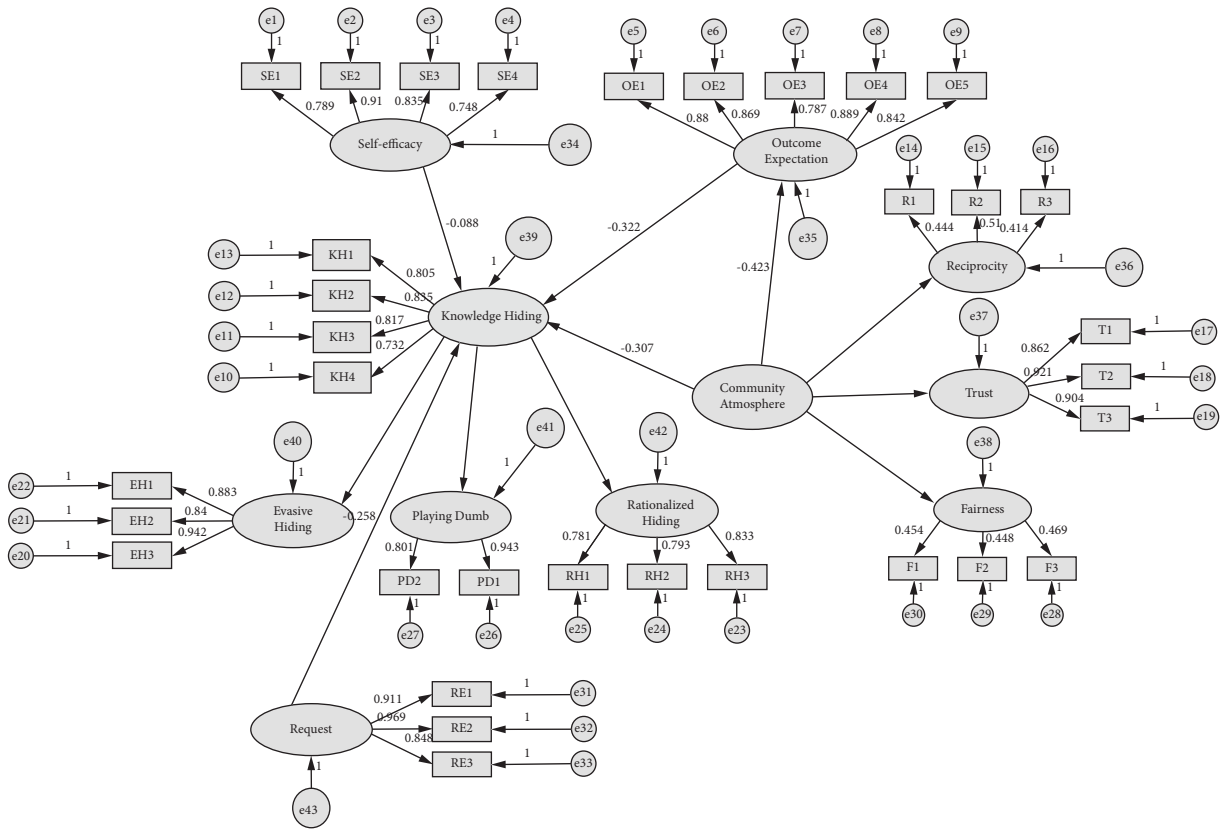


Figure 3: structural equation model normalized path coefficients.

Table 11: Hypothesis test results.

Hypothesis	Path	Path coefficient	P value	Test result
H1	Self-efficacy ⟶ knowledge hiding	−0.088	0.298	Not established
H1a	Self-efficacy ⟶ evasive hiding	−0.316	≤0.001	Established
H1b	Self-efficacy ⟶ playing dumb	−0.193	0.042**	Established
H1c	Self-efficacy ⟶ rationalized hiding	−0.257	0.003***	Established
H2	Outcome expectation—> knowledge hiding	−0.322	≤0.001	Established
H2a	Outcome expectation ⟶ evasive hiding	−0.379	≤0.001	Established
H2b	Outcome expectation ⟶ playing dumb	−0.259	0.004***	Established
H2c	Outcome expectation ⟶ rationalized hiding	−0.423	≤0.001	Established
H3	Community atmosphere ⟶ knowledge hiding	−0.307	0.002***	Established
H3a	Community atmosphere ⟶ evasive hiding	−0.202	0.026***	Established
H3b	Community atmosphere—> playing dumb	−0.251	0.020**	Established
H3c	Community atmosphere ⟶ rationalized hiding	−0.261	0.008**	Established
H4	Community atmosphere ⟶ outcome expectation	0.423	≤0.001	Established
H6	Request ⟶ knowledge hiding	−0.258	≤0.001	Established
H5	Community atmosphere ⟶ outcome expectation ⟶ knowledge hiding		≤0.001	Established

***, **, * represent the significance levels of 1%, 5%, and 10%, respectively.

calculate the standardized path coefficients between potential variables. As shown in Figure 3, by studying the standardized path regression coefficients between model variables, the causality hypothesis of each potential variable of the knowledge hiding model is verified, and the results of the hypothesis verification are more intuitively and clearly explained.

As can be seen from Table 11, the hypothesis proposed in this study is partially valid and the research model of knowledge hiding behavior constructed indicates that outcome expectation, community atmosphere, and request negatively affect knowledge hiding behavior; self-efficacy, outcome expectation, and community atmosphere and all of these negatively influence the three different types of

knowledge hiding; there is a positive effect between community atmosphere and outcome expectation, with outcome expectation mediating significantly between community atmosphere and knowledge hiding behavior. The P values indicate that the model is significant at all path levels and all paths are valid (except H1).

5. Conclusion

Based on social cognitive theory, this study establishes a research model of the influencing factors of users' knowledge hiding behavior in socialized Q&A communities. Six hypotheses are proposed through the analysis of literature on knowledge hiding. In this study, the influencing factors of knowledge hiding behavior are divided into four aspects: self-efficacy, outcome expectation, community atmosphere, and request where the community atmosphere is split into three dimensions of reciprocity, trust, and fairness. Moreover, it explores the influence of the three factors of SCT on the three types of knowledge hiding behavior, namely, evasive hiding, playing dumb, and rationalized hiding. Knowledge hiding behavior of "Zhihu" users in epidemic prevention and control under the background of post-epidemic era is studied through multiple factors, and the structural equation model is used to verify the model. Five of the six hypotheses in the model are significantly supported.

(1) The results show that outcome expectation, community atmosphere, and request negatively affect users' knowledge hiding behavior. As Constant et al. [34] and Fu et al. [40] believe, users' personal output is related to their perceived environment and perceived benefits. When users are in an environment with a low sense of fairness, reciprocity, and trust, and think that the benefits brought by their efforts are not higher than expected; then, they often lose their willingness to share, which is consistent with the research conclusion of Gu [68] and Han [69]. In addition, different from previous studies, this study expands on the influencing factors of knowledge hiding behavior and finds that request also negatively affects knowledge hiding. Users' knowledge requirements for epidemic prevention and control topics are often based on reliability and authenticity, and they hope that the acquired knowledge can play a defensive role. When these requirements cannot be met, users will lose their desire to share and communicate, hide their knowledge, and form knowledge-hiding behavior. According to Xie [70], promoting situational regulation and controlling the community atmosphere play a good role in regulating knowledge hiding. Therefore, to improve users' participation in the socialized Q&A communities, it is suggested that promoting and maintaining a virtuous circle of positive reciprocity in the community and actively paying attention to meeting user requirements are vital, so that users can fully trust and rely on the community. Knowledge-sharing behavior should be promoted by improving the organizational reward mechanism [71], thus it is suggested that the community regularly reward users who actively share knowledge publicly so that users feel respected and recognized. Certain rewards will also become the motivation for users to actively share knowledge.

(2) The results show that self-efficacy, outcome expectation, and community atmosphere negatively affect evasive hiding, playing dumb, and rationalized hiding. On one hand, according to the definition of knowledge hiding's three types from Connelly et al. [11], the occurrence of different types of knowledge hiding behavior may be affected by external incentives and user's benefits. If users lack formal contractual relationships or certain external incentives, they will not have a high willingness to share. When other users in the community ask questions, users will consciously expect to be in a mutually beneficial state. Once this does not happen, they may hide knowledge and automatically make evasive behavior. On the other hand, when answering other people's questions, some users with low self-efficacy will reduce the expectation of successfully contributing knowledge in the network knowledge space to avoid disappointing results, which leads to the occurrence of three types of knowledge hiding. The insecurity of the environment affects knowledge hiding through emotional exhaustion [72]. Hence, a fair and open environment in the communities should be ensured and monitoring channels for community managers and service providers should be established. At the same time, for some divers, the community can set appropriate restrictions, such as reading permission restrictions, to effectively reduce the users' knowledge hiding behaviors.

(3) The results show that community atmosphere positively affects outcome expectation, and outcome expectation plays a significant intermediary effect between community atmosphere and knowledge hiding behavior. According to Bandura's social cognitive theory [29], environmental factors have a certain impact on individual factors. Outcome expectation is an individual behavior, and users will make a subjective judgment on whether their input is directly proportional to their income. When users first enter the community, they are not sure whether they can get the same return through knowledge sharing because the new environment is

unfamiliar, and they do not receive any benefit from it. Based on the mentality of mutual benefit, the user's expected reward is 0, so the user will be more inclined to knowledge hiding behavior at the beginning. When users have a deeper understanding of their community atmosphere, if the community atmosphere itself is not ideal, that is., there is a fraud, inadequate incentives, and uneven distribution of material resources, people may hold a negative attitude toward reciprocity. Therefore, they are not willing to trust others easily, and their expectation of knowledge sharing also decreases. On the contrary, when the user is in an atmosphere with a strong sense of fairness, trust, and reciprocity, they will have a good expectation of the consequences of sharing knowledge. Then, the user will spend more time in Q&A communities and will be more willing to share knowledge rather than hiding knowledge. A good community atmosphere will strengthen the communication and contact between users, thus affecting users' judgment of the outcome expectation [39]. Moreover, with the communication between users, the reciprocal exchange behavior between them becomes more and more frequent, users believe that when they need help, others will take the initiative to lend a helping hand so as to achieve a satisfactory response. This will encourage both sides to produce sustained and stable knowledge sharing and contributary behavior. These findings explain the mediating effect of outcome expectation between community atmosphere and knowledge hiding behavior, and this conclusion is consistent with the research conclusions of Zhu et al. [73].

(4) The results of the model data rejected the original hypothesis H1, that is, self-efficacy negatively affects knowledge hiding behavior which is contrary to the results. This is due to the anonymity and high openness of the Q&A community. Users can hide their true identity and speak freely during use. Therefore, even if the user has a low sense of self-efficacy and does not have enough confidence to provide high-quality answers, anonymity will add a protective film to the user's psychology, thereby reducing the impact of self-efficacy on knowledge hiding. In addition, the result of this hypothesis is also related to the fact that the users of the questionnaire may not be strict with the way they handle the questionnaire, or because the questionnaire is aimed at a small number of users and only represents the views of some people.

This study synthesizes the previous research studies on knowledge hiding in a socialized Q&A community based on the social cognitive theory, expands its influencing factors to request, and expands the research boundary and theoretical knowledge related to knowledge hiding behavior. It understands and enriches the research on the behavior of the Q&A community from a dual perspective. From the user's point of view, it explores the reasons for knowledge hiding based on the user's thoughts, so as to reduce the chance of knowledge hiding generation by making more users participate in the interaction of the Q&A community, thus promoting the occurrence of knowledge sharing behavior. The reduction of knowledge hiding behavior is also conducive to further promoting the benign development of the Q&A community, so as to create a fair, interactive, and open community environment for users, thereby deepening the connotation and extension of knowledge hiding and strengthening the closeness between users and communities. In terms of health information, this study will help the public to timely master the real-time epidemic prevention and control knowledge, reduce the obstacles that may be encountered in the dissemination of relevant knowledge, help the public accurately grasp the characteristics of public health information needs, and provide an effective reference for the society to better serve the public. It is also of great significance to boost users' usage experience and to optimize knowledge community ecology.

The limitations of this study are mainly reflected in the sample data of the empirical research stage. In the phase of data collection, only 151 valid questionnaires were collected due to the restrictions of time, manpower, and other objective reasons. In the future, we will consider expanding the number of sample size and combining different interview methods, such as focus groups and one-on-one in-depth interviews. In addition, the coverage of the questionnaire in this study is insufficient and the research data are all from Zhihu, which cannot cover all types of communities. The universality of the research results for other Q&A communities needs to be studied. Therefore, further data from multiple platforms should be considered in future research studies to explore the applicability of research results and enhance the credibility of the research study. In addition, knowledge hiding is also closely related to team structure. The heterogeneity between users will affect the size of the difference. When the difference is small, knowledge hiding may also be affected. Therefore, in future research studies, we should pay more attention to diversity. At the same time, this study has less discussion on the three types of knowledge hiding. In the future, we should increase the discussion on the connotation and dimensions of knowledge hiding behavior and more comprehensively discuss whether users' knowledge hiding behavior is active or passive in order to strengthen the consideration of "tacit knowledge." [74].

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Conceptualization and methodology was done by J. -F. C. and X. F.; formal analysis was conducted by J. -F. C. and W. -Z. L.; the project was supervised by Q. Y.; project administration was handled by X. F. and W. -Z. L.; the original draft was prepared by J.-F. C. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 11905042), Natural Science Foundation of Hebei Province (No. G2021203011), and Project of Social Science Development of Hebei Province (No. 20210501003).

References

[1] Q. Q. Meng, H. X. Xiong, and Z. R. Yang, "Analysis on the health information needs and evolution of Internet users in the post epidemic period," *Library Magazine*, vol. 41, pp. 119–127, 2022.

[2] J. J. Li and S. L. Guo, "Influencing factors of knowledge adoption behavior of users in socialized Q&A community from the perspective of cognition -- an exploratory analysis based on grounded theory," *Information Science*, vol. 40, pp. 91–98, 2022.

[3] Y. Chen, L. Wang, and T. Y. Chen, "Research on user portrait of socialized Q&A platform based on social network analysis," *Journal of Information*, vol. 40, pp. 414–423, 2021.

[4] J. Gao and X. X. Wang, "Study on the influence of harmonious state experience of network community members on diving intention," *Nankai Management Review*, vol. 30, p. 1, 2021.

[5] Y. Tian, Q. L. Cao, and L. H. Mao, "Research on the theoretical origin and core content of organizational knowledge hiding," *Technical Economy and Management Research*, vol. 42, pp. 57–62, 2021.

[6] H. Liu and L. R. An, "Research on the impact of expected return and knowledge power perception on employees' knowledge hiding behavior," *Scientific Decision*, vol. 28, pp. 81–89, 2021.

[7] L. Shi, G. Song, G. Cheng, and X. Liu, "A user-based aggregation topic model for understanding user's preference and intention in social network," *Neurocomputing*, vol. 413, pp. 1–13, 2020.

[8] M. M. Shi, "Research on the influencing factors of knowledge sharing behavior of virtual learning community members from the perspective of social cognition," *China's Collective Economy*, vol. 34, pp. 54-55, 2018.

[9] Z. Shi, D. R. Lin, and L. Ding, "The impact of workplace negative gossip on hotel employees' willingness to share tacit knowledge," *Tourism Tribune*, vol. 37, pp. 108–120, 2022.

[10] X. Q. Zhao, Q. Q. Wang, and Q. Cai, "An analysis of knowledge sharing behavior characteristics of opinion leaders in online Q&A community: a case study of Zhihu "Travel" topic," *Information Science*, vol. 39, pp. 68–74, 2021.

[11] C. E. Connelly, D. Zweig, J. Webster, and J. P. Trougakos, "Knowledge hiding in organizations," *Journal of Organizational Behavior*, vol. 33, no. 1, pp. 64–88, 2011.

[12] J. Nielsen, *The 90-9-1 Rule for Participation Inequality in Social media and Online Communities*, https://www.nngroup.com/articles/participation-inequality/, 2006.

[13] T. M. Nguyen, A. Malik, and P. Budhwar, "Knowledge hiding in organizational crisis: the moderating role of leadership," *Journal of Business Research*, vol. 139, pp. 161–172, 2022.

[14] M. A. Fauzi, "A review of knowledge hiding in team: evaluation of critical research streams," *Team Performance Management: International Journal*, vol. 28, no. 5/6, pp. 281–305, 2022, ahead-of-print.

[15] V. Jafari-Sadeghi, H. Amoozad Mahdiraji, A. Devalle, and A. C. Pellicelli, "Somebody is hiding something: disentangling interpersonal level drivers and consequences of knowledge hiding in international entrepreneurial firms," *Journal of Business Research*, vol. 139, pp. 383–396, 2022.

[16] Y. Y. Huang, "Research on the mechanism of the sense of excess qualification on employees' knowledge hiding -- a regulated intermediary model," *Operations Management*, vol. 40, pp. 101–109, 2022.

[17] J. W. Li and Q. Ke, "An empirical study on the influencing factors of users' knowledge hiding behavior in virtual Q&A community," *Journal of Agricultural Library and Information*, vol. 34, pp. 1–15, 2022.

[18] X. Y. Lu, X. Q. Xu, and X. L. Wang, "Research on knowledge hiding behavior configuration of socialized Q&A community based on FsQCA," *Knowledge Management Forum*, vol. 7, pp. 37–48, 2022.

[19] M. A. Hamza, S. Rehman, A. Sarwar, and K. N. Choudhary, "Is knowledge a tenement? The mediating role of team member exchange over the relationship of big five personality traits and knowledge-hiding behavior," *VINE Journal of Information and Knowledge Management Systems*, vol. 51, 2021.

[20] A. Anand, P. Centobelli, and R. Cerchione, "Why should I share knowledge with others? A review-based framework on events leading to knowledge hiding," *Journal of Organizational Change Management*, vol. ahead-of-print, no. ahead-of-print, pp. 379–399, 2020.

[21] T. Alam, Z. Ullah, F. S. AlDhaen, E. AlDhaen, N. Ahmad, and M. Scholz, "Towards explaining knowledge hiding through relationship conflict, frustration, and irritability: the case of public sector teaching hospitals," *Sustainability*, vol. 13, no. 22, Article ID 12598, 2021.

[22] L. Zhang, J. A. Pang, and Y. X. Wang, "Research on evaluation of knowledge hiding willingness of researchers in virtual academic community based on SVM," *Information Exploration*, vol. 33, pp. 9–16, 2019.

[23] D. H. Schunk and M. K. Dibenedetto, "Motivation and social cognitive theory," *Contemporary Educational Psychology*, vol. 60, Article ID 101832, 2020.

[24] F. X. Fang, "Brief introduction to the development of children 's social cognition," *Psychological Science*, vol. 8, 1986.

[25] R. H. Shi, *Modern Social Psychology*, East China Normal University Press, Shanghai, 1989.

[26] M. H. Hsu, T. L. Ju, C. H. Yen, and C. M. Chang, "Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations," *International Journal of Human-Computer Studies*, vol. 65, no. 2, pp. 153–169, 2007.

[27] Y. Cai and W. Shi, "The influence of the community climate on users' knowledge-sharing intention: the social cognitive

theory perspective," *Behaviour & Information Technology*, vol. 41, no. 2, pp. 307–323, 2022.

[28] X. J. Zhan and T. R. Xiong, "The observed influence of supervisor's uncivilized behavior on employee's uncivilized behavior -- based on social cognitive theory," *Contemporary Finance & Economics*, vol. 41, pp. 92–102, 2021.

[29] A. Bandura, "Self-Efficacy:toward a unifying theory of behavioral change," *Psychological Review*, vol. 84, no. 2, pp. 191–215, 1977.

[30] W. Y. Zhang and S. Zhu, "Review of domestic research on emotion regulation self-efficacy," *Publica Standardization*, vol. 40, pp. 124–126, 2021.

[31] K. C. Lee, I. H. Chang, I. L. Wang, and R. S. Chen, "Effects of knowledge sharing on sustainable happiness of preschool teachers: the mediating roles of self-efficacy and helping behavior," *Current Psychology*, vol. 41, pp. 1–10, 2022.

[32] X. M. Yang and X. B. Li, "The influence of big five personality on knowledge sharing in virtual community -- taking self-efficacy as the intermediary variable," *Information Exploration*, vol. 35, pp. 1–6, 2021.

[33] H. C. Zhao and X. Q. Li, "Approaches to the influence of perceived overqualification on organizational knowledge sharing behavior - based on the moderating effect of Chinese traditional cultural values," *Journal of Zhengzhou University*, vol. 55, pp. 51–55, 2022.

[34] D. Constant, S. Kiesler, and L. Sproull, "What's mine is ours, or is it?a study of attitudes about information sharing," *Information Systems Research*, vol. 5, no. 4, pp. 400–421, 1994.

[35] G. W. Bock, R. W. Zmud, Y. G. Kim, and Lee, "Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological forces, and organizational climate," *MIS Quarterly*, vol. 29, no. 1, pp. 87–111, 2005.

[36] Y. H. Tang and J. H. Mao, "The influence mechanism of individual perception difference and workplace exclusion on knowledge sharing behavior," *Scientific Research Management*, vol. 41, pp. 200–208, 2020.

[37] M. Zhang, Z. Ma, and Y. Zhang, "The formation path of users' subjective knowledge hiding behavior in online health community," *Information Theory and Practice*, vol. 41, pp. 111–117+53, 2018.

[38] Y. G. Dai, *An Empirical Study on the Impact of Knowledge Hiding on Temporary Team Performance*, Chongqing University of Technology, Chongqing, China, 2020.

[39] Y. H. Shang, S. Z. Ai, and F. Y. Wang, "An empirical study on knowledge sharing behavior of virtual community members based on social cognitive theory," *Scientific and Technological Progress and Countermeasures*, vol. 29, pp. 127–132, 2012.

[40] S. H. Fu, Z. F. Shen, and Y. Y. Jiao, "Research on the connotation, antecedents and consequences of open innovation community atmosphere," *Scientific Research (New York)*, vol. 38, pp. 2293–2304, 2020.

[41] E. J. Kim and S. Park, "Transformational leadership, knowledge sharing, organizational climate and learning: an empirical study," *The Leadership & Organization Development Journal*, vol. 41, no. 6, pp. 761–775, 2020.

[42] H. Lu, X. X. Zhang, and Y. W. Wu, "Multiple motivations and effects of users ' knowledge contribution behavior in professional virtual community," *Library Construction*, vol. 45, pp. 131–140, 2022.

[43] J. L. Yin, F. Yang, and N. Wang, "Research on the influence mechanism of perceived community value on consumers' knowledge sharing behavior: the intermediary role of reciprocal norms," *Business and Economic Research*, vol. 40, pp. 36–40, 2021.

[44] W. B. Xia, J. Zhai, and K. L. He, "Review of Reciprocal Altruistic Behavior in Supply Chain Management," *Journal of Fujian Business University*, vol. 22, pp. 41–47, 2019.

[45] H. F. Lin, "Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions," *Journal of Information Science*, vol. 33, no. 2, pp. 135–149, 2007.

[46] X. W. Cao, "Mechanism analysis of the influence of cross-border e-commerce platform empowerment on customer locking: the intermediary effect of customer participation and trust," *Business and Economic Research*, vol. 41, pp. 106–110, 2022.

[47] X. H. Dang and Y. L. Sun, "Research on the impact of technological innovation network location on network practices -- taking inter organizational trust as an intermediary variable," *Scientific Research Management*, vol. 34, pp. 1–8, 2013.

[48] M. H. Liu, *Research on Knowledge Sharing and protection Based on Evolutionary Game*, Hebei University of Economics and Trade, Hebei, China, 2022.

[49] M. Chi, X. H. Bi, and Y. S. Xu, "Research on the impact of governance mechanism of mobile academic virtual community on knowledge sharing behavior -- taking trust as the intermediary variable," *Management Review*, vol. 33, pp. 164–175, 2021.

[50] T. Lin, S. M. Zhang, and W. W. Han, "A meta-analysis of the differences in the effect of trust on supply chain knowledge sharing," *Business and Economic Research*, pp. 9–12, 2019.

[51] Y. Zhang, R. Zhang, Q. Ma et al., *ISA Transactions*, vol. 100, pp. 210–220, 2020.

[52] D. H. Yang, H. T. Xie, and Y. Q. Dong, "The influence path of knowledge-based leadership on knowledge sharing behavior of scientific research team members -- a Study on the mediating effect of social exchange and the moderating effect of organizational justice," *Technology and Innovation Management*, vol. 42, pp. 409–416, 2021.

[53] Y. R. Hao, *Research on the Occurrence Mechanism of Employees' Knowledge Sharing Behavior Based on Fairness Perception*, Hebei University of Economics and Business, Hebei, China, 2021.

[54] J. M. Becker, K. Klein, and M. Wetzels, "Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models," *Long Range Planning*, vol. 45, no. 5-6, pp. 359–394, 2012.

[55] C. J. Chen and S. W. Hung, "To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities," *Information & Management*, vol. 47, no. 4, pp. 226–236, 2010.

[56] J. B. Rotter, "A new scale for the measurement of interpersonal trust1," *Journal of Personality*, vol. 35, no. 4, pp. 651–665, 1967.

[57] J. A. Colquitt, "On the dimensionality of organizational justice: a construct validation of a measure," *Journal of Applied Psychology*, vol. 86, no. 3, pp. 386–400, 2001.

[58] M. Yi, J. J. Song, and B. Yang, "Research on user demand hierarchy of network knowledge community," *Information Science*, vol. 35, pp. 22–26, 2017.

[59] F. Lin and H. Huang, "Why people share knowledge in virtual communities? The use of Yahoo! Kimo Knowledge+ as an example," *Internet Research*, vol. 23, no. 2, pp. 133–159, 2013.

[60] H. L. Li and L. R. An, "Review of research on knowledge hiding behavior in organizations," *Financial Economy*, vol. 37, pp. 178-179, 2018.
[61] W. Liang, Z. P. Lu, and G. F. Liu, *Research on Maker Knowledge Needs Based on Grounded Theory*, Library Information Work, vol. 62, , pp. 10–17, 2018.
[62] X. X. Zhang, Z. M. Li, S. L. Guo, and Q. Social, "A community users 'knowledge needs and their dynamic evolution," *Information Theory and Practice*, vol. 41, pp. 38–44+50, 2018.
[63] M. Li, Y. Li, and Q. Zhou, "Knowledge supply and demand in network Q & A community based on TF-PIDF," *Data Analysis and Knowledge Discovery*, vol. 42, pp. 106–115, 2021.
[64] H. F. Kaiser, "An index of factorial simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31–36, 1974.
[65] B. A. Cerny and H. F. Kaiser, "A study of a measure of sampling adequacy for factor analytic correlation matrices," *Multivariate Behavioral Research*, vol. 12, no. 1, pp. 43–47, 1977.
[66] J. F. Hair and B. J. Black, *Multivariate Data Analysis*, Prentice-Hall, Upper Saddle River, 2014.
[67] A. Akpinar, *Factors Influencing the Use of Urban Greenways: A Case Study of Aydın, Turkey*, Urban Forestry & Urban Greening, vol. 16 , pp. 123–131, 2016.
[68] K. Gu, *Research on Influencing Factors of Knowledge Hiding Behavior of Users in Virtual Community*, Shandong University of Finance and Economics, Shandong, China, 2018.
[69] R. Han, *Research on the Influence Mechanism of Organizational Injustice on Knowledge Hiding*, Jilin University, China, 2021.
[70] W. Y. Xie, *Research on the Influence of Organizational Motivation Atmosphere on Employees' Knowledge Hiding Behavior*, Xi'an University of Architecture and Technology, Xian, China, 2021.
[71] Y. Y. Su, C. Fu, and F. Y. Wei, "How to reduce knowledge hiding behavior in organizations," *China Social Sciences Journal*, vol. 20, 2022.
[72] M. Li, "Research on the influence mechanism of job insecurity on employees' knowledge hiding," *Operations Management*, vol. 40, pp. 1–15, 2022.
[73] H. Y. Zhu, W. T. Cao, and J. Chen, "The influence of class atmosphere on medical students' knowledge hiding -- the mediating effect of interpersonal trust," *Journal of Medical Informatics*, vol. 43, pp. 53–58, 2022.
[74] J. M. Wang, *Sample Size Calculation* , 2020, https://wenku.baidu.com/view/71aead7efb0f76c66137ee06eff9aef8951e4859.html.

Hindawi
Complexity
Volume 2022, Article ID 1047309, 17 pages
https://doi.org/10.1155/2022/1047309

Research Article

Establishment of Dynamic Evolving Neural-Fuzzy Inference System Model for Natural Air Temperature Prediction

Suraj Kumar Bhagat,[1] **Tiyasha Tiyasha**,[1] **Zainab Al-khafaji**,[2] **Patrick Laux**,[3] **Ahmed A. Ewees**,[4,5] **Tarik A. Rashid**,[6] **Sinan Salih**,[7] **Roland Yonaba**,[8] **Ufuk Beyaztas**,[9] **and Zaher Mundher Yaseen**[10]

[1]*Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh, Vietnam*
[2]*Building and Construction Techniques Engineering Department, AL-Mustaqbal University College, Hillah 51001, Iraq*
[3]*Institute of Meteorology and Climate Research (IMK-IFU), Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany*
[4]*Department of Information Systems, College of Computing and Information Technology, University of Bisha, Bisha 61922, Saudi Arabia*
[5]*Department of Computer, Damietta University, Damietta, Egypt*
[6]*Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil, KRI, Iraq*
[7]*Artificial Intelligence Research Unit (AIRU), Dijlah University College, Al-Dora, Baghdad, Iraq*
[8]*Laboratoire Eaux, Hydro-Systèmes et Agriculture (LEHSA), Institut International D'Ingénierie de L'Eau et de L'Environnement (2iE), 01 P. O. Box 594, Ouagadougou 01, Burkina Faso*
[9]*Department of Statistics, Marmara University, Istanbul, Turkey*
[10]*Civil and Environmental Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*

Correspondence should be addressed to Roland Yonaba; roland.yonaba@gmail.com

Received 10 June 2022; Revised 4 August 2022; Accepted 11 August 2022; Published 23 September 2022

Academic Editor: Gonzalo Farias

Air temperature (AT) prediction can play a significant role in studies related to climate change, radiation and heat flux estimation, and weather forecasting. This study applied and compared the outcomes of three advanced fuzzy inference models, i.e., dynamic evolving neural-fuzzy inference system (DENFIS), hybrid neural-fuzzy inference system (HyFIS), and adaptive neurofuzzy inference system (ANFIS) for AT prediction. Modelling was done for three stations in North Dakota (ND), USA, i.e., Robinson, Ada, and Hillsboro. The results reveal that FIS type models are well suited when handling highly variable data, such as AT, which shows a high positive correlation with average daily dew point (DP), total solar radiation (TSR), and negative correlation with average wind speed (WS). At the Robinson station, DENFIS performed the best with a coefficient of determination (R^2) of 0.96 and a modified index of agreement (md) of 0.92, followed by ANFIS with R^2 of 0.94 and md of 0.89, and HyFIS with R^2 of 0.90 and md of 0.84. A similar result was observed for the other two stations, i.e., Ada and Hillsboro stations where DENFIS performed the best with R^2: 0.953/0.960, md: 0.903/0.912, then ANFIS with R^2: 0.943/0.942, md: 0.888/0.890, and HyFIS with R^2: 0.908/0.905, md: 0.845/0.821, respectively. It can be concluded that all three models are capable of predicting AT with high efficiency by only using DP, TSR, and WS as input variables. This makes the application of these models more reliable for a meteorological variable with the need for the least number of input variables. The study can be valuable for the areas where the climatological and seasonal variations are studied and will allow providing excellent prediction results with the least error margin and without a huge expenditure.

1. Introduction

One of the commonly measured weather parameters is the air temperature (A_T), which measures the relative motion/kinetic energy of the component gases that constitute air. It increases when the molecules of a gas are moving more quickly and vice versa. A_T estimation is an important process for several applications, such as in studying vector-borne diseases [1, 2], weather forecasting, climate change [3–5], epidemic forecasting [6], veterinary uses, radiation [7], and heat flux estimation [8], estimation of water potential and vapour pressure deficit [9, 10], ecology [11–13], wastewater treatment [14–16], hydrology [17], urban land use, and urban heat island [18]. The estimation of A_T is usually conducted by weather metrological stations and is considered an essential weather parameter, which is usually measured with high accuracy [19].

1.1. Application of Classic Machine Learning Models. Improvement of the accuracies of various high-impact weather prediction models using machine learning (ML) models have been the focus of most research activities recently [20–23]. This is based on the nonreliance of ML models on input variables' multicollinearity; hence, they can process numerous input variables [24]. The development of ML-based models for a multitude of stations is achievable and as such, it is possible to monitor the spatial distribution of the prediction such as A_T, when the ML models are fed with spatially continuous input parameters [25, 26]. The postprocessing of the hourly temperature outputs of the Advanced Regional Prediction System (ARPS) using an artificial neural network (ANN) has been investigated by Marzban [27]. The study achieved an average of 40% decline in the mean squared error (MSE) for the validated weather stations. Various ANN-based models for A_T prediction during winter periods have been developed by Jain et al. [28]. The training of the developed models involved the use of patterns that included 6-hours of previous weather information, such as WS, relative humidity (RH), A_T, time of the day, and TSR. In another study by Jang et al. [29], the authors predicted A_T in Southern Quebec (Canada) based on the use of the ANN model and AVHRR images. The employed ML model was trained using Levenberg–Marquardt backpropagation (LM-BP) while the LM-BP was improved using the early stopping method to ensure the generalization of the learning process of the networks. As per Smith et al. [30], the prediction performance of A_T models during winter periods can be improved by incorporating seasonal information in the input pattern, followed by an extension of the duration of previous data to at least 24 hours. The monthly mean A_T prediction performance of ANN and Support Vector Regression (SVR) have been studied by Salcedo-Sanz et al. [31] based on the previously measured values in New Zealand and Australia. The models were also used to predict the climate indices of importance within the studied region. From the results, the SVR model outperformed the ANN model in terms of prediction performance. However, the authors reported that last years of the test set do not allow the consistency of the prediction performance of different algorithms due to the high fluctuations. Various models, ranging from simple correction (i.e., mean bias) to ML models (such as ANN and random forest (RF)), have been investigated by Eccel et al. [32] for improving the minimum A_T prediction performance of two numerical models for weather prediction. The outcome of the comparative study showed that the RF model in comparison to the other models achieved the best performance in terms of being easier to automate. An establishment of ANN-based models for A_T prediction has been developed by Smith et al. [33]. The models were developed for A_T prediction throughout the year using the data collected since 2005. The ability of the polynomial neural network to bias-correct the National Oceanic and Atmospheric Administration (NOAA) mesoscale model for hourly A_T prediction has been reported by Vashani et al. [34] while another study by Şahin [35] reported monthly mean A_T prediction using remote sensing dataset and ANN model in 20 Turkish cities. The performance of the developed ANN model in monthly mean A_T modelling using remote sensing data was reported as efficient and accurate. Moreover, deciding those hyperparameters is challenging to the nonstationary data.

1.2. Application of Hybrid Machine Learning Models. The trend of hybrid model application is growing year by year as per its scientific advantages and higher robustness. The ANFIS and ANN models have been evaluated for effectiveness in long-term monthly AT prediction at 30 Iraqi weather stations Kisi and Shiri [36]. The models were trained using the monthly data of 20 weather stations while the data for the remaining 10 stations were used for model validation. The models were also compared against each other in terms of prediction performance and the outcome showed that the ANN model performed better than the ANFIS model in the test period. Moreover, the authors suggested further investigations with other techniques and data management scenarios for the generalization of the application such as other important climatologic variables. Besides, a couple of studies applied a hybridization of adaptive neurofuzzy inference system with optimization methods using mutation Salp Swarm Algorithm as well as Grasshopper Optimization Algorithm (ANFIS-mSG) and particle swarm optimization (ANFIS-PSO) to simulate the soil temperature using univariate independent variables and the high-strength concrete shear strength using multiple independent variables [37, 38]. Both studies reported marginal performance gains compared to the performance of the ANFIS standalone model. Also, both studies reported the hybrid model is limited to the univariate, i.e., AT scenario, and needs to use more derivative data from the primary character. The study by Yi et al. [39] focused on improving the AT prediction accuracy of the Local Data Assimilation and Prediction System (LDAPS) model used in Seoul, South Korea. The study deployed SVR and linear regression models for this purpose and found that the prediction accuracy of the SVR model was higher

than that of the linear regression model. A hybrid model consisting of a regularized extreme learning machine (RELM) and a global climate model has been presented by Shin et al. [40] for seasonal prediction of field-scale daily mean AT. The hybrid model was found capable of performing accurate long-term field-scale AT prediction. The authors advised examining the appropriateness of other regression models to replace the base model. Besides, this can be applied for long-range prediction of other meteorological variables, such as solar radiation, humidity, and rainfall, which are critical meteorological variables in agricultural management. The use of various models (RF, SVR, ANN, and a multimodel ensemble (MME)) to correct the output of LDAPS models when predicting 2-day maximum and minimum AT in South Korea has been reported by Cho et al. [41]. From the results of the analysis, the MME model achieved the best generalization compared to the other three single ML models. Also, the authors suggested applying a more refined ensemble technique (i.e., weighted) for operational purposes. Moreover, [42] applied DENIFS for modelling coagulant dosage rates using an online and offline approach. The authors selected 6 features to perform that and found online approach stands alone as per R (0.80).

1.3. Research Motivation. Following the reported literature on the AT simulation, the implementation of ML models has progressed remarkably over the past decade. Yet, there is no single generalized ML that can be applied for diverse regional characteristics. Conceptually, AT phenomena highlight stochastic and nonstationary process as it is highly correlated with several synoptic climate features and hydrometeorological parameters. The introduction of a new ML model for AT is still an interesting topic for hydrology and climate scientists. Investigation of new paradigms that are reliable and robust in mimicking the AT trends is an open research domain. Thus, the current research has selected three stations, i.e., Robinson, Ada, and Hillsboro located in the USA where AT was predicted by implementing three advanced fuzzy inference system models which are ANFIS, DENFIS, and HyFIS. The selection of those three different meteorological stations is to test the feasibility of the proposed more with the variant trend of AT as those stations are located in different coordinates. Also applying the long-range prediction of other meteorological variables, such as solar radiation and others as a feature to predict AT, is the necessity of the research. Worth to mention, DENFIS and HyFIS models were modelled over the literature for different hydrometrological parameters and confirmed their feasibility such as pan evaporation [43], rainfall [44, 45], evapotranspiration [46], land surface temperature [47], crops suitability [48], and energy consumption [49].

1.4. Research Objectives. The main motivation of the current research is to investigate advanced inference system models for AT prediction. To the best of our knowledge, application of that neurofuzzy algorithm especially DENFIS in the field of AT of the specific location has never been used. The modelling procedure was adopted based on the construction of different input combinations to predict AT. The paper has been divided into four sections: The first section covers the introduction which is followed by the methodology section comprising data description, model concept, and statistical analysis. The third section covers the results and discussions based on statistical analysis done among the models and for three station datasets. Section four presents the conclusion along with recommendations for future studies.

2. Materials and Methods

This section has displayed the explanation of the simulated dataset and the applied predictive models for the A_T prediction.

2.1. Dataset Overview. In the current research, North Dakota (ND) is selected as the case study site for the A_T prediction. The climate of this region is featured by climatic variation and land use-land cover changes due to biofuel production. It is situated in the central northern great plain of North America and can be distributed into our ecoregions, i.e., the lake of Agassiz plain, the northern glaciated plains, the north-western glaciated plain, and the northwestern great plains [50]. As per the fourth national assessment report published in 2018, the northern great plains present a challenge for researchers because of their intense changes in elevation throughout the area leading to geological, ecological, and climatological fluctuations. Besides, due to the substantial increment in the temperature and change in precipitation pattern over the last decades. These climate changes may lead to an increase in temperature up to 2°F–4°F by 2050 [51]. The study has selected daily data for three stations at ND from 2015 to 2019. The selected station includes Robinson situated in the southern part of ND at latitude 47° 8′ 35.1384″, longitude −99° 46′ 44.8644″, and an elevation of 1829 m a.s.l., the second is Ada located at latitude 47° 19′ 15.96″, longitude −96° 30′ 50.04″, and an elevation of 910 m a.s.l., and Hillsboro is at latitude 47° 21′ 10.8″, longitude −96° 55′ 19.2″, and an elevation of 886 m a.s.l., as shown in Figure 1.

The study has selected four metrological characteristics of the selected areas for modelling which are average A_T expressed in degree Fahrenheit (°F), average dew point (DP) expressed in °F, total solar radiation (TSR) expressed in Langley (Ly), and average wind speed (WS) expressed in meters per hour (mph). The dataset used for stations has a sample of size $n = 1827$ and the descriptive statistics are presented in Table 1. Furthermore, Figure 2 presents the correlation analysis between the variables for three stations. As per Figure 2, A_T shows a high positive correlation with DP which is 0.98, 0.97, and 0.97 for Robinson (see Figure 2(a)), Ada (see Figure 2(b)), and Hillsboro (see Figure 2(c)) station, respectively. Similarly, the results show that WS is negatively correlated, such as −0.13, −0.18, and −0.27 for Robinson, Ada, and Hillsboro station, respectively.

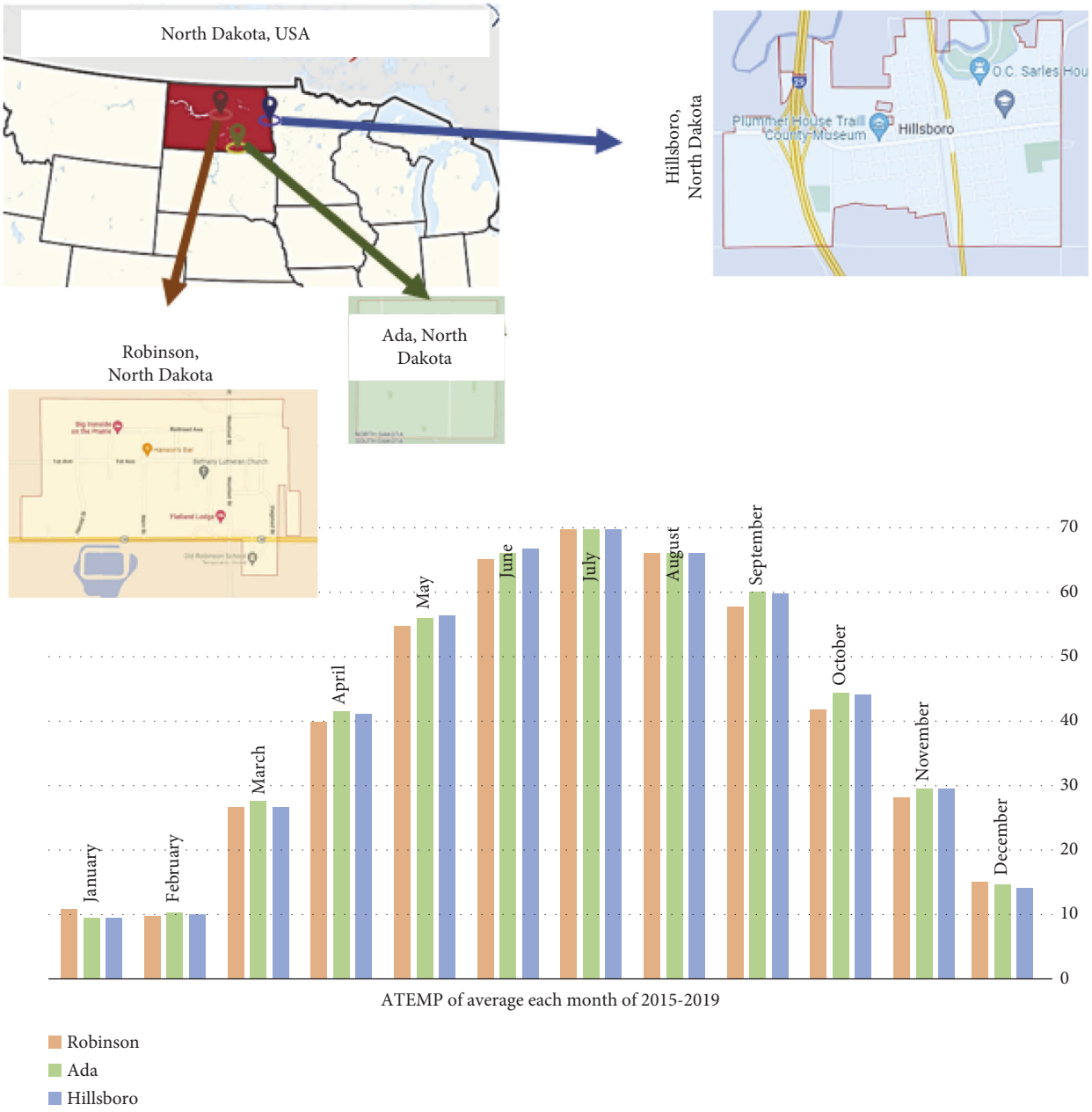


Figure 1: Study area: Ada, Hillsboro, and Robinson stations, North Dakota, North America.

2.2. Applied Predictive Models. The proposed methodology is displayed in the form of a flowchart and presented in Figure 3. Figure 3 shows three interferences in fuzzy AI predictive models. While several attempts attain to score, the best hyperparameters for the rule-based nodes of the fuzzy AI algorithms established the best target values. A detailed explanation of each method is given in the following subsections.

The individual result analysis shows that DENFIS has the highest R^2 (0.968) values when plotted in the scattered diagram (see Figure 8(a)) in comparison to ANFIS (R^2 0.949) and HyFIS (R^2 0.903) performance shown in Figures 8(b) and 7 at Robinson station. In addition to that, it is worth mentioning that Figure 8(c) shows scattered results and in some cases far from the trend line. The accuracy of the models was also evaluated in terms of Nash and MD and DENFIS performance was excellent during both the training and validation phase (Nash: 0.968 and MD: 0.919). This study has used a modified version of the Willmott formula to overcome the issues created by the presence of the outliers in the dataset which helped the study to better evaluate the model performance.

Thus, DENFIS showed the highest fitness for the Robinson station with the least prediction error (i.e., MAE) for all the considered models. The model error rates were near zero with the least outliers which shows it can handle such data with more ease than others. The scatter plot also supports the conclusion which shows the least variation

Table 1: Descriptive statistical parameters for the selected variables in the applied dataset for the analytical approach for the applied models fit.

Parameters	AT	DP	TSR	WS
Mean	40.65	32.32	13.77	9.42
Standard error	0.55	0.50	0.19	0.10
Median	43.23	32.80	12.62	8.61
Mode	60.99	58.88	6.57	5.86
Standard deviation	23.56	21.43	8.07	4.47
Sample variance	555.51	459.50	65.26	20.06
Kurtosis	−0.76	−0.46	−1.17	2.31
Skewness	−0.41	−0.45	0.33	1.19
Range	110.65	106.52	30.15	34.99
Minimum	−27.22	−33.26	0.98	0.95
Maximum	83.43	73.26	31.12	35.94
Sum	74241.17	59017.50	25159.40	17202.90
Count	1826.00	1826.00	1826.00	1826.00

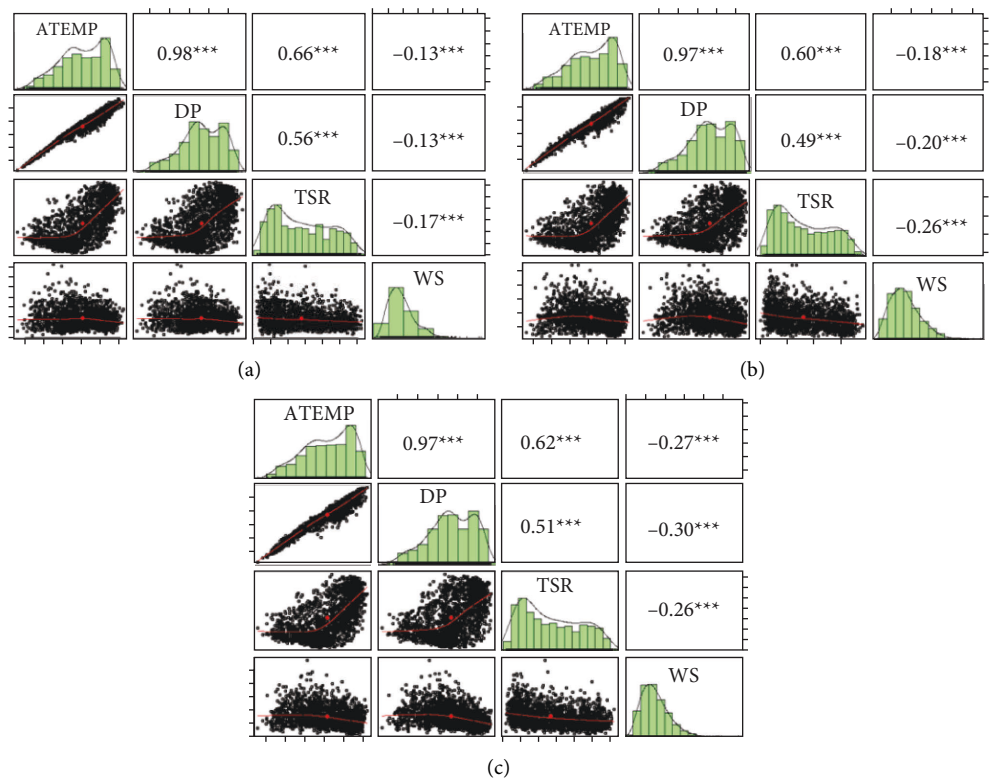


Figure 2: Statistical analysis of the applied dataset in terms of pairs plot for (a) Robinson station; (b) Ada station; (c) Hillsboro station.

from the trend line whereas the ANFIS plot is broader and HyFIS predicted values for A_T were scattered and disordered. Even though all the three models' training and validation result variation is about 10%, DENFIS was able to produce a consistent result compared to ANFIS and HyFIS.

2.2.1. Dynamic Evolving Neural-Fuzzy Inference System (DENFIS). One of the recently developed versions of neurofuzzy models is the dynamic evolving neural-fuzzy inference system (DENFIS) which, according to [52], is an extended version of the original evolving fuzzy neural networks (EFuNN). DENFIS is one of the emerging connectionist systems and its structural arrangement stemmed from the original NF models in terms of the arrangement in various layers while a block of rules made up the main core [53]. A major attribute of DENFIS is the use of a clustering procedure for input space partitioning, which is done in the original NF model using various clustering techniques, such as fuzzy c-mean clustering and grid partition (GP), subtractive clustering, etc. DENFIS relies on the so-called evolving clustering method (ECM) for input space partitioning into various regions [54, 55]. Furthermore, DENFIS uses only Takagi-Sugeno-Kang for fuzzy rule base system and triangular fuzzy membership functions (MFs) generation [56]. A recursive clustering

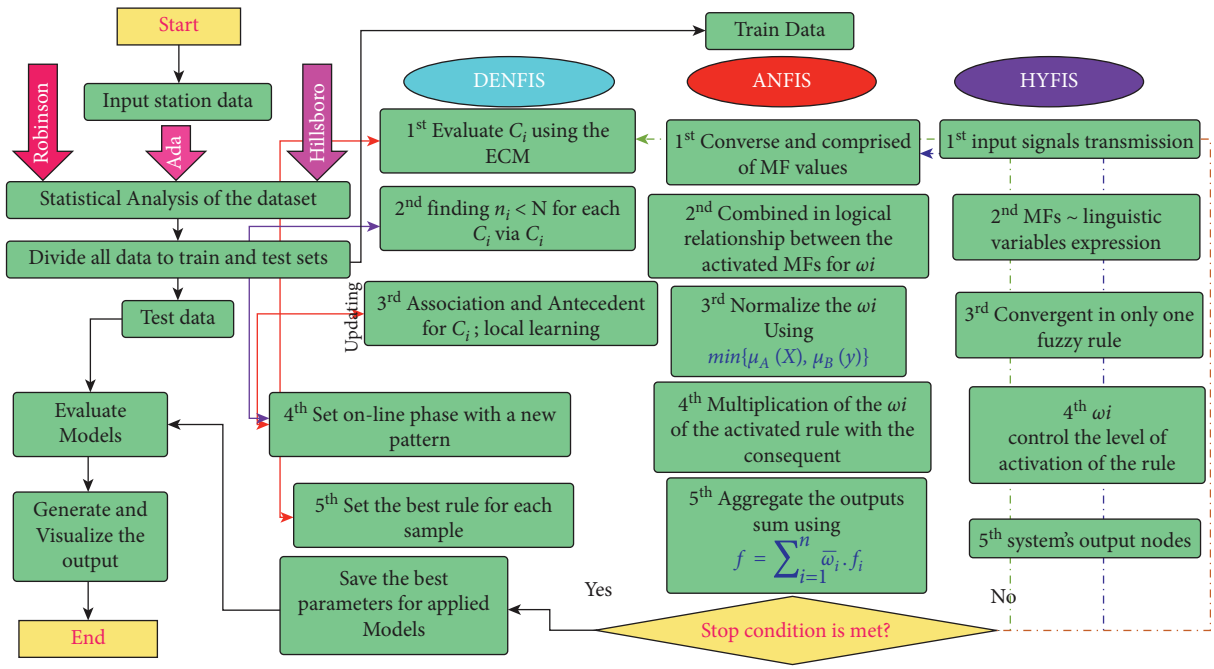


FIGURE 3: The proposed methods for the interferences fuzzy predictive models, statistical evaluators for three-station data, i.e., Robinson, Ada, and Hillsboro stations.

algorithm is used to create the rule bases. The DENFIS model can be mathematically expressed as in (1) and (2) [57]; thus:

$$\text{Rule1: if } X_1 \text{is } R_{11} \text{and } X_2 \text{is } R_{12}, \ldots, X_q \text{is } R_{1q} \text{then y is } f_1\left(X_1, X_2, X_2, \ldots, X_q\right), \tag{1}$$

$$\text{Rule2: if } X_1 \text{is } R_{21} \text{and } X_2 \text{is } R_{22}, \ldots, X_q \text{is } R_{2q} \text{then y is } f_2\left(X_1, X_2, X_2, \ldots, X_q\right), \tag{2}$$

where the predictor or input variable is represented by X_i while y represents the model output or dependent variable; R_{ij} represents the fuzzy sets while the consequent aspect of the fuzzy rules is represented [52, 54]. In the standard NF models, there is a fixed number of fuzzy rules which does not change during the training process, but in the DENFIS model, the fuzzy rules are generated, meaning that only the MFs parameters can change [52, 54]. As such, the calculation of the output of the DENFIS model only considers an aspect of the fuzzy rule base called activated rules [52, 54]. The first phase of the training process of DENFIS is the use of the ECM to cluster the input space and build the fuzzy rules. This involves two major steps which are (i) the first is formation of the antecedent part of the rules via finding the best MFs combination that will activate the cluster centre and improve the MFS efficiency; hence, the selection and formation of the antecedent part are achieved; (ii) the second part is to use the least mean estimation method to fix the consequent part of the fuzzy rules in consideration of the existing pattern within the cluster; hence, one cluster is used for each rule [52, 54, 58]. The DENFIS model involves the following steps [59]: (i) presentation of the first N samples and establishment of the cluster centre using the ECM, (ii) searching and finding $n_i < N$ example for each cluster centre C_i via closely linking to one of the cluster centers C_i, (iii) association of the fuzzy rules to the C_i with equality (rules = cluster), followed by creation of the antecedent aspects of the rules, (iv) local learning approach-based calculation of the antecedent linear parameters, (v) initiation of the first online phase with a new pattern presentation, (vi) updating the cluster partition using step (iii), (vii) creation of a new rule upon the establishment of a new cluster, followed by creation of the new consequent part, (viii) updating the linear parameters upon creation of a new cluster, (ix) carrying out the required adaptation of the related parameters, and (x) finally, reverting to step (v) for each new sample. The architecture of DENFIS is shown in Figure 4.

2.2.2. Adaptive Neurofuzzy Inference System (ANFIS). Numerous computational techniques exist which combine artificial neural networks with fuzzy systems to form new systems that are generically referred to as neurofuzzy systems [60]. The study by Jang [61] developed the ANFIS

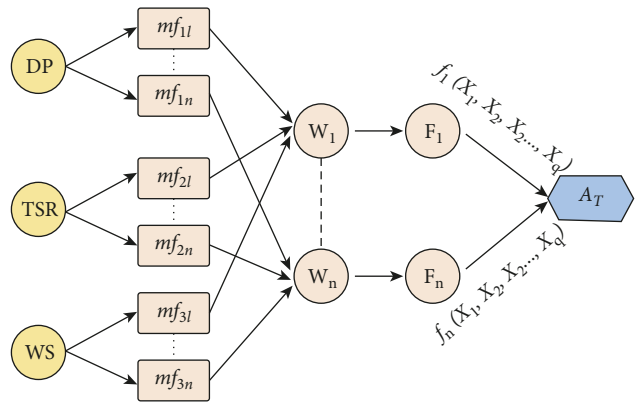


Figure 4: A schematic diagram for working flow of the DENFIS algorithm for the specific dataset.

model as a popular variant of the neurofuzzy system which mimics the human way of reasoning by combining the Takagi-Sugeno fuzzy inference systems with the RBF neural network [62]. Neurofuzzy systems rely on decision rules and fuzzy sets to deal with the impreciseness of input data and domain knowledge; it also allows quick approximation of the expected solutions [63]. Hence, these intelligent systems perform well in function approximation, real-time applications, pattern classification, etc. [60, 64]. The architecture of ANFIS is shown in Figure 5.

The fuzzification of the mode input values (x and y) is the objective in Layer 1; this implies the conversion of a set of numerical values into the equivalent fuzzy sets [65]. In this layer, the output is comprised of a set of membership values that correspond to the activation level of each MFs of the input variables: $\{\mu_{A1}(X), \ldots, \mu_{Am}(X)\}$ and $\{\mu_{B1}(y), \ldots, \mu_{B2}(y)\}$. Each node in Layer 2 corresponds to the previous part of the inference rule and depicts the likely combinations between the MFs of the first layer. In this layer, the objective is to establish the logical relationships between the activated MFs for the weight (ω_i) of each rule to be determined. The activation degree of each inference rule is calculated by applying a t-norm operator, such as minimum or algebraic product, as captured, respectively, in equations (1) and (2) discussed earlier. The objective in Layer 3 is to normalize the weights of the activated rules using equation (3) [61, 66].

$$\min\{\mu_A(X), \mu_B(y)\}, \tag{3}$$

$$\mu_A(X) \cdot \mu_B(y), \tag{4}$$

$$\overline{\omega}_i = \frac{\omega_i}{\omega_1 + \cdots + \omega_i + \cdots \omega_n}. \tag{5}$$

A set of adaptive nodes is made up of Layer 4; these nodes represent the inference rule's consequents and provide each rule's outputs. A linear function or a constant value is used to represent each consequent. In the first case, the parameters of the function are the crisp values of the input variables (x and y); the computation of the output of each Layer 4 node 4 is achieved via multiplication of the weight of the activated rule with the consequent. Lastly, Layer 5 aggregates the outputs of each node of Layer 4 nodes (using equations (4) and (5)) by computing the weighted sum; this provides the final system output as in equation (6) [61, 67].

$$f = \sum_{i=1}^{n} \overline{\omega}_i \cdot f_i. \tag{6}$$

2.2.3. Hybrid Neural-Fuzzy Inference System (HyFIS). There are two learning phases in the HyFIS [68]. Phase one is structure learning which involves the use of the knowledge acquisition module to establish the rules. Phase two is the learning of the parameters for tuning the fuzzy MFs [69] to ensure the expected level of performance will be achieved. This approach is most beneficial because the fuzzy rule base can be updated with ease when new data sets are available [70]. A new rule is created for any new set of available data pairs, followed by updating of the fuzzy rule base by this new rule (see Figure 6).

The learning phase of the neurofuzzy model in the HyFIS employs a gradient descent learning algorithm-based MLP network for adapting the fuzzy model parameters [71]. The model structure simplifies knowledge acquisition, approximate reasoning, and learning from data; it allows the use of both fuzzy rules and numerical data which brings about the benefits of the two data sources. In the HyFIS, the proposed neurofuzzy model is a multilayered ANN that combined numerous fuzzy systems. As captured in Figure 6, there are five layers in the system. In this structure, the input node is the input state signal while the output node is the output control/decision signal. The MFs and the rules are represented by the nodes in the hidden layers.

The nodes in the first layer are the inputs; their major role is input signals transmitted to the next layer. The second and fourth layers have the term nodes that serve as MFs for the input-output fuzzy linguistic variables expression. The fuzzy sets defined in this layer for the input-output variables are denoted as large (L), medium (M), and small (S). For the third layer, each of the nodes is a rule node that represents only one fuzzy rule. The certainty factor of the associated rules between Layers 4 and 5 is represented by the connection weights between the layers, meaning that the weight values control the level of activation of each rule. Finally, the

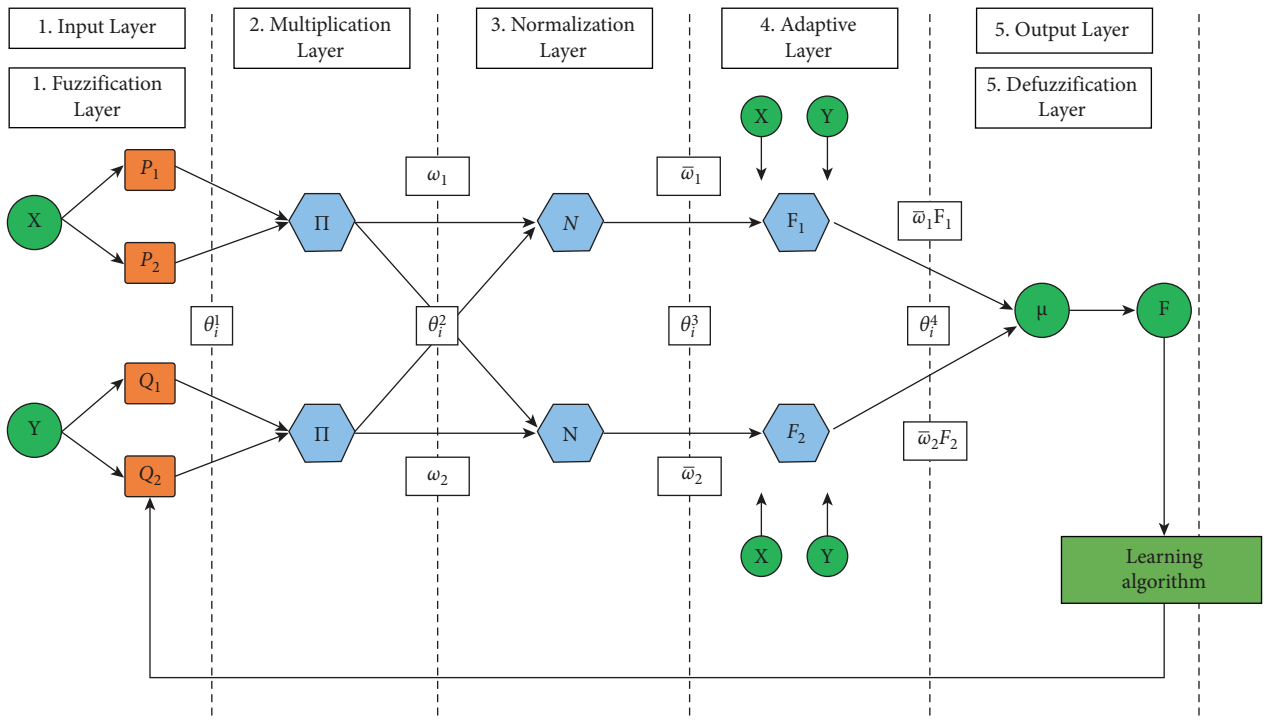


Figure 5: A schematic diagram for working flow of the ANFIS algorithm for the specific dataset.

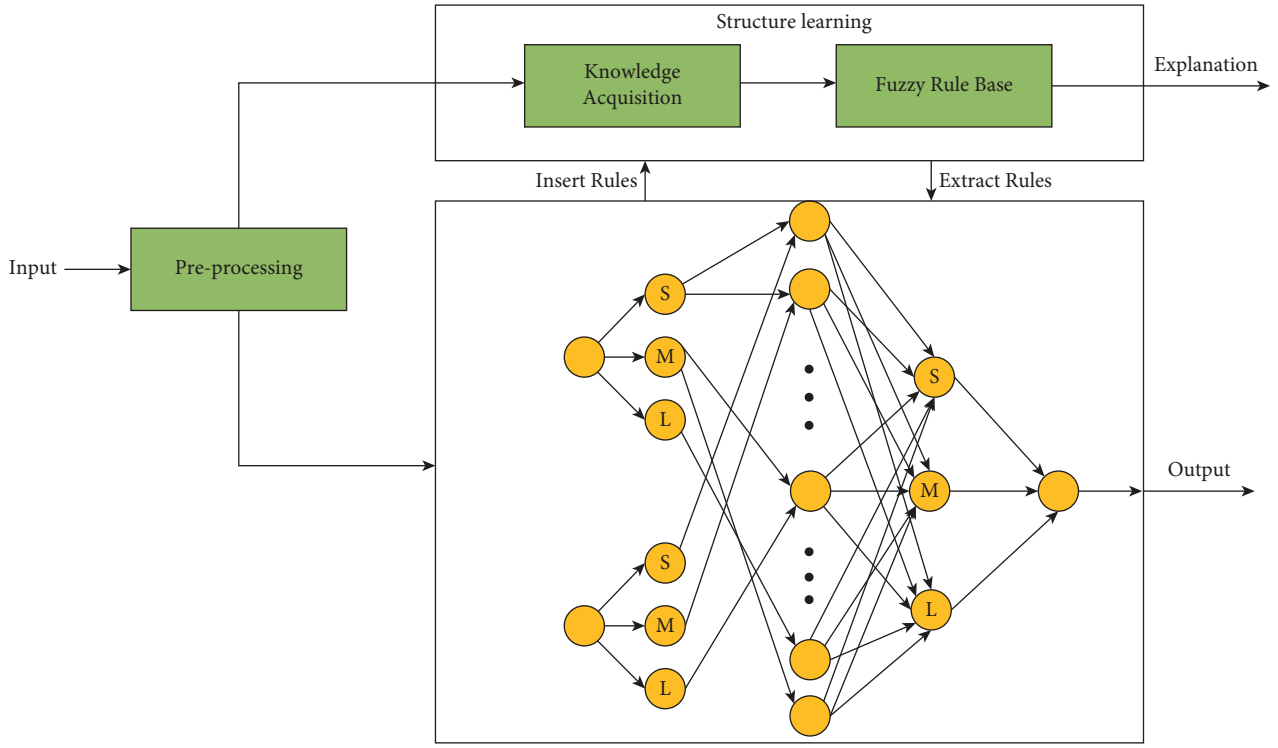


Figure 6: A schematic diagram for working flow of the HyFIS algorithm for the specific dataset.

nodes that represent the system's output are the nodes of the fifth layer.

2.3. Performance Metrics. Model competence and performance can be measured based on numerous metrics. Various performance metrics have been employed for assessing river WQ data modelling in the past two decades [72]. To gain more insight into the model performance, it is important to include the goodness of fit and absolute error measures [73]. This study applied seven commonly used metrics which are coefficient of determination (R^2), root-mean-squared error (RMSE), Nash-Sutcliffe efficiency (NSE), modified index of agreement (md), mean absolute error (MAE), and mean absolute percentage error (MAPE) [74–76] as represented in equations (7)–(12):

$$R^2 = 1 - \frac{\sum (a_i - p_i)^2}{\sum (a_i - \mu_a)^2}, \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (a_i - p_i)^2}, \tag{8}$$

$$NSE = 1 - \frac{\sum^{n} (\widehat{a}_i - p_i)^2}{\sum (p_i - \overline{Y})^2}, \tag{9}$$

$$md = 1.0 - \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |a_i - \overline{p}| + |p_i - \overline{p}|}, \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |a_i - p_i|, \tag{11}$$

$$MAPE = \sum_{i=1}^{n} |(p_i - a_i) \div a_i| \times 100 \div n, \tag{12}$$

where n is the total number of data: a denotes the output values, p denotes the real values, and μ_a is the mean value of the values, and n is the number of observations. In the current research, several statistical metrics were computed to have a more informative visualization of the applied predictive models. This is due to the limitation of some statistics such as RMSE which does not provide a sufficient error distribution. Hence, investigating more than one couple of statistical metrics can provide a more comprehensive prediction evaluation.

R software has been used for building the applied models and the statistical measurement. The applied libraries are caret, plyr, recipes, dplyr, hydroGOF, and zoo. The method CV and LOOCV have been applied. The best values of the hyperparameters have been selected.

3. Application Results and Analysis

3.1. Robinson Station. Each model performed differently based on the dataset gathered from each station. Model performance can be evaluated at different levels such as accuracy or error generated by the models. As shown in Figure 7(a), the boxplot presents the relative error (RE) produced by the three models and it can be observed that the DENFIS result shows median RE value nearest to zero with the least number of outliers. On the other hand, even though ANFIS generated an RE value closest to zero, it produced a lot of outliers in the lower quartile area. However, in the case of HyFIS, results show a high amount of RE, a huge deviation from zero, and extended whiskers due to a lot of outliers. In terms of correlation and standard deviation results, DENFIS scored the best and is thus the nearest to the actual value as presented in Figure 7(b), followed by ANFIS and HyFIS models. Updating the cluster partition using step in case of DENFIS makes it stands at the top. Furthermore, it can be concluded that DENFIS is capable of producing fewer errors in terms of RMSE: 4.031 MAE: 3.077, and MAPE: 0.159, whereas ANFIS and HyFIS generated more errors of RMSE: 5.142 and 7.271, MAE: 3.870 and 5.954, and MAPE: 0.354 and 0.277, respectively (see Table 2).

3.2. Ada Station. In the Ada station, it can be observed that the model behaviour is slightly different than the results observed in the Robinson station. The error produced by the model has a huge impact on the overall performance and as per the RMSE values, DENFIS can produce the least error and then ANFIS and HyFIS, i.e., 4.979, 5.502, and 7.025, respectively. Similarly, when testing the A_T predicting error, the MA error values were highest for HyFIS and then ANFIS and lowest for DENFIS, i.e., 5.666, 4.129, and 3.729, respectively (see Table 3). Similarly, RE values presented as a boxplot in Figure 9(a) show the deviation of the RE produced by the models from the desired value, zero. Unlike the previous model RE performance (i.e., Robinson station), all three models generated values near zero; however, all produced outliers in the low quartile of the sample population. When ranked, HyFIS show higher percentage samples in the lower quartile and similarly more outliers leading to be ranked as last whereas sample population distribution was more equally distributed for DENFIS, including the outliers. The overall model performance correlation assessment can be done using Taylor diagram in Figure 9(b) where DENFIS and ANFIS show almost the same correlation and slight diffidence compared to standard deviation results from the actual value.

To estimate the robustness and accuracy of the model in prediction A_T, Nash metrics were estimated. As presented in Table 3, DENFIS outperformed ANFIS and HyFIS with Nash values of 0.952, 0.941, and 0.904, respectively. However, Nash is sensitive to outliers; thus, it is relevant to measure the model performance with other metrics such as R^2 and Md. As visualized in Figures 10(a)–10(c), DENFIS showed best-fit values when the scatter plot was done with an R^2 value of 0.953; ANFIS showed little variation from the trend line with an R^2 value of 0.943; on the contrary, the HyFIS plot was more dispersed with an R^2 value of 0.908. In the case of DENFIS, the creation of a new rule upon the establishment of a new cluster, followed by the creation of the new consequent part, made it outperform ANFIS and HyFIS.

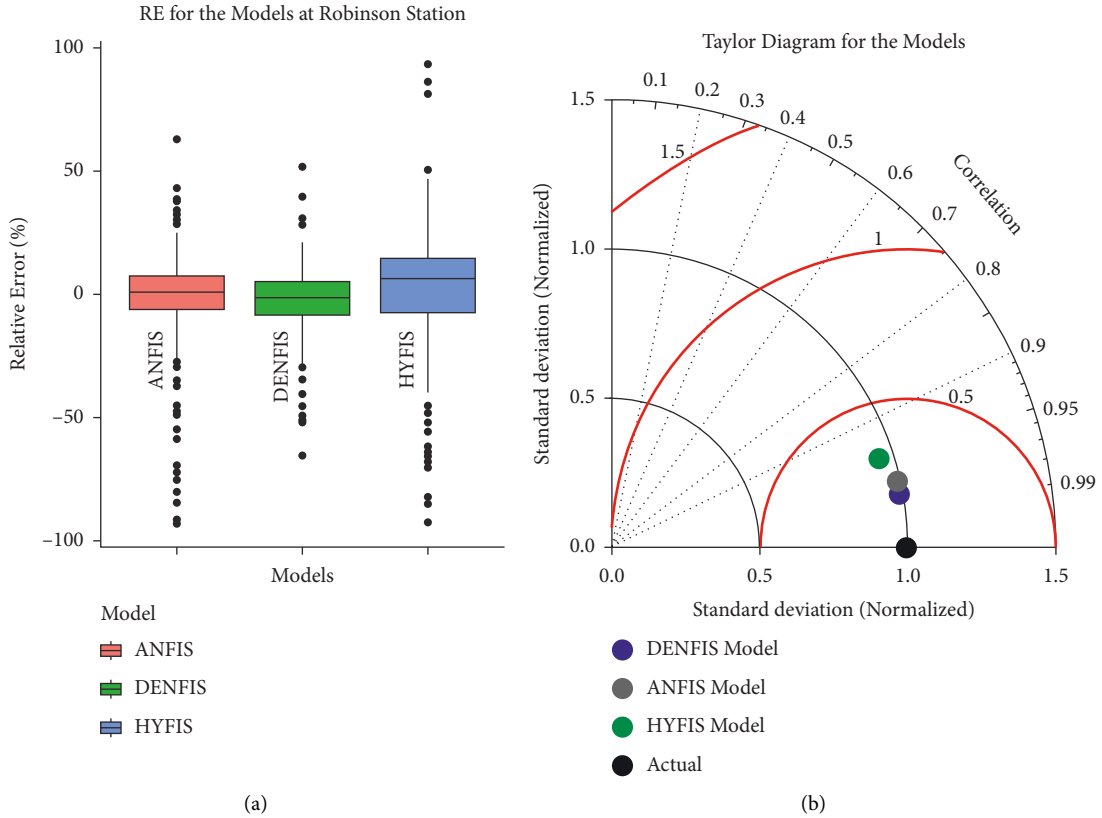


Figure 7: (a) Boxplot of residual error produced by all the models at Robinson station; (b) Taylor diagram with the comparative performance of the models at Robinson station.

Table 2: Performance metrics at Robinson station for A_T modelling.

			Training			
Models	R^2	RMSE	MAE	MAPE	Nash	MD
DENFIS	**0.971**	**4.082**	**3.076**	**0.134**	**0.971**	**0.923**
ANFIS	0.949	5.485	4.225	0.370	0.947	0.894
HyFIS	0.919	7.157	5.769	0.312	0.909	0.856
			Testing			
Models	R^2	*RMSE*	*MAE*	*MAPE*	*Nash*	*MD*
DENFIS	**0.968**	**4.031**	**3.077**	**0.159**	**0.968**	**0.919**
ANFIS	0.949	5.142	3.870	0.354	0.949	0.899
HyFIS	0.904	7.271	5.954	0.277	0.897	0.845

It can be observed from the training and validation result evaluation that DENFIS outperformed others, and ANFIS performance was a little behind DENFIS; nonetheless, HyFIS performance improved from the training to validation phase in terms of error production with MAPE decreased from 0.961 to 0.375 and accuracy of MD was improved from 0.797 to 0.845. For the Ada station dataset, DENFIS is the highest performing model and ANFIS is a good and robust model.

3.3. Hillsboro Station. The model error rate for the Hillsboro dataset for A_T prediction showed a similar pattern as discussed for other stations and DENFIS and ANFIS mean values were near zero in comparison to HyFIS. Figure 11(a) clearly shows that HyFIS sample population distribution is skewed and more deviated towards the lower quadrant. It can also be observed that when dealing with this dataset the model produced lots of outliners in both extents of the quadrants. Furthermore, when Figure 11(b) is perceived, it is apparent that the Taylor diagram shows that DENFIS is exceedingly correlated with actual value, even though ANFIS is not far behind.

In terms of accuracy, Figure 12 was able to specify the individual performance of the model when predicting A_T. Figures 12(a) and 12(b) evaluations show that the values are near the trend line and among all DENFIS show the best fit with R^2 of 0.960. Contrarily, HyFIS shows a more random and disorganized pattern and is away from the trend line (see Figure 12(c)). This result can be supported by the evaluation results produced by Nash and Md as in Table 4. DENFIS accuracy was highest with Nash: 0.960 and MD: 0.912, followed by ANFIS and HyFIS with Nash: 0.941 and 0.873 and MD: 0.890 and 821, respectively. Regarding the other error metrics such as RMSE, MAE, and MAPE, HyFIS generated the maximum number of errors during the prediction with RMSE: 8.162, MAE: 6.693, and MAPE: 1.716. On the contrary, the error caused by the DENFIS and ANFIS was almost 50% less than HyFIS concerning MAE and RMSE.

In the overall assessment between testing and training runs, DENFIS and ANFIS gave similar results except for the

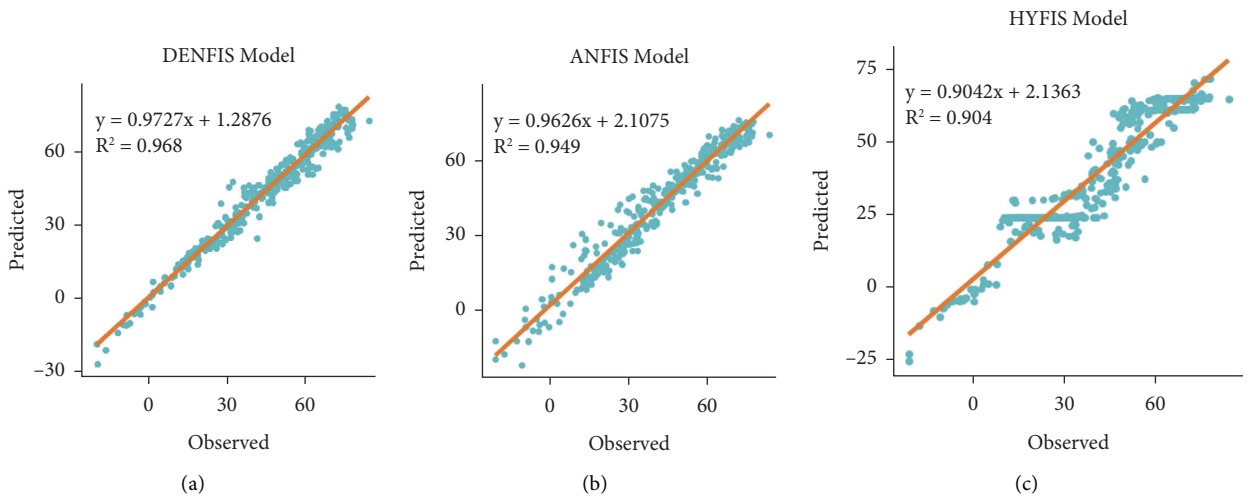


Figure 8: Scatter plot for (a) DENFIS; (b) ANFIS; (c) HyFIS models at Robinson station.

Table 3: Performance metrics at Ada station for A_T modelling.

			Training			
Models	R^2	RMSE	MAE	MAPE	Nash	MD
DENFIS	**0.963**	**4.671**	**3.454**	**0.229**	**0.963**	**0.916**
ANFIS	0.949	5.852	4.392	0.333	0.942	0.889
HyFIS	0.821	10.320	7.914	0.961	0.821	0.797
			Testing			
Models	R^2	*RMSE*	*MAE*	*MAPE*	*Nash*	*MD*
DENFIS	**0.953**	**4.979**	**3.729**	0.319	**0.952**	**0.903**
ANFIS	0.943	5.502	4.129	**0.270**	0.941	0.889
HyFIS	0.908	7.025	5.666	0.375	0.904	0.845

MAPE value of DENFIS. MAPE was much higher in a testing run but other errors were slightly less. DENFIS generated RMSE: 4.596, MAE: 3.417, and MAPE: 0.734 followed by ANFIS with RMSE: 5.578, MAE: 4.161 and little more MAP (1.707) error in comparing the other two errors for evaluation. The error devised by MAPE can be due to the high forecast in this study and since MAPE has no upper limit it can sometimes lead to difficulty in the assessment. The marginal uplifted value by DENFIS might be due to updating of the linear parameters upon the creation of a new cluster, following carrying out the required adaptation of the related parameters.

4. Discussion and Comparative Analysis

Among all stations, the DENFIS model worked well for Robinson and possibly applied the Willmott formula to overcome the issues created by the presence of the outliers in the dataset and the lower correlation in case of WS and the marginal difference in case of DP and TSR. Also, it has been observed that ANFIS worked better than HyFis in the case of Ada and Hillsboro due to a lower correlation with WS; so an upper than 27 in negative relation makes ANFIS work better. Few previous studies have been done where A_T was predicted using other models and has been discussed in this study to assess the possible future aspect of utilizing FIS type of models. A study conducted by Karthika and Deka [77] predicted A_T by applying wavelet-ANFIS and ANFIS at Bhadra station, Karnataka, India. The result showed the highest R^2: 0.95 for Db4 Gauss wavelet-ANFIS, and ANFIS produced poor performance, i.e., R^2: 0.39. On the contrary, in this study, DENFIS performed the best with R^2: 0.953–0.968 and for ANFIS R^2 was 0.942–0.949.

Similarly, another study reported the prediction of minimum, mean, and maximum AT over southwest Asia by applying ANFIS with genetic algorithm (GA), particle swarm optimization (PSO), and ant colony optimization for continuous domains (ACOR), and differential evolution (DE). The performance of these models, i.e., ANFIS, ANFIS-ACOR, ANFIS-GA, ANFIS-DE, and ANFIS-PSO in predicting max AT in terms of R^2 was 0.88, 0.95, 0.93, 0.94, and 0.90, and for min AT the R^2 were 0.72, 0.93, 0.93, 0.93, and 0.93, and for mean AT R^2 were 0.55, 0.88, 0.92, 0.90, and 0.91 [78]. It is evident from this hybrid ANFIS model that performance varied between 0.88 and 0.95 and the conventional ANFIS performance fluctuated between 0.55 and 0.88; at the same time, the ANFIS model in this study performed between 0.942 and 0.949 which shows the model accuracy was considerably improved in the current research. In addition to that, the new model DENFIS and HyFIS also performed well when handling different datasets, i.e., R^2 0.953–0.968 and 0.904–0.908, respectively. Another study set ANFIS (R^2 0.945) better than the dynamic thermal exchange model, i.e., energy balance equation (EBE) (R^2 0.743), respectively, with the small size data, though this research suits the reliable proposing DENFIS along with ANFIS and HyFIS [79]. Moreover, [80] reported that the SVR (R^2 0.95) outperformed the ANN model too with the limited scenarios of the applied data, where the current research fills the gap by performing those adequately created scenarios to set the reliable application of the DENFIS to the real world. Those overcome could be possible due to several possible advantages of the DENFIS algorithm such as fuzzy rules that are generated, meaning that only the MFs parameters are

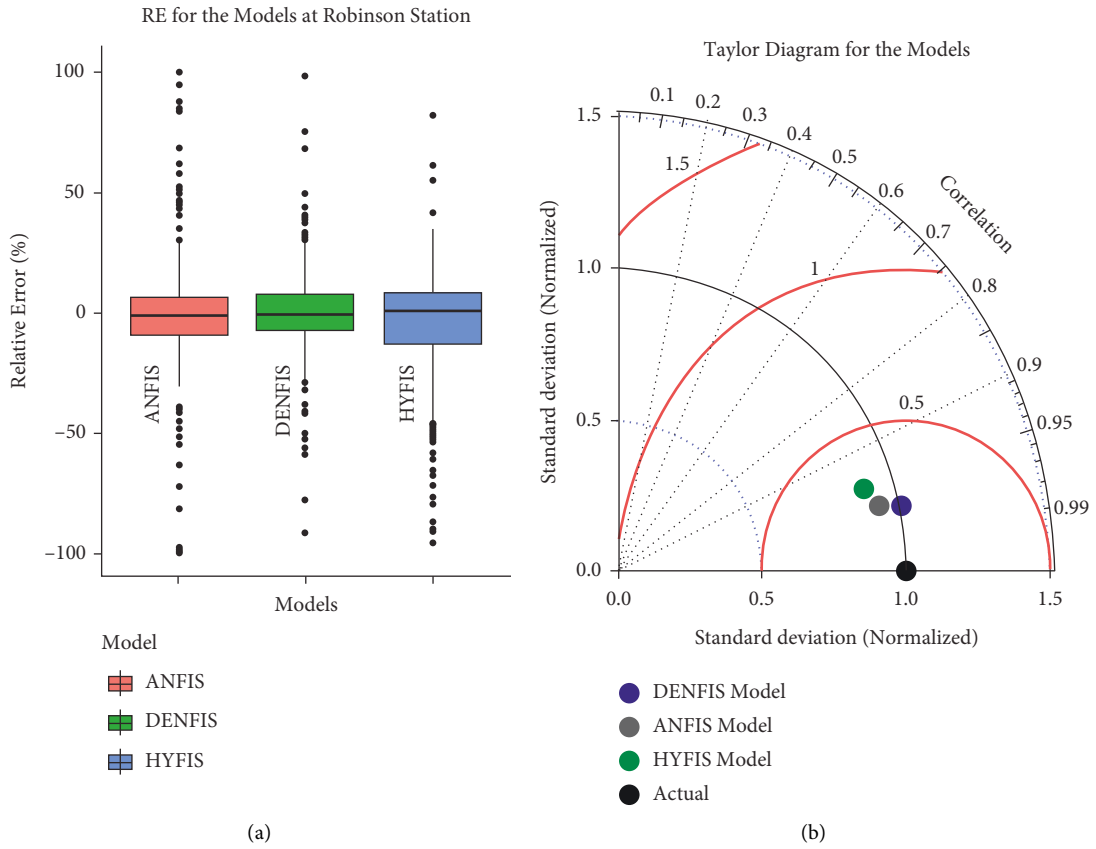


Figure 9: (a) Boxplot of residual error produced by all the models at Ada station; (b) Taylor diagram with the comparative performance of the models at Ada station.

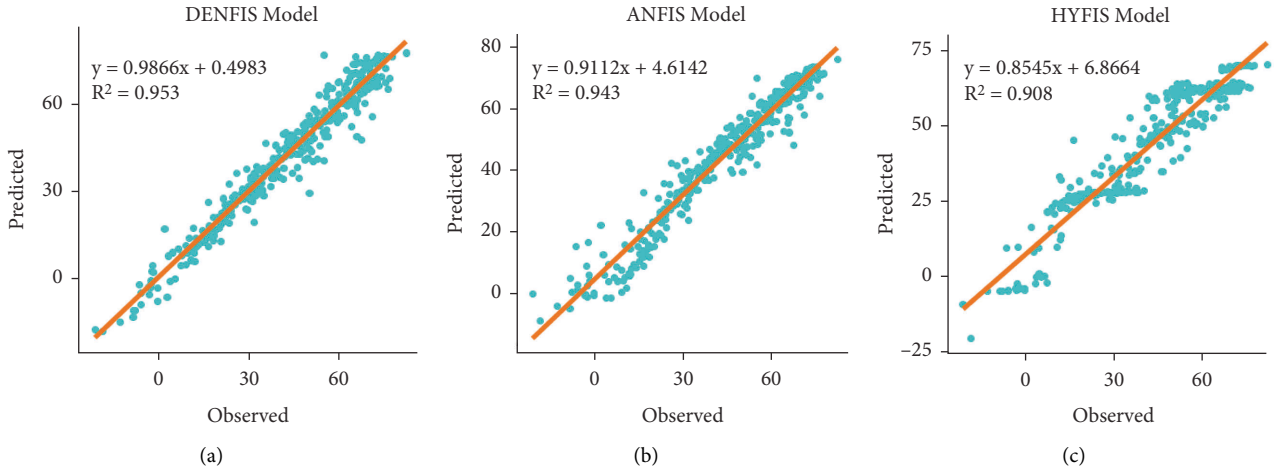


Figure 10: Scatter plot for (a) DENFIS; (b) ANFIS; (c) HyFIS models at Ada station.

calculated adequately and the activated rules for the fuzzy rules with the best performance have been investigated. Recently, [81] reported the ANFIS and DENFIS with a marginal difference using short-ranged data (of soil moisture) compared with the current study using long-ranged data. The authors also mentioned that ANFIS (step size: 0.001, membership type: Gaussian) and DENFIS (Max iteration: 3000, Step size: 0.01) were overcome with the HyFIS architect. Also, the Gaussian membership function of ANFIS remains the best performer. Another study reveals that the hybridization of DENFIS with two advanced metaheuristic optimization algorithms (i.e., Whale Optimization Algorithm (WOA) and Bat Algorithm (BA)) showed the potential predictive capacity as per the R^2

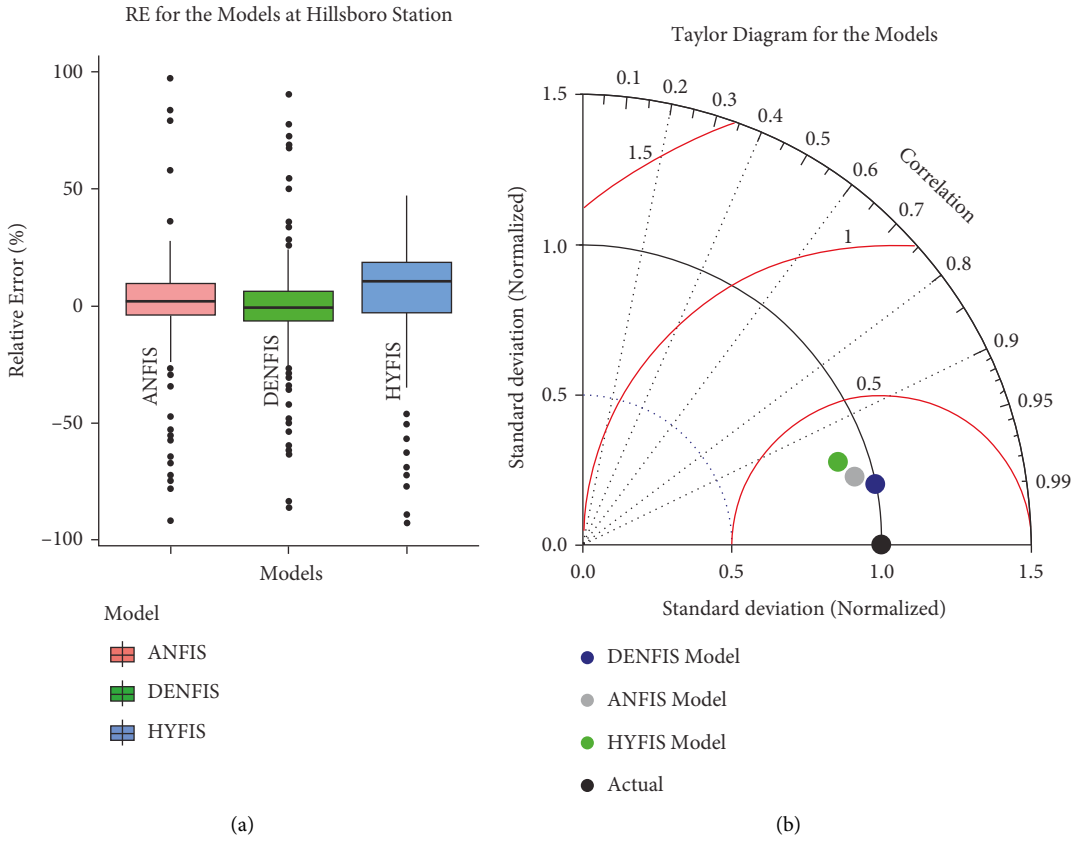


Figure 11: (a) Boxplot of residual error for models at Hillsboro station; (b) Taylor diagram of models at Hillsboro station.

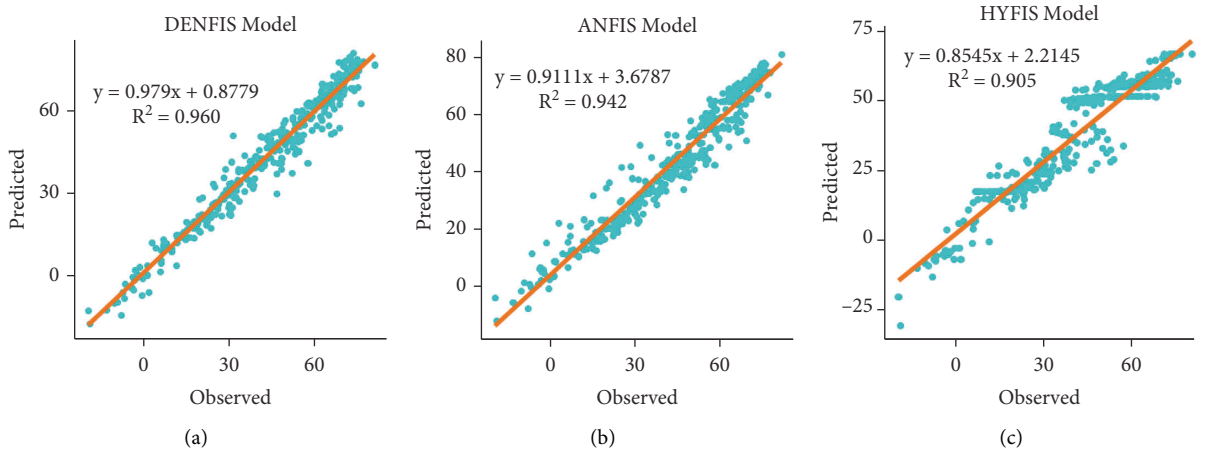


Figure 12: Scatter plot for (a) DENFIS; (b) ANFIS; (c) HyFIS models at Hillsboro station.

Table 4: Performance metrics at Hillsboro station for A_T modelling.

			Training			
Models	R^2	RMSE	MAE	MAPE	Nash	MD
DENFIS	**0.964**	**4.672**	**3.457**	**0.217**	**0.964**	**0.917**
ANFIS	0.946	5.924	4.504	0.658	0.941	0.888
HyFIS	0.917	8.373	6.760	0.338	0.883	0.831
			Testing			
Models	R^2	*RMSE*	*MAE*	*MAPE*	*Nash*	*MD*
DENFIS	**0.960**	**4.596**	**3.417**	**0.734**	**0.960**	**0.912**
ANFIS	0.942	5.578	4.161	1.707	0.941	0.890
HyFIS	0.905	8.162	6.693	1.716	0.873	0.821

(0.85–0.94) against the counterpart MARS to predict the daily scale evapotranspiration for three different coastal locations [46].

5. Conclusion

The current research reviewed the prediction competencies of advanced FIS type models such as DENFIS, ANFIS, and HyFIS. The models were able to successfully predict AT (target variable) with high accuracy and less error while using only three input variables, i.e., DP, TSR, and WS. The models were applied for three datasets acquired from three stations in North Dakota, USA, i.e., Robinson, Ada, and Hillsboro, from 2015 to 2019. Among the three applied models, DENFIS outperformed the others, followed by ANFIS and HyFIS for all three stations. The performance efficiency of DENFIS in Robinson, Ada, and Hillsboro stations was excellent with R^2: 0.97/0.95/0.96, RMSE: 4.0/4.9/4.6, and md: 0.91/0.90/0.92, respectively. Following DENFIS results, ANFIS also performed well with R^2: 0.95/0.94/0.94, RMSE: 5.1/5.5/5.6, and md: 0.90/0.89/0.89, respectively. Lastly, HyFIS similarly performed well with R^2: 0.90/0.90/0.87, RMSE: 7.3/7.0/8.2, and md: 0.84/0.79/0.82, respectively. North Dakota has reported significantly several AT complementary relations with other parameters of the sciences and engineering; for example, the evapotranspiration has been increasing over the period [82], in borehole paleoclimatology directly linked with AT [83], and snowpack control alteration [84], which lead many challenging to resources management. A report says that the lower-income area has much potential for green emission of air pollution [85]. The study is helpful to design the accurate decision priory and appropriately as per the local geographical location. Even the study applied the black box type models which have their limitations; however, they can simplify the assessment and prediction method when dealing with the meteorological data which plays an imperative role in various environmental, climatological, and meteorological studies. One of the current burning topics in these fields is climate change and such studies which applied ML methods for data modelling with similar statistical characteristics should be studied more to make a more precise projection of the future changes which will change the Earth environment. It is worth mentioning that more research should be done considering different geological conditions, diverse model types, and various diverse input variables. Also, another different meteorological parameter can be used to support the sustainable water resources and agricultural systems.

Data Availability

Data can be shared upon request from the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest in this study

References

[1] M. C. Thomson, S. J. Connor, P. J. M. Milligan, and S. P. Flasse, "The ecology of malaria—as seen from Earth-observation satellites," *Annals of Tropical Medicine and Parasitology*, vol. 90, no. 3, pp. 243–264, 1996.

[2] S. J. Goetz, S. D. Prince, and J. Small, "Advances in satellite remote sensing of environmental variables for epidemiological applications," *Advances in Parasitology*, vol. 47, pp. 289–307, 2000.

[3] C. J. Kucharik, S. P. Serbin, S. Vavrus, E. J. Hopkins, and M. M. Motew, "Patterns of climate change across Wisconsin from 1950 to 2006," *Physical Geography*, vol. 31, no. 1, pp. 1–28, 2010.

[4] D. Bocchiola and G. Diolaiuti, "Evidence of climate change within the adamello glacier of Italy," *Theoretical and Applied Climatology*, vol. 100, no. 3–4, pp. 351–369, 2010.

[5] B. Halder, M. Haghbin, and A. A. Farooque, "An assessment of urban expansion impacts on land transformation of rajpur-sonarpur municipality," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 3, pp. 34–53, 2021.

[6] L. Bian, L. Li, and G. Yan, "Combining global and local estimates for spatial distribution of mosquito larval habitats," *GIScience and Remote Sensing*, vol. 43, no. 2, pp. 128–141, 2006.

[7] A. Sharafati, K. Khosravi, P. Khosravinia et al., "The potential of novel data mining models for global solar radiation prediction," *International journal of Environmental Science and Technology*, vol. 16, no. 11, pp. 7147–7164, 2019.

[8] N. A. Brunsell, D. B. Mechem, and M. C. Anderson, "Surface Heterogeneity Impacts on Boundary Layer Dynamics via Energy Balance Partitioning," *Atmospheric Chemistry and Physics*, vol. 11, 2011.

[9] K. Aasamaa and A. Sõber, "Stomatal sensitivities to changes in leaf water potential, air humidity, CO2 concentration and light intensity, and the effect of abscisic acid on the sensitivities in six temperate deciduous tree species," *Environmental and Experimental Botany*, vol. 71, no. 1, pp. 72–78, 2011.

[10] B. Bickici Arikan, L. Jiechen, I. Sabbah, A. Ewees, R. Homsi, and S. O. Sulaiman, "Dew point time series forecasting at the north Dakota," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 2, pp. 24–34, 2021.

[11] S. W. Myint, A. Brazel, G. Okin, and A. Buyantuyev, "Combined effects of impervious surface and vegetation cover on air temperature variations in a rapidly expanding desert city," *GIScience and Remote Sensing*, vol. 47, no. 3, pp. 301–320, 2010.

[12] L. C. Smith, "Agents of change in the new north," *Eurasian Geography and Economics*, vol. 52, no. 1, pp. 30–55, 2011.

[13] S. Heding, L. Kai, C. Hanchun, C. Xianlong, H. Yongan, and S. Zhiyi, "Experimental ecology and hibernation of onchidium struma (gastropoda: pulmonata: systellommatophora)," *Journal of Experimental Marine Biology and Ecology*, vol. 396, no. 2, pp. 71–76, 2011.

[14] Z. M. Yaseen, T. T. Zigale, S. Q. Salih et al., "Laundry wastewater treatment using a combination of sand filter, biochar and teff straw media," *Scientific Reports*, vol. 9, no. 1, pp. 18709–18711, 2019.

[15] S. K. Bhagat, Tiyasha, and D. N. Bekele, "Economical approaches for the treatment and Re utilization of laundry wastewater - a review," *Journal of Industrial Pollution Control*, vol. 34, no. 2, pp. 2164–2178, 2018.

[16] S. K. Bhagat and Tiyasha, "Impact of millions of tones of effluent of textile industries: analysis of textile industries effluents in Bhilwara and an approach with bioremediation," *International Journal of ChemTech Research*, vol. 5, no. 3, pp. 1289–1298, 2013.

[17] S. K. Jain, S. K. Jain, V. Hariprasad, and A. Choudhry, "Water balance study for a basin integrating remote sensing data and GIS," *Journal of the Indian Society of Remote Sensing*, vol. 39, no. 2, pp. 259–270, 2011.

[18] S. Cheval, A. Dumitrescu, and A. Bell, "The urban heat island of Bucharest during the extreme high temperatures of July 2007," *Theoretical and Applied Climatology*, vol. 97, no. 3–4, pp. 391–401, 2009.

[19] M. Xu, S. Kang, H. Wu, and X. Yuan, "Detection of spatio-temporal variability of air temperature and precipitation based on long-term meteorological station observations over Tianshan Mountains, Central Asia," *Atmospheric Research*, vol. 203, pp. 141–163, 2018.

[20] S. Sim, J. Im, S. Park, H. Park, M. H. Ahn, and P. W. Chan, "Icing detection over East Asia from geostationary satellite data using machine learning approaches," *Remote Sensing*, vol. 10, no. 4, p. 631, 2018.

[21] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 155–170, 2019.

[22] K. Ahmed, D. A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan, "Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms," *Atmospheric Research*, vol. 236, Article ID 104806, 2020.

[23] R. C. Deo and M. Şahin, "Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia," *Atmospheric Research*, vol. 153, pp. 512–525, 2015.

[24] S. K. Bhagat, T. Tiyasha, T. M. Tung, R. R. Mostafa, and Z. M. Yaseen, "Manganese (Mn) removal prediction using extreme gradient model," *Ecotoxicology and Environmental Safety*, vol. 204, Article ID 111059, 2020.

[25] R. F. Chevalier, G. Hoogenboom, R. W. McClendon, and J. A. Paz, "Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks," *Neural Computing & Applications*, vol. 20, no. 1, pp. 151–159, 2011.

[26] Q. Fang, Z. Li, Y. Wang, M. Song, and J. Wang, "A neural-network enhanced modeling method for real-time evaluation of the temperature distribution in a data center," *Neural Computing & Applications*, vol. 31, no. 12, pp. 8379–8391, 2019.

[27] C. Marzban, "Neural networks for postprocessing model output: ARPS," *Monthly Weather Review*, vol. 131, no. 6, pp. 1103–1111, 2003.

[28] A. Jain, R. W. McClendon, G. Hoogenboom, and R. Ramyaa, "Prediction of Frost for Fruit protection Using Artificial Neural Networks," *The American Society of Agricultural and Biological Engineers*, pp. 3–3075, 2003.

[29] J.-D. Jang, A. A. Viau, and F. Anctil, "Neural network estimation of air temperatures from AVHRR data," *International Journal of Remote Sensing*, vol. 25, no. 21, pp. 4541–4554, 2004.

[30] B. A. Smith, R. W. McClendon, and G. Hoogenboom, "Improving air temperature prediction with artificial neural networks," *International Journal of Computational Intelligence*, vol. 3, no. 3, pp. 179–186, 2006.

[31] S. Salcedo-Sanz, R. C. Deo, L. Carro-Calvo, and B. Saavedra-Moreno, "Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms," *Theoretical and Applied Climatology*, vol. 125, no. 1-2, pp. 13–25, 2015.

[32] E. Eccel, L. Ghielmi, P. Granitto, R. Barbiero, F. Grazzini, and D. Cesari, "Prediction of Minimum Temperatures in an alpine Region by Linear and Non-linear post-processing of Meteorological Models," *Nonlinear Processes in Geophysics*, vol. 14, 2007.

[33] B. A. Smith, G. Hoogenboom, and R. W. McClendon, "Artificial neural networks for automated year-round temperature prediction," *Computers and Electronics in Agriculture*, vol. 68, no. 1, pp. 52–61, 2009.

[34] S. Vashani, M. Azadi, and S. Hajjam, "Comparative Evaluation of different post processing methods for numerical prediction of temperature forecasts over Iran," *Research Journal of Environmental Sciences*, vol. 4, no. 3, pp. 305–316, 2010.

[35] M. Şahin, "Modelling of air temperature using remote sensing and artificial neural network in Turkey," *Advances in Space Research*, vol. 50, no. 7, pp. 973–985, 2012.

[36] O. Kisi and J. Shiri, "Prediction of long-term monthly air temperature using geographical inputs," *International Journal of Climatology*, vol. 34, no. 1, pp. 179–186, 2014.

[37] L. Penghui, A. A. Ewees, B. H. Beyaztas et al., "Metaheuristic optimization algorithms hybridized with artificial intelligence model for soil temperature prediction: novel model," *IEEE Access*, vol. 8, pp. 51884–51904, 2020.

[38] A. Sharafati, M. Haghbin, M. S. Aldlemy et al., "Development of advanced computer aid model for shear strength of concrete slender beam prediction," *Applied Sciences*, vol. 10, no. 11, p. 3811, 2020.

[39] C. Yi, Y. Shin, and J.-W. Roh, "Development of an urban high-resolution air temperature forecast system for local weather information services based on statistical downscaling," *Atmosphere*, vol. 9, no. 5, p. 164, 2018.

[40] J.-Y. Shin, K. R. Kim, and J.-C. Ha, "Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management," *Agricultural and Forest Meteorology*, vol. 281, Article ID 107858, 2020.

[41] D. Cho, C. Yoo, J. Im, and D. Cha, "Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas," *Earth and Space Science*, vol. 7, no. 4, 2020.

[42] S. Heddam and N. Dechemi, "A New Approach Based on the Dynamic Evolving Neural-Fuzzy Inference System (DENFIS) for Modelling Coagulant Dosage (Dos): Case Study of Water Treatment Plant of Algeria," *Desalin. Water Treat*, vol. 53, 2015.

[43] O. Eray, C. Mert, and O. Kisi, "Comparison of multi-gene genetic programming and dynamic evolving neural-fuzzy inference system in modeling pan evaporation," *Hydrology Research*, vol. 49, no. 4, pp. 1221–1233, 2017.

[44] A. Talei, L. H. C. Chua, and C. Quek, "Hydrological modelling with a dynamic neural fuzzy inference system," in *HIC 2012: Understanding Changing Climate and Environment and Finding Solutions: Proceedings of the 10th International Conference on Hydroinformatics*p. 1, Hamburg, Germany, 2012.

[45] A. Anand, A. S. Dinesh, P. K. Srivastava, S. K. Chaudhary, A. K. Varma, and P. Kumar, "Rainfall rate estimation over

India using global precipitation measurement's microwave imager datasets and different variants of fuzzy information system," *Geocarto International*, pp. 1–19, 2021.

[46] L. Ye, M. M. A. Zahra, N. K. Al-Bedyry, and Z. M. Yaseen, "Daily scale evapotranspiration prediction over the coastal region of southwest Bangladesh: new development of artificial intelligence model," *Stochastic Environmental Research and Risk Assessment*, vol. 36, no. 2, pp. 451–471, 2022.

[47] E. K. Mustafa, Y. Co, G. Liu et al., "Study for predicting land surface temperature (LST) using landsat data: a comparison of four algorithms," *Advances in Civil Engineering*, vol. 2020, Article ID 7363546, 16 pages, 2020.

[48] K. B. Dang, B. Burkhard, W. Windhorst, and F. Müller, "Application of a hybrid neural-fuzzy inference system for mapping crop suitability areas and predicting rice yields," *Environmental Modelling & Software*, vol. 114, pp. 166–180, 2019.

[49] A. Jozi, T. Pinto, I. Praça, F. Silva, B. Teixeira, and Z. Vale, "Energy consumption forecasting based on hybrid neural fuzzy inference system," in *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–5, Athens, Greece, December 2016.

[50] R. Li and J. W. Merchant, "Modeling vulnerability of groundwater to pollution under future scenarios of climate change and biofuels-related land use change: a case study in North Dakota, USA," *Science of the Total Environment*, vol. 447, pp. 32–45, 2013.

[51] USGCRP, *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment*, U.S. Global Change Research Program, Washington, DC, USA, 2016.

[52] N. K. Kasabov and Q. Song, "DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 144–154, 2002.

[53] A. Amirkhani, H. Nasiriyan-Rad, and E. I. Papageorgiou, "A novel fuzzy inference approach: neuro-fuzzy cognitive map," *International Journal of Fuzzy Systems*, vol. 22, no. 3, pp. 859–872, 2020.

[54] N. Kasabov, *Evolving Connectionist Systems: The Knowledge Engineering Approach*, Springer Science \& Business Media, Berlin, Germany, 2007.

[55] M. J. Watts, "A Decade of Kasabov's Evolving Connectionist Systems: A Review," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, vol. 39, 2009.

[56] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[57] H. Tao, A. A. Ewees, A. O. Al-Sulttani et al., "Global solar radiation prediction over North Dakota using air temperature: development of novel hybrid intelligence model," *Energy Reports*, vol. 7, pp. 136–157, 2021.

[58] N. K. Kasabov, "Evolving connectionist systems for adaptive learning and knowledge discovery: trends and directions," *Knowledge-Based Systems*, vol. 80, pp. 24–33, 2015.

[59] E. Lughofer, "EFS approaches for regression and classification," in *Evolving Fuzzy Systems--Methodologies, Advanced Concepts and Applications*pp. 94–164, Berlin, Heidelberg, 2011.

[60] S. Kar, S. Das, and P. K. Ghosh, "Applications of neuro fuzzy systems: a brief review and future outline," *Applied Soft Computing*, vol. 15, pp. 243–259, 2014.

[61] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. Syst. Man. Cybern.*vol. 23, no. 3, pp. 665–685, 1993.

[62] O. E. Omeje, H. S. Maccido, Y. A. Badamasi, and S. I. Abba, "Performance of hybrid neuro-fuzzy model for solar radiation simulation at abuja, Nigeria: a correlation based input selection technique," *Knowledge-Based Eng. Sci.*vol. 2, no. 3, pp. 54–66, 2021.

[63] M. Ko, A. Tiwari, and J. Mehnen, "A review of soft computing applications in supply chain management," *Applied Soft Computing*, vol. 10, no. 3, pp. 661–674, 2010.

[64] A. Mohiyuddin, A. R. Javed, C. Chakraborty, M. Rizwan, M. Shabbir, and J. Nebhen, "Secure cloud storage for medical IoT data using adaptive neuro-fuzzy inference system," *International Journal of Fuzzy Systems*, vol. 24, no. 2, pp. 1203–1215, 2021.

[65] R. Tur and S. Yontem, "A comparison of soft computing methods for the prediction of wave height parameters," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 1, pp. 31–46, 2021.

[66] A. F. GüNeri, T. Ertay, and A. YüCel, "An approach based on ANFIS input selection and modeling for supplier selection problem," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14907–14917, 2011.

[67] P. Aengchuan and B. Phruksaphanrat, "Comparison of fuzzy inference system (FIS), FIS with artificial neural networks (FIS+ ANN) and FIS with adaptive neuro-fuzzy inference system (FIS+ ANFIS) for inventory control," *Journal of Intelligent Manufacturing*, vol. 29, no. 4, pp. 905–923, 2018.

[68] J. Kim and N. Kasabov, "HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems," *Neural Networks*, vol. 12, no. 9, pp. 1301–1319, 1999.

[69] W. Pedrycz and F. Gomide, *Fuzzy Systems Engineering: Toward Human-Centric Computing*, John Wiley & Sons, Hoboken, NJ, USA, 2007.

[70] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. Syst. Man. Cybern.*vol. 22, no. 6, pp. 1414–1427, 1992.

[71] K. Rudd, G. D. Muro, and S. Ferrari, "A constrained backpropagation approach for the adaptive solution of partial differential equations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 571–584, 2014.

[72] S. K. Bhagat, T. Tiyasha, S. M. Awadh, T. M. Tung, A. H. Jawad, and Z. M. Yaseen, "Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models," *Environmental Pollution*, vol. 268, Article ID 115663, 2021.

[73] S. K. Bhagat, T. M. Tung, and Z. M. Yaseen, "Heavy metal contamination prediction using ensemble model: case study of Bay sedimentation, Australia," *Journal of Hazardous Materials*, vol. 403, Article ID 123492, 2021.

[74] C. J. Willmott, "On the validation of models," *Physical Geography*, vol. 2, no. 2, pp. 184–194, 1981.

[75] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," *International Journal of Forecasting*, vol. 15, no. 4, pp. 405–408, 1999.

[76] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: empirical comparisons," *International Journal of Forecasting*, vol. 8, no. 1, pp. 69–80, 1992.

[77] B. S. Karthika and P. C. Deka, "Prediction of air temperature by hybridized model (Wavelet-ANFIS) using wavelet decomposed data," *Aquat. Procedia*, vol. 4, pp. 1155–1161, 2015.

[78] A. Azad, H. Kashi, S. Farzin et al., "Novel approaches for air temperature prediction: a comparison of four hybrid evolutionary fuzzy models," *Meteorological Applications*, vol. 27, no. 1, 2019.

[79] Q. Xie, J.-Q. Ni, J. Bao, and Z. Su, "A thermal environmental model for indoor air temperature prediction and energy consumption in pig building," *Building and Environment*, vol. 161, Article ID 106238, 2019.

[80] X. Zhang, Y. Wang, X. He et al., "Prediction of vehicle driver's facial air temperature with SVR, ANN, and GRU," *IEEE Access*, vol. 10, pp. 20212–20222, 2022.

[81] S. K. Chaudhary, P. K. Srivastava, D. K. Gupta et al., "Machine learning algorithms for soil moisture estimation using Sentinel-1: model development and implementation," *Advances in Space Research*, vol. 69, no. 4, pp. 1799–1812, 2022.

[82] D. O. Rosenberry, D. I. Stannard, T. C. Winter, and M. L. Martinez, "Comparison of 13 equations for determining evapotranspiration from a prairie wetland, Cottonwood Lake area, North Dakota, USA," *Wetlands*, vol. 24, no. 3, pp. 483–497, 2004.

[83] W. L. Schmidt, W. D. Gosnold, and J. W. Enz, "A decade of air–ground temperature exchange from Fargo, North Dakota," *Global and Planetary Change*, vol. 29, no. 3–4, pp. 311–325, 2001.

[84] A. Grundstein, P. Todhunter, and T. Mote, "Snowpack control over the thermal offset of air and soil temperatures in eastern North Dakota," *Geophysical Research Letters*, vol. 32, no. 8, Article ID L08503, 2005.

[85] R. T. Carson, Y. Jeon, and D. R. McCubbin, "The relationship between air pollution emissions and income: US data," *Environment and Development Economics*, vol. 2, no. 4, pp. 433–450, 1997.

Hindawi
Complexity
Volume 2022, Article ID 8177307, 10 pages
https://doi.org/10.1155/2022/8177307

Research Article

Multifractal Early Warning Signals about Sudden Changes in the Stock Exchange States

Andrey Dmitriev,[1,2] **Andrey Lebedev**,[1] **Vasily Kornilov**,[1] **and Victor Dmitriev**[1]

[1]*National Research University Higher School of Economics, Moscow, Russia*
[2]*University of Bernardo O'Higgins, Santiago, Chile*

Correspondence should be addressed to Andrey Dmitriev; a.dmitriev@hse.ru

Received 16 August 2022; Accepted 7 September 2022; Published 22 September 2022

Academic Editor: Andrea Murari

Critical phenomena in stock exchange are regularly occurring and difficult to predict events, often leading to disastrous consequences. The presented paper is devoted to the search and research of early warning signals of critical transitions in stock exchange based on the results of a multifractal analysis of a series of transactions in shares of public companies. We have proposed and justified the use of certain features of behavior of multifractal spectrum shape parameters such as signals. As model time series, on which methods of multifractal analysis were tested, we used a series of the number of unstable sites of the sandpile automaton on the random Erdős–Rényi graph, self-organizing into critical and bistable states. It was found that the early warning signals for both cellular automata and stock exchanges are an increase in the magnitude of the maximum position, a decrease in the width, and a decrease, followed by a sharp increase, in the value of the spectrum asymmetry parameter.

1. Introduction

Most complex systems, regardless of their origin, are scale-invariant, have heterogeneity and nonstationary behavior, and contain internal mechanisms of self-organization. Therefore, dynamic processes in such systems are usually nonlinear. In such systems, abrupt changes in states can occur; such changes are often called critical transitions. An example is a phase transition accompanied by a radical change in the properties of the system at the macro level. As a result of a phase transition, the system acquires completely new and unexpected properties that are not reducible to the properties of individual parts.

Ordinary critical phenomena, such as phase transitions of the second kind, are observed only when the control parameter reaches a certain critical value. In other words, a critical state is created artificially by tuning a control parameter to a critical value. For example, if a control parameter such as temperature is adjusted to a critical value, then an order parameter such as magnetization will reach a zero value, and the paramagnetic-ferromagnetic phase transition will occur in the system. The parameters of the system at the critical point are characterized by power laws.

For most complex macroscopic systems and processes of natural origin, it is impossible to adjust the value of a control parameter to a critical value, but, despite this, such systems, while in a critical state, are characterized by power laws. Examples of such systems and processes are financial markets with crashes and crises, seismic activity with catastrophic earthquakes, social networks with information cascading, and other systems (e.g., see papers [1–4]). The answer to the question of how critical transitions occur in such systems was given by Per Bak, Chao Tang, and Kurt Wiesenfeld only in the late 1980s. They discovered the phenomenon of self-organized criticality (SOC) and proposed a theory that explains how such systems reach a critical state without tuning the control parameter (e.g., see papers [5, 6]). It turned out that a critical state, in which even a minor event can lead to a catastrophe, can not only be created artificially (e.g., in laboratory conditions), but also

arise as a result of the self-organization of the system. In such a state, the system acquires properties that its elements did not have, demonstrating complex holistic behavior.

The basic model of the SOC theory is a sandpile into which grains of sand fall from time to time (e.g., see papers [2, 7]). At first, the pile simply grows, and in those places where the local slope is greater than the stability threshold, sand grains crumble down the slope to neighboring surface areas. If the average surface slope (z) is small, the set of chaotically directed microcurrents of sand grains is mutually balanced and the macroscopic sand current $J = 0$. If z exceeds some critical value (z_c), then there is a spontaneous sand flow ($J \neq 0$) across the surface of the heap, which increases as z increases. The value of z_c separates the subcritical ($z < z_c$) and supercritical ($z > z_c$) phases, which are resistant to small perturbations. If $z = z_c$, then a single fallen grain of sand can cause avalanches of any size. Thus, the sandpile self-organizes into a critical state at $J = 0^+$, corresponding to a phase transition of the second kind with a control parameter z and a parameter of order J. It should be noted that the sandpile model also allows us to explain the self-organization in the bistable state, corresponding to a phase transition of the first kind. For this purpose, a model of facilitated sandpile was proposed (see the paper [8]), which can demonstrate self-organized bistability (SOB) (e.g., see papers [9, 10]).

There are many studies that substantiate the concept of the similarity of the mechanisms of behavior of economic systems, stock markets, and financial time series with the behavior of the model variables (e.g., see papers [11–17]), as well as studies on the search for early warning signals (EWS) for critical transitions in financial and stock markets (e.g., see papers [18–24]). The determination of the time interval preceding the occurrence of a critical transition in the system not only has important theoretical value, but also has important applied value. The studies we know are mainly focused on the detection of SOC mechanisms in financial systems. Also, there are studies devoted to finding EWS for financial crises using mainly measures of correlation theory (autocorrelation function, skewness, kurtosis, variance, and other measures) according to the results of the analysis of financial series in the selected range. A significant limitation in using such measures for the study of financial series scaling is their applicability only to stationary and fractal time series (e.g., see the paper [25]).

At the moment, we are not aware of any works that present studies of the dynamics of stock exchange self-organization in SOC and SOB states, based on the results of multifractal analysis of the time series of the number of deals made on the shares of companies (volume indicator), as well as multifractal EWS for the corresponding critical transitions. To address this gap, we investigated the possibilities and limitations of multifractal EWS for critical transitions using the results of multifractal stochastic dynamics analysis of volume indicators, using the numbers of unstable cells of the sandpile cellular automaton as the basic (in a sense reference) time series. Research results are presented in this paper.

The paper is structured as follows: Section 2 is devoted to the description of mechanisms of functioning of sandpile cellular automaton and substantiation of similarities in the behavior of such automata and stock markets. Methods for generating time series of the number of unstable automaton nodes and methods for obtaining time series of the number of deals made on shares of companies are also presented. The rationale for the necessity of application and methods of calculation of parameters of a multifractal spectrum of time series as measures of early detection of critical transitions is presented; Section 3 presents and discusses the results of calculations of parameters of multifractal spectra of time series used as measures of early detection of critical transitions; and Section 4 presents the main conclusions, possible practical applications of the obtained results, and the prospects for further research.

2. Data Set and Methods

2.1. Model and Real-Time Series. The self-organized critical sandpile behavior considered in Section 1 can be described using sandpile cellular automata (e.g., see papers [7, 8, 26]). Among the many sandpile cellular automata models, we chose the Manna model (see the paper [27]) on the Erdős–Rényi random graph (e.g., see papers [28, 29]) as the most relevant model of avalanche-like changes in the number of traded shares of companies.

We built three Erdős–Rényi random graphs with a number of sites N equal to 500, 1500, and 2500 by connecting any two sites v_i and v_j with edge e_{ij} with probability p independently of all other pairs of sites.

In this case, the Manna model is a random graph with the number of sites N, the sites of which are assigned integer non-negative numbers $z_i(v_j)$. These numbers are traditionally interpreted as the number of sand grains. If $z_i(v_j)$ is not less than the set threshold z_{cv_j}, then site v_j is unstable and "topples." This removes z_c sand grains from it, each of which is transferred to one of the randomly chosen neighboring sites. If a site is on the edge of the graph, the sand grains transferred for it are irreversibly lost. Each neighboring site receives a random number of sand grains δ_k. If there are several unstable sites, they "topple" simultaneously—during one time step.

The elementary event that causes the system to move from one steady state to another is initiated by adding a grain of sand to one of the sites. If the addition of a grain of sand causes a site to lose stability, then the grains of sand transferred to neighboring sites during its toppling may violate their stability. The chain reaction of toppling that continues as long as unstable sites remain in the system will be called an avalanche.

Regardless of the initial state of the system, after a certain number of events, the system reaches a critical state (SOC state), in which the processes occurring are scale-invariant, and all characteristics of avalanches correspond to power distributions.

The rules of the standard Manna model on the Erdős–Rényi random graph with the number of sites N, which demonstrates the output of the system in the SOC state, corresponding to the phase transition of the second kind, have the following form:

$$
\begin{aligned}
&z_i\left(v_j\right) \geq z_{cv_j} > 1,\\
&z_{i+1}\left(v_j\right) \longrightarrow z_{i+1}\left(v_j\right) - z_{cv_j},\\
&z_{i+1}\left(\mathrm{Ne}\right) \longrightarrow z_{i+1}\left(\mathrm{Ne}\right) + \delta_k, \sum_{k=1}^{z_{cv_j}} \delta_k = z_{cv_j}, \quad \delta_k \geq 0,
\end{aligned}
\tag{1}
$$

where Ne denotes the nearest neighboring site to the site v_j.

As noted in Section 1, sandpile cellular automata are also capable of self-organization into a bistable state (SOB state) corresponding to a first-order phase transition. The rules of the automaton allowing to bring it into SOB state are known as "facilitated rules" (see papers [8, 26]). A site v_j of the facilitated automaton is unstable when $z_i(v_j) \geq z_{cv_j}$ and when $f_{i-1}(v_j) \geq 2$ (f_{i-1} is the number of hits to site v_j at the previous iteration). This is the main difference between the facilitated and the standard automaton. Thus, the rules of the facilitated Manna model have the following form:

$$
\begin{aligned}
&z_i\left(v_j\right) \geq z_{cv_j} > 1 \vee f_i\left(v_j\right) \geq 2,\\
&z_i\left(v_j\right) \geq z_{cv_j}: \begin{cases} z_{i+1}\left(v_j\right) \longrightarrow z_{i+1}\left(v_j\right) - z_{cv_j}, \\ z_{i+1}\left(\mathrm{Ne}\right) \longrightarrow z_{i+1}\left(\mathrm{Ne}\right) + \delta_k, \sum_{k=1}^{z_{cv_j}} \delta_k = z_c, \quad \delta_k \geq 0, \\ f_{i+1}\left(\mathrm{Ne}\right) \longrightarrow f_{i+1}\left(\mathrm{Ne}\right) + \delta_k, \quad \delta_k > 0, \end{cases}\\
&z_i\left(v_j\right) < z_{cv_j}: \begin{cases} z_{i+1}\left(x, y\right) \longrightarrow z_{i+1}\left(x, y\right) - z_i\left(x, y\right), \\ z_{i+1}\left(\mathrm{Ne}\right) \longrightarrow z_{i+1}\left(\mathrm{Ne}\right) + \delta_k, \sum_{k=1}^{z_{cv_j}} \delta_k = z_i\left(v_j\right), \quad \delta_k \geq 0, \\ f_{i+1}\left(\mathrm{Ne}\right) \longrightarrow f_{i+1}\left(\mathrm{Ne}\right) + \delta_k, \quad \delta_k > 0. \end{cases}
\end{aligned}
\tag{2}
$$

The avalanche-like propagation of sand grains between sites of the considered sandpile cellular automata in the critical state is a good qualitative econophysical model demonstrating the general regularities of the origin of the avalanche-like change in the number of deals made on the stocks of companies. Indeed, the nodes of the graph can be associated with the agents of the stock market; the edges of the graph, along which the movement of sand grains from unstable sites occurs, can be associated with the deals made between agents; and the random addition of sand grains to the sites can be associated with the market pumping (e.g., information pumping from media, quarterly reports, news feeds, and others). Then, the change in time of the number of unstable sites on the graph corresponds to the change in the number of deals made on the shares of companies. Therefore, we further use the time series of the number of unstable sites with known critical transition times as test series to determine the capabilities and limitations of a particular method of multifractal analysis in the selection and evaluation of EWS for the critical transitions.

We believe it is important to note that besides sandpile cellular automata with Manna model rules, there are other models of self-organized critical cellular automata that cannot be adequate models of stock market transactions. For example, the Bak–Tang–Wiesenfeld model and the Feder–Feder model assume that a nonrandom equal number of sand grains are transferred from an unstable site; the Dhar–Ramaswamy model (e.g., see the paper [30]) and the Pastor-Satorras–Vespignani model (e.g., see the paper [31]) are directional models in which the unstable site has only underlying neighboring sites. In addition, all models can also be realized on square lattices, which implies that there are only four nearest neighbors with an unstable site. Another well-known model that demonstrates self-organized critical behavior is the forest-fire model (e.g., see papers [32, 33]). This model is one of the most popular for simulating sociopolitical and historical processes since it simulates the spread of arousal in some environments.

Time series volume indicators were selected for companies whose shares are listed on any of the stock exchanges. In the stock trading volume data, information is available for 1-day intervals. The exchanges where these companies are traded represent four regions: Asia (Sony Group Corporation, Subaru Corporation); Russia (PJSC Aeroflot—Russian Airlines, Sberbank of Russia); the USA (Apple Inc., Meta Platforms, Inc., and Tesla, Inc.); and Europe (Airbus SE, Allianz SE, Deutsche Lufthansa AG).

2.2. Multifractal Analysis of the Time Series. It is now generally accepted that many financial time series have a complex fractal structure (e.g., see papers [34–37]). In particular, fractal analysis is effectively used to predict market crashes in financial series (e.g., see papers [38–40]). In addition, the universality of multifractal analysis has determined the success of its application to the analysis of time series depicting the dynamics of critical transitions (e.g., see papers [26, 41]).

The features of time series scaling can be studied using different approaches, starting with the classical correlation (or spectral) analysis. Among the obvious drawbacks of such approaches is their applicability only to stationary time

series. Since most processes in nature are highly heterogeneous and nonstationary, the attractiveness of the choice of one or another method of analysis is largely determined by its universality and the possibility of its effective application to real processes of any origin.

The most popular methods for analyzing the multifractal structure of nonstationary time series are multifractal detrended fluctuation analysis (MF-DFA) (e.g., see papers [42, 43]); wavelet transform modulus maxima (WTMM), based on continuous wavelet transform (e.g., see papers [44, 45]); and wavelet leaders (WL), based on discrete wavelet transform (e.g., see the paper [46]).

The MF-DFA is a variant of variance analysis of univariate random walks. The method algorithm analyzes the root mean square error of linear approximation ($F^2(s)$) of the generalized random walk model from the size (s) of the approximated area. The analyzed time series is multifractal if the scaling relation is observed for all s:

$$F_q(s) = \left\{ \frac{1}{N_s} \sum_{i=1}^{N_s} \left[F_i^2(s) \right]^{q/2} \right\}^{1/q} \sim s^{H_q}, \tag{3}$$

where N_s is the number of approximated sections, and H_q are generalized Hurst exponents if $q \in (-\infty, +\infty)$.

The multifractal spectrum ($D(H)$) has the following form (see the paper [41]):

$$D_q = \frac{qH_q - 1}{q - 1}, \tag{4}$$

where D_q are the generalized multifractal dimensions.

The WTMM method assumes the existence of the following scaling relation for multifractal time series:

$$Z(q, s) = \sum_{l \in L(s)} \left(\sup_{s' \le s} |W(s', t_l(s'))| \right)^q \sim s^{\tau_q}. \tag{5}$$

In equation (5), $Z(q, s)$ is the structural function; $L(s)$ is the set of all lines l of maximum modules of wavelet coefficients existing at scale s; $t_l(s')$ characterizes the location of the maximum at scale s, relating to line l; $W(\cdot)$ are the coefficients of the continuous wavelet transform; and τ_q are the scaling exponents.

The multifractal spectrum ($D(h)$) has the following form (see the paper [44]):

$$D_q = qh_q - \tau_q, \tag{6}$$

where D_q are the generalized multifractal dimensions, and h_q are the Hölder exponents.

The WL method assumes the existence of the following scaling relation for a multifractal time series:

$$Z(q, s) = \frac{1}{n_s} \sum_{k=1}^{n_s} L(s, k)^q \sim s^{\tau_q}. \tag{7}$$

In equation (6), $L(k, s) = \sup_{\lambda' \subset 3\lambda_{s,k}} |d(k, s)|$ are the leaders of the wavelet coefficients in which the 2^s scales are translated into the $2^s k$ time positions; $d(k, s)$ are the coefficients of the discrete wavelet transform; k is the time shift; and s is the scale.

The multifractal spectrum ($D(h)$) is defined by the following decomposition:

$$D(h) = d + \frac{c_2}{2!}\left(\frac{h - c_1}{c_2}\right)^2 + \frac{-c_3}{3!}\left(\frac{h - c_1}{c_2}\right)^3 + \cdots, \tag{8}$$

where c_1, c_2, and c_3 are the log-cumulants. c_1 corresponds to the position of the spectrum maximum, c_2 characterizes the width of the spectrum, and c_3 characterizes the asymmetry of the spectrum. The triplet c_1, c_2, c_3 contains the basic information about the multifractal structure of the studied time series.

As will be shown in Section 3, the studied time series are multifractal series, which require an infinite spectrum of fractal dimensions for a complete description. Therefore, as an EWS for critical transitions in the sandpile cellular automata and stock markets, we use the features of changes in the multifractal spectra ($D(H)$ and $D(h)$) of the studied time series as the systems approach critical points.

We used three main spectrum shape parameters as early warning measures for critical transitions $(\cdot)$: the position of the spectrum maximum (H_0, h_0 and c_1); spectrum width ($W = H_{\max} - H_{\min}$, $W = h_{\max} - h_{\min}$ and c_2); and spectrum asymmetry ($S = H_{\max} - H_0/H_0 - H_{\min}$, $S = h_{\max} - \overline{h_0/h_0} - h_{\min}$ and c_3). The spectrum was calculated at $q = \overline{-5, 5}$ in increments of 0.1.

We calculated the time series of early warning measures (m_t) with a fixed left window boundary corresponding to the first value (x_1) of the studied time series $x_t, t = \overline{1, n}$, and a sliding right window boundary (τ) corresponding to some selected value (x_τ) of the studied time series. As a result, we obtained series of early warning measures $m_t, t = \overline{\tau, n}$, with $\tau = 1000$ for the time series of the number of toppled cells of the sandpile cellular automata and $\tau = 50$ for the time series of the volume indicators.

3. Results and Their Discussion

3.1. Time Series of Unstable Sites. The time series of the number of unstable sites of automata with standard rules, which lead to the output of the automaton in SOC state, and with clothed rules, which lead to the output of the automaton in SOB state, of the Manna model are shown in Figure 1. The figure shows time series of automata whose random graphs contain $N = 2500$ sites. Time series for $N = 500$ и $N = 1500$ have similar appearance. The series differ only in the time it takes for the system to enter the critical state (subcritical time) and in the maximum values of the number of unstable sites of cellular automata in the critical state. These values are presented in Table 1.

The time series demonstrate the presence of subcritical phase (SubC phase) and critical state (SOC state and SOB state) of the sandpile cellular automata (see Figure 1). The SubC phase corresponds to the noncatastrophic behavior. The sandpile cellular automaton, being in this chaotic phase, is stable to small perturbations. Only at the critical point (SOC state), catastrophes are possible (see Figure 1(a)) since

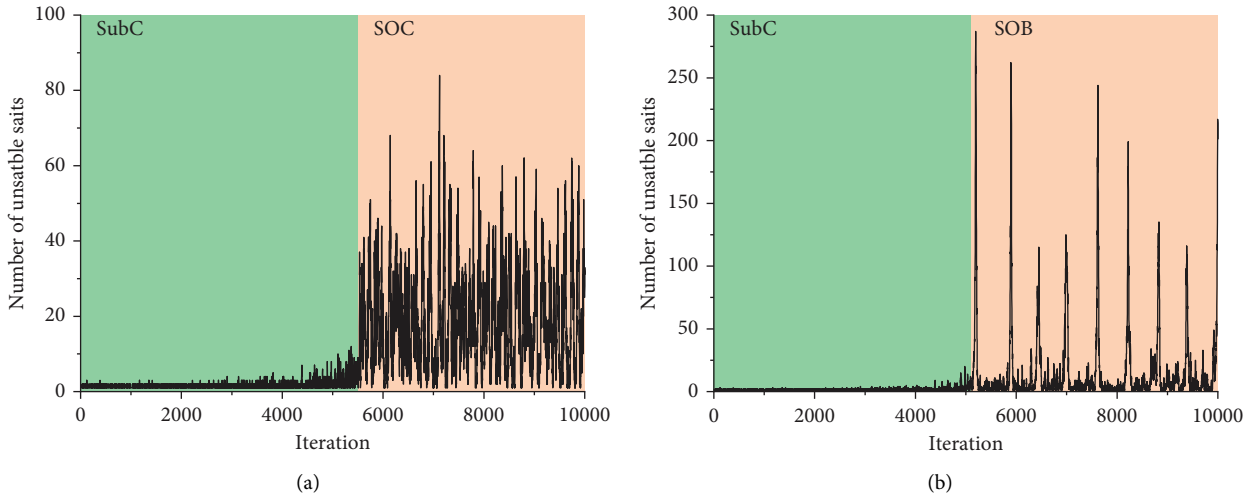


Figure 1: Time series of the number of unstable sites of sand-cell automata. (a) Standard rules of the Manna model and (b) facilitated rules of the Manna model.

Table 1: Subcritical time (t_{SubC}) and the maximum number of unstable sites (A_{max}) of cellular automata in the critical state.

Number of sites	Standard model		Facilitated model	
	t_{SubC}	A_{max}	t_{SubC}	A_{max}
500	2750	38	4320	196
1500	5050	62	4630	238
2500	5500	80	5100	280

a single added grain of sand in any site of the automaton can cause an avalanche of grains of any size. The SOB state is also characterized by an avalanche of sand grains of any size with the appearance of periodic bursts of activity (see Figure 1(b)). So, Buendia and coauthors state in their paper [9] that "the probability distributions for both avalanche size and duration are bimodal": small avalanches coexist with extremely large ones that span the whole system. These latter "anomalous" outbursts of activity, which are also called "king" avalanches, occur in an almost periodic way. The size of avalanches (the maximum number of unstable sites) increases with the total number of sites of the sandpile cellular automata (see Table 1).

An important characteristic of the process of the system reaching a critical state is the subcritical time (t_{SubC}). It is known (e.g., see papers [2, 5]) that different types of SOC systems have different t_{SubC}. The greatest t_{SubC} is characteristic of the evolution of the Earth's crust and of biological evolution. For many other types of SOC systems, the t_{SubC} is much smaller.

In our opinion, the reason for such large differences in t_{SubC} values is the different levels of complexity of individual SOC systems. Differences in the value of t_{SubC} allow to distinguish different levels of complexity in SOC systems. This extends the applicability of SOC theory, as well as SOB, far beyond the characteristic power laws for the distribution of avalanche size and power spectral density as $1/f$ noise. Given that t_{SubC} increases with the size of the cellular automata (see Table 1), we can use t_{SubC} as a measure of the complexity of the system capable of a critical transition. Note that we previously found a similar change in t_{SubC} with changes in the size of the cellular automata (see the paper [26]). Besides, other things being equal, the value of t_{SubC} for SOB systems is lower than the value of t_{SubC}, characteristic of SOC systems.

Perhaps the formation of a more complex SOC system initially requires a larger value of t_{SubC}, but when such a system is already formed, the corresponding t_{SubC} at the next level is already much smaller.

3.2. Multifractal Measures for Early Detection of Critical Transitions in Sandpile Cellular Automata. The scaling relation (3) of the MF-DFA method is not met for any values of the right boundary x_τ of the sliding window. Therefore, this method cannot be used as a method for calculating measures of early detection of critical transitions in sandpile cellular automata based on the results of multifractal analysis of the number of unstable sites series. The reason why the MF-DFA method does not allow revealing the multifractal structure of model time series is the presence of a large number of repeating values in such series.

In contrast, the scaling relations (see equations (5) and (7)) of wavelet transform-based methods are satisfied for all $\tau \in [1000, 10000]$. This is connected with the fact that these methods do not require the extraction of local trends in repeating values of the time series. The time series of multifractal early warning measures of critical transitions in sandpile cellular automata with a number of 2500 tiles obtained by the WTMM method are presented in Figure 2,

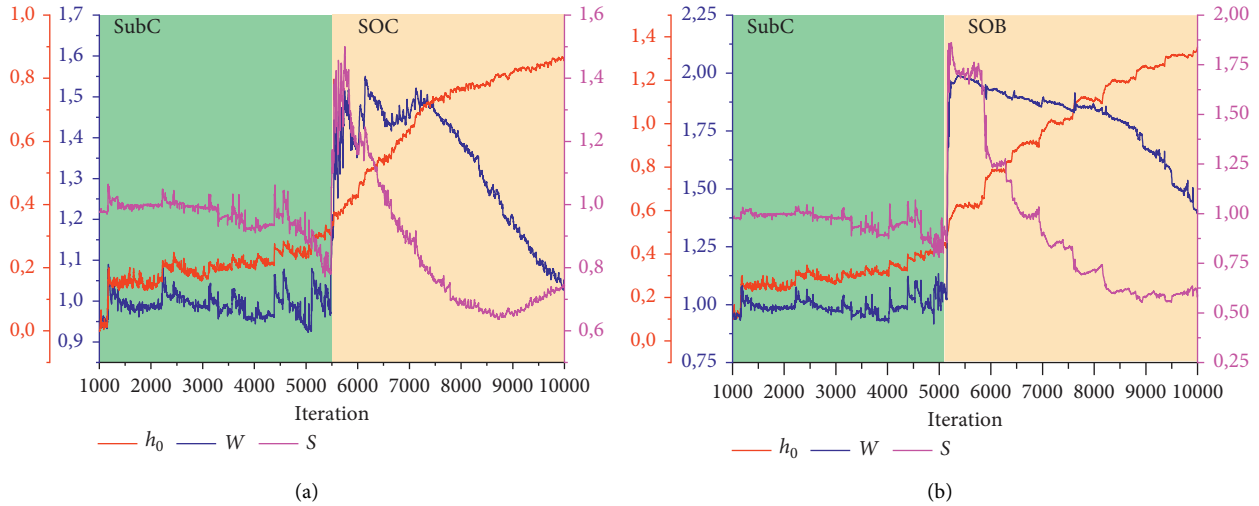


FIGURE 2: Time series of the maximum position (h_0), width (W), and asymmetry (S) of the multifractal spectrum for the sandpile cellular automata. (a) Standard rules of the Manna model and (b) lightweight rules of the Manna model.

and those obtained by the WL method are presented in Figure 3. The conditional time corresponding to the iteration step is used.

The values h_0 (see Figure 2) and c_1 (see Figure 3), which characterize the positions of the maximum of the multifractal spectrum of unstable sandpile cellular automata, increase as the automata approach the critical state. This increase is typical for both standard cellular automata and facilitated cellular automata. Consequently, as the automaton approaches the critical state, the time series of the number of its unstable tiles becomes more "smooth" or less "jagged." Note that a sharp increase in the position of the maximum of the singularity spectrum is also observed in the vicinity of the critical point of the phase transition of the second kind in the Ising model (e.g., see the paper [41]).

The width of the multifractal spectrum (W) of the time series of unstable sites computed by the WTMM method decreases as the standard and facilitated automata approach the critical state (see Figure 2). Also, the value of W calculated by the WL method decreases or is equivalent to the absolute value of the second log-cumulant $|c_2|$ (see Figure 3). The equivalence of W and $|c_2|$ follows from a simple analysis of equation (8). The increasing W value is observed in the SOC state and SOB state. It follows from the decreasing value of W that as automata approach the critical state, the time series of unstable sites become more homogeneous fractal series, with a more uniform distribution of series values. A similar decrease in the width of the spectrum in the vicinity of the critical point is characteristic of a second-order phase transition in the Ising model (e.g., see papers [41, 47]).

The value of the spectrum asymmetry parameter S calculated by the WTMM method first decreases, then sharply increases as the automata approach the critical state (see Figure 2). Consequently, large fluctuations (strong singularities) in the number of unstable tiles of automata as they approach the critical state prevail in the time series. Similar behavior of the parameter S is also observed in the Ising model (e.g., see the paper [41]). The log-cumulant c_3, which characterizes the asymmetry of the spectrum, decreases not only in the vicinity of the time of the critical transition, but also in some noncritical time interval (see Figure 3). In our opinion, such behavior of parameter c_3 contradicts the existence of one of the precursors of the critical transition, known as the critical slowing down (e.g., see papers [19, 21, 24, 48]). Perhaps the incorrect estimation of the asymmetry parameter is one of the drawbacks of the expansion (8) or, moreover, a drawback of the WL method.

The critical slowing down is the phenomenon that when a system approaches a critical point, it relaxes more slowly after small perturbations. It is known (e.g., see papers [48, 49]) that time series showing a critical slowing down are characterized by increases in autocorrelation (or increases in h_0 and decreases in W, as we found), dispersion (or increases in S, as we found), kurtosis and skewness, and the β of the power spectral density $1/f^{\beta}$ (or increases in h_0, as we found). Consequently, of the three multifractal analysis methods, only the WTMM method allows us to obtain correct estimates of the multifractal spectrum shape parameters, at least for the time series of the number of unstable tiles of cellular automata. Recall that the time-varying WL estimation (c_3) for the asymmetry parameter does not explain the critical slowdown.

3.3. Multifractal Measures for Early Detection of Critical Transitions in Stock Exchange. As shown in Subsection 3.2, the multifractal early warning signals are an increase in the magnitude of the maximum position (h_0, c_1) of the multifractal spectrum $D(h)$, a decrease in the spectrum width (W, $|c_2|$), and a decrease followed by a sharp increase in the spectrum asymmetry parameter (S). We also remind that the MF-DFA method did not reveal a multifractal structure in the time series of the number of unstable tiles.

In this subsection, we demonstrate the results of calculations of these three shape parameters of $D(h)$ for the

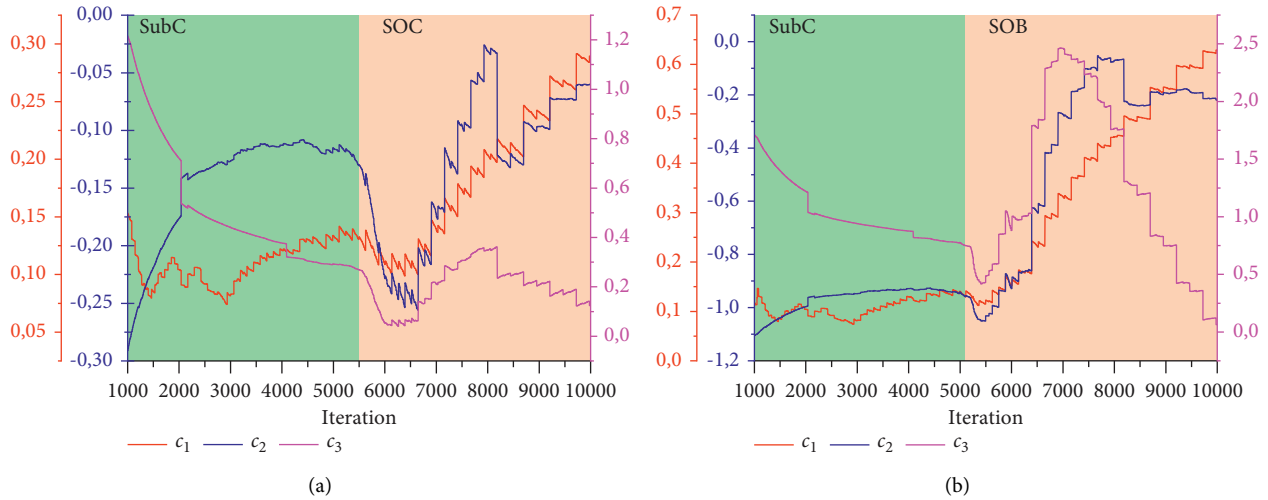


FIGURE 3: Time series of the log-cumulants ($c_i, i = 1, 2, 3$) for the sandpile cellular automata. (a) Standard rules of the Manna model and (b) facilitated rules of the Manna model.

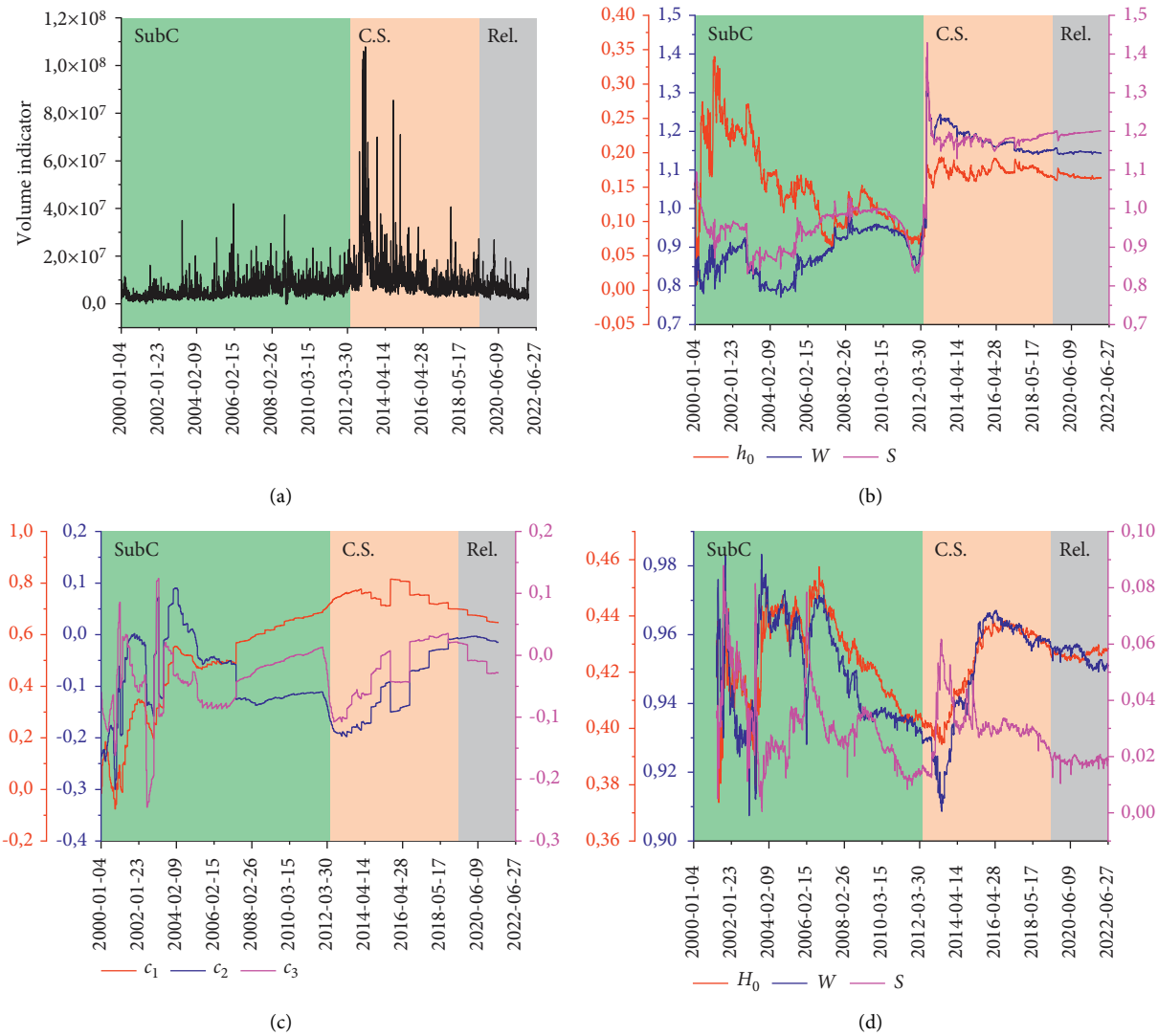


FIGURE 4: Financial time series and time series of multifractal spectrum shape parameters for Sony Group Corporation. The symbol "C.S." denotes the critical state, and the symbol "Rel." denotes the relaxation. (a) Volume indicator, (b) WTMM method, (c) WL method, and (d) MF-DFA method.

Table 2: Multifractal early warning signals and the critical transition date (t_C) of the stock exchange volumes.

Public company	WTMM method			WL method			MF-DFA method			t_C
	h_0	W	S	c_1	c_2	c_3	H_0	W	S	
Sony Group Corporation	+	+	+	+	+	+	–	+	+	May 25, 2012
Subaru Corporation	+	+	+	+	+	+	–	+	+	February 20, 2016
Apple Inc.	+	+	+	+	+	+	–	+	–	September 20, 2014
PJSC Aeroflot—Russian Airlines	+	+	+	+	+	+	–	+	+	March 17, 2002
Airbus SE	+	+	+	+	+	+	+	–	+	July 25, 2007
Allianz SE										April 19, 2004
Meta Platforms, Inc.	+	+	+	+	+	+	–	+	+	December 3, 2021
Deutsche Lufthansa AG	+	+	+	+	+	+	–	+	+	August 11, 2014
Sberbank of Russia	+	+	+	+	+	+	–	+	+	November 25, 2021
Tesla Inc.	+	+	+	+	+	+	–	+	–	January 8, 2020

volume indicator series, using MF-DFA, WTMM, and WL methods. As an illustrative example, Figure 4 shows the time series of the spectrum shape parameters calculated for the time series of the number of transactions on Sony Group Corporation shares. These figures also show the subcritical phase and the subcritical time (t_c), and the possible critical state and relaxation intervals of the segment of the stock exchange whose agents are involved in Sony Group Corporation stock transactions. Hereinafter, the term "stock exchange segment" refers to a stock exchange with agents involved in transactions in the stock of a particular public company, such as Sony Group Corporation.

As shown by the results of the WTMM calculation of changes in the shape parameters (see Figure 4(b)), the corresponding market segment on May 25, 2012, self-organizes into a critical state. Accordingly, the time it takes for a market segment to enter the critical state (subcritical time) is 3035 days.

The volume indicator series shown in this figure demonstrates another interesting phenomenon, which in our opinion cannot be determined by the multifractal early warning signals. This is a market segment relaxation that begins around March 2, 2019 (see Figure 4(a)). In this time interval, the number of transactions on shares decreases, which is probably due to the loss of interest of market players in the corresponding transactions.

The results of the WL method are similar to those of the WTMM method even for the spectrum asymmetry parameter, although the increase in c_3 occurs at the moment of critical transition (see Figure 4(c)). Recall that for time series of the number of unstable tiles of automata, the behavior of $S(t)$ and $c_3(t)$ when approaching the critical point in time is not consistent (see Figures 2 and 2b).

The MF-DFA method made it possible to reveal the multifractal structure in the volume indicator series and, consequently, to give estimates of the spectrum shape parameters (see Figure 4(d)). The behavior of the parameters W and S when approaching the critical point is similarly consistent with the behavior of the same parameters calculated by the WTMM method, and this behavior is characteristic of the critical deceleration. Yet, in spite of this, the parameter H_0 decreases when approaching the critical point. This contradicts with the results of calculations of parameter H_0 by methods based on wavelet transform. Consequently, the critical transition of the exchange segment associated with the trades in Sony Group Corporation shares cannot be detected in advance when calculating the spectrum parameters using the MF-DFA method.

In order not to overload our paper with unnecessary graphical information, for the remaining nine time series studied, we presented the results of calculations of multifractal early warning signals and the most important features of the time series in Table 2. The symbol "+" means that the change in time series values for the corresponding measure when approaching the time of critical transition is similar to the behavior of the time series for sandpile cellular automata. The symbol "–" means that there is no such analogy.

The results presented in Table 2 suggest that all of the considered public companies self-organize into a critical state. At the same time, the time for companies to reach a critical state (t_{SubC}) is different. Methods based on wavelet transform give similar results and allow us to give estimates of the values of the parameters of the shape spectra, acceptable for their application as early warning signals.

4. Conclusion

The time series of the number of unstable tiles of sandpile automata and the time series of the number of transactions in shares of public companies with one-day increments are multifractal. Such series admit decomposition into segments with different local scaling properties, so their quantitative description requires a whole spectrum of fractal dimensions, such as a multifractal spectrum in the form of $D(h)$, sometimes called singularity spectrum.

For early detection of the time moment of the systems reaching a critical state based on the results of analysis of multifractal series generated by such systems, analysis of the change in the shape of the multifractal spectrum as the system approaches the point of critical transition is required. A change in the shape of the spectrum with a good degree of accuracy is determined by a change in its three parameters. These parameters are the position of the spectrum maximum, spectrum width, and spectrum asymmetry.

As the sandpile cellular automata and stock exchange volume approach the time point of critical transition, the value of the maximum position increases, the width

decreases, and the value of the spectrum asymmetry parameter decreases, followed by a sharp increase. Such behavior of the spectrum shape in the vicinity of the critical transition point corresponds to the critical slowdown of the system as it approaches the critical transition point. Indeed, in the vicinity of a critical point, the time series becomes more regular and homogeneous, with larger fluctuations prevailing in the value of the number of unstable tiles and the value of the number of stock transactions. Therefore, the indicated behavior of the values of the spectrum shape parameters calculated by the WTMM method is reliable early warning signals for critical transitions.

The sandpile cellular automaton is a very coarse model of the stock exchange, but, despite this, its time series have similar behavior, demonstrating subcritical phase and critical state, and similar behavior of the values of spectrum shape parameters when approaching the time moment of critical transition. Therefore, the series of the number of unstable tiles can be used as reference series for testing various measures of early detection of critical transitions and the method for their calculation. In our opinion, these analogies of time series are caused by multifractality of random graphs with colored (unstable) tiles and stock exchange transaction network-multifractal structures generate multifractal series.

There are different types of self-organized criticality (in particular, self-organized criticality and self-organized bistability), which differ from each other in the level of complexity depending on the characteristic value of subcritical time. The more complex the system of self-organized criticality, in particular, the more tiles the random graph contains, the greater its subcritical time. Also, the subcritical time is different for different public companies. Perhaps this difference is not only due to the different number of market players involved in transactions with the shares of a particular public company, but also due to the different mechanisms of self-organized criticality. In any case, subcritical time can be viewed as one measure of system complexity, along with power laws for the probability density function and autocorrelation function of avalanche size, as well as $1/f$ noise.

In conclusion, we note that not all segments of the stock exchange are capable of self-organization into a critical state; perhaps for some market segments, the time moment of critical transition has not yet arrived. Yet despite this, the possibility of early detection of critical transitions should not be underestimated. In particular, this is due to the irreversibility of a segment of a stock exchange as it approaches a critical point, which can have catastrophic consequences for a company. Multifractal early warning signals will give company managers information about the need to take precritical measures if there is enough time to take such measures.

Data Availability

The time-series data of stock trading volumes used to support the findings of this study are available at https://finance.yahoo.com

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work was an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). The work was partially supported by the Russian Foundation for Basic Research (Project 20-07-00651 A).

References

[1] B. Tadić and R. Melnik, "Self-organised critical dynamics as a key to fundamental features of complexity in physical, biological, and social networks," *Dynamics*, vol. 1, no. 2, pp. 181–197, 2021.

[2] N. W. Watkins, G. Pruessner, S. C. Chapman, N. B. Crosby, and H. J. Jensen, "25 Years of self-organized criticality: concepts and controversies," *Space Science Reviews*, vol. 198, no. 1-4, pp. 3–44, 2016.

[3] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[4] J. P. Sethna, "Power laws in physics," *Nature Reviews Physics*, vol. 4, no. 8, pp. 501–503, 2022.

[5] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality," *Physical Review A*, vol. 38, no. 1, pp. 364–374, 1988.

[6] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality: an explanation of the 1/f noise," *Physical Review Letters*, vol. 59, no. 4, pp. 381–384, 1987.

[7] A. A. Járai, "The sandpile cellular automaton," in *Probabilistic Cellular Automata. Emergence, Complexity and Computation*, P. Y. Louis and F. Nardi, Eds., vol. 27, pp. 79–88, 2018.

[8] S. di Santo, R. Burioni, A. Vezzani, and M. A. Muñoz, "Self-organized bistability associated with first-order phase transitions," *Physical Review Letters*, vol. 116, no. 24, Article ID 240601, 2016.

[9] V. Buendía, S. di Santo, J. A. Bonachela, and M. A. Muñoz, "Feedback mechanisms for self-organization to the edge of a phase transition," *Frontiers in Physics*, vol. 8, p. 333, 2020.

[10] V. Buendía, S. di Santo, P. Villegas, R. Burioni, and M. A. Muñoz, "Self-organized bistability and its possible relevance for brain dynamics," *Physical Review Research*, vol. 2, no. 1, Article ID 013318, 2020.

[11] M. Bartolozzi, D. B. Leinweber, and A. W. Thomas, "Self-organized criticality and stock market dynamics: an empirical study," *Physica A: Statistical Mechanics and Its Applications*, vol. 350, no. 2-4, pp. 451–465, 2005.

[12] P. Bak, "Catastrophes and self organized criticality," *Computers in Physics*, vol. 5, no. 4, p. 430, 1991.

[13] N. Zachariou, P. Expert, M. Takayasu, and K. Christensen, "Generalised sandpile dynamics on artificial and real-world directed networks," *PLoS ONE*, vol. 10, no. 11, Article ID e0142685, 2015.

[14] B. Rao, D. Yi, and C. Zhao, "Self-organized criticality of individual companies: an empirical study," in *Procedings of the Third International Conference on Natural Computation*, pp. 481–487, Haikou, China, August 2007.

[15] M. Bartolozzi, D. Leinweber, and A. Thomas, "Scale-free avalanche dynamics in the stock market," *Physica A:*

Statistical Mechanics and Its Applications, vol. 370, no. 1, pp. 132–139, 2006.

[16] C. Tebaldi, "Self-organized criticality in economic fluctuations: the age of maturity," *Frontiers in Physics*, vol. 8, Article ID 616408, 2021.

[17] A. E. Biondo, A. Pluchino, and A. Rapisarda, "Modeling financial markets by self-organized criticality," *Physical Review E*, vol. 92, no. 4, Article ID 042814, 2015.

[18] J. Jurczyk, T. Rehberg, A. Eckrot, and I. Morgenstern, "Measuring critical transitions in financial markets," *Scientific Reports*, vol. 7, no. 1, Article ID 11564, 2017.

[19] M. S. Ismail, M. S. Md Noorani, M. Ismail, and F. Abdul Razak, "Early warning signals of financial crises using persistent homology and critical slowing down: evidence from different correlation tests," *Frontiers in Applied Mathematics and Statistics*, vol. 8, Article ID 940133, 2022.

[20] M. S. Ismail, M. S. M. Noorani, M. Ismail, F. A. Razak, and M. A. Alias, "Early warning signals of financial crises using persistent homology," *Physica A: Statistical Mechanics and Its Applications*, vol. 586, Article ID 126459, 2022.

[21] C. Diks, C. Hommes, and J. Wang, "Critical slowing down as an early warning signal for financial crises?" *Empirical Economics*, vol. 57, no. 4, pp. 1201–1228, 2019.

[22] M. Kozłowska, M. Denys, M. Wiliński et al., "Dynamic bifurcations on financial markets," *Chaos, Solitons & Fractals*, vol. 88, pp. 126–142, 2016.

[23] H. Wen, M. P. Ciamarra, and S. A. Cheong, "How one might miss early warning signals of critical transitions in time series data: a systematic study of two major currency pairs," *PLoS ONE*, vol. 13, no. 3, Article ID e0191439, 2018.

[24] J. P. L. Tan and S. S. A. Cheong, "Critical slowing down associated with regime shifts in the US housing market," *The European Physical Journal B*, vol. 87, no. 2, p. 38, 2014.

[25] J. W. Kantelhardt, "Fractal and multifractal time series," in *Encyclopedia of Complexity and Systems Science*, R. Meyers, Ed., Springer, NY, USA, 2009.

[26] A. Dmitriev, V. Kornilov, V. Dmitriev, and N. Abbas, "Early warning signals for critical transitions in sandpile cellular automata," *Frontiers in Physics*, vol. 10, Article ID 839383, 2022.

[27] S. S. Manna, "Two-state model of self-organized criticality," *Journal of Physics A: Mathematical and General*, vol. 24, no. 7, pp. L363–L369, 1991.

[28] C. K. Tse, J. Liu, and F. C. M. Lau, "A network perspective of the stock market," *Journal of Empirical Finance*, vol. 17, no. 4, pp. 659–667, 2010.

[29] M. F. B. Granha, A. L. M. Vilela, and C. Wang, "Opinion dynamics in financial markets via random networks," 2022, https://arxiv.org/abs/2201.07214.

[30] D. Dhar and R. Ramaswamy, "Exactly solved model of self-organized critical phenomena," *Physical Review Letters*, vol. 63, no. 16, pp. 1659–1662, 1989.

[31] R. Pastor-Satorras and A. Vespignani, "Universality classes in directed sandpile models," *Journal of Physics A: Mathematical and General*, vol. 33, no. 3, L39 pages, 2000.

[32] A. Lara-Sagahón, T. Govezensky, R. Mendez-Sanchez, and M. Jose, "A lattice-based model of rotavirus epidemics," *Physica A: Statistical Mechanics and Its Applications*, vol. 359, pp. 525–537, 2006.

[33] B. Drossel and F. Schwabl, "Forest-fire model with immune trees," *Physica A: Statistical Mechanics and Its Applications*, vol. 199, no. 2, pp. 183–197, 1993.

[34] S. Zhang and W. Fang, "Multifractal behaviors of stock indices and their ability to improve forecasting in a volatility clustering period," *Entropy*, vol. 23, no. 8, p. 1018, 2021.

[35] E. Canessa, "Multifractality in time series," *Journal of Physics A: Mathematical and General*, vol. 33, no. 19, pp. 3637–3651, 2000.

[36] Z. Q. Jiang, W. J. Xie, W. X. Zhou, and D. Sornette, "Multifractal analysis of financial markets: a review," *Reports on Progress in Physics*, vol. 82, no. 12, Article ID 125901, 2019.

[37] E. Green, W. Hanan, and D. Heffernan, "The origins of multifractality in financial time series and the effect of extreme events," *The European Physical Journal B*, vol. 87, no. 6, p. 129, 2014.

[38] F. M. Siokis, "Multifractal analysis of stock exchange crashes," *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 5, pp. 1164–1171, 2013.

[39] Y. Li, "Multifractal view on China's stock market crashes," *Physica A: Statistical Mechanics and Its Applications*, vol. 536, Article ID 122591, 2019.

[40] W. Schadner, "On the persistence of market sentiment: a multifractal fluctuation analysis," *Physica A: Statistical Mechanics and Its Applications*, vol. 581, Article ID 126242, 2021.

[41] L. Zhao, W. Li, C. Yang, J. Han, Z. Su, and Y. Zou, "Multifractality and network analysis of phase transition," *PLoS ONE*, vol. 12, no. 1, Article ID e0170467, 2017.

[42] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," *Physica A: Statistical Mechanics and Its Applications*, vol. 316, no. 1-4, pp. 87–114, 2002.

[43] L. Rydin Gorjão, G. Hassan, J. Kurths, and D. Witthaut, "MFDFA: efficient multifractal detrended fluctuation analysis in python," *Computer Physics Communications*, vol. 273, Article ID 108254, 2022.

[44] J. F. Muzy, E. Bacry, and A. Arneodo, "Multifractal formalism for fractal signals: the structure-function approach versus the wavelet-transform modulus-maxima method," *Physical Review E*, vol. 47, no. 2, pp. 875–884, 1993.

[45] A. N. Pavlov and V. S. Anishchenko, "Multifractal analysis of complex signals," *Physics-Uspekhi*, vol. 50, no. 8, 819 pages, 2007.

[46] H. Wendt and P. Abry, "Multifractality tests using bootstrapped wavelet leaders," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 4811–4820, 2007.

[47] W. Jeżewski, "Multifractal and critical properties of the Ising model," *Physical Review B*, vol. 57, no. 17, pp. 10240–10243, 1998.

[48] F. Nazarimehr, S. Jafari, M. Perc, and J. C. Sprott, "Critical slowing down indicators," *Europhysics Letters*, vol. 132, no. 1, Article ID 18001, 2020.

[49] M. Scheffer, S. R. Carpenter, T. M. Lenton et al., "Anticipating critical transitions," *Science*, vol. 338, no. 6105, pp. 344–348, 2012.

Hindawi
Complexity
Volume 2022, Article ID 4989344, 19 pages
https://doi.org/10.1155/2022/4989344

Research Article

CIMA: A Novel Classification-Integrated Moving Average Model for Smart Lighting Intelligent Control Based on Human Presence

Aji Gautama Putrada,[1] **Maman Abdurohman**,[2] **Doan Perdana**,[1] **and Hilal Hudan Nuha**[2]

[1]*Advanced and Creative Networks Research Center, Telkom University, Bandung, Indonesia*
[2]*School of Computing, Telkom University, Bandung, Indonesia*

Correspondence should be addressed to Maman Abdurohman; abdurohman@telkomuniversity.ac.id

Received 17 March 2022; Accepted 20 August 2022; Published 21 September 2022

Academic Editor: Gonzalo Farias

Smart lighting systems utilize advanced data, control, and communication technologies and allow users to control lights in new ways. However, achieving user comfort, which should be the focus of smart lighting research, is challenging. One cause is the passive infrared (PIR) sensor that inaccurately detects human presence to control artificial lighting. We propose a novel classification-integrated moving average (CIMA) model method to solve the problem. The moving average (MA) increases the Pearson correlation (PC) coefficient of motion sensor features to human presence. The classification model is for a smart lighting intelligent control based on these features. Several classification models are proposed and compared, namely, *k*-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), näive Bayes (NB), and ensemble voting (EV). We build an Internet of things (IoT) system to collect movement data. It consists of a PIR sensor, a NodeMCU microcontroller, a Raspberry Pi-based platform, a relay, and LED lighting. With a sampling rate of 10 seconds and a collection period of 7 days, the system achieved 56852 data records. In the PC test, movement data from the PIR sensor has a correlation coefficient of 0.36 to attendance, while the MA correlation to attendance can reach 0.56. In an exhaustive search of an optimum classification model, KNN has the best and the most robust performance, with an accuracy of 99.8%. It is more accurate than direct light control decisions based on motion sensors, which are 67.6%. Our proposed method can increase the correlation value of movement features on attendance. At the same time, an accurate and robust KNN classification model is applicable for human presence-based smart lighting control.

1. Introduction

Smart lighting systems utilize advanced data, control, and communication technologies and allow users to control lights in new ways [1]. Smart lighting products are already on the market, where their global revenue is up to US$600 million in 2020 [2]. The main issue of smart lighting research is energy efficiency, in which until 2021, 232 out of 384 papers on smart lighting try to solve this problem [3]. The main targets for smart lighting installations are on roads, offices, and housings [4]. Noting the needs of such targets, user comfort and security also become important in smart lighting. However, achieving user comfort is still challenging because the passive infrared (PIR) sensor, a low-price movement sensor, inaccurately detects human presence to control artificial lighting [5].

A smart thing device such as smart lighting should be able to co-operate with its users and environment intelligently [6]. Gartner stated that intelligence is one of five key factors in smart lighting. Activity recognition is an example of intelligence implementation, where it detects human activity based on machine learning applications on several types of sensors [7]. Intelligence can also be applied to improve uncertainty problems in conventional control systems, hence creating an intelligent control system [8].

Several previous studies have tried to overcome the problem of motion sensors to improve accuracy for smart lighting intelligent control based on human presence. Jin

et al. [9] used a time-series-artificial neural network (TS-ANN) on historical PIR sensor data and got up to 97% accuracy in human presence predictive control based on human presence. Fakhruddin et al. [10] used activity recognition to detect five activities using four PIR sensors installed in the house using the principal component analysis-k-nearest neighbor (PCA-KNN) method and to get an accuracy of 94%. Lupion et al. [11] made another study that uses activity recognition and utilizes feature extraction from sliding windows on various sensor data used to produce 99.26% accuracy in detecting 14 activities using the random forest classification method. Park et al. [12] used reinforcement learning (RL) on the PIR sensor and several other sensors to get smart lighting that is adaptive to user needs and also energy-efficiency.

Reconsidering [9, 11], we can think of human presence as a type of activity. On the other hand, we can also consider historical data as a sliding window feature extraction. A moving average (MA) concept can substitute the sliding window feature extraction method in this intuition. Usually, MA is a method for smooth fluctuating data and, among others, can be used as a noise filtering method for time-series data [13]. In some research, MA is used to increase the Pearson correlation (PC) coefficient of machine learning features [14]. Furthermore, we can conduct a comprehensive test to find the optimum classification model. Several studies use some well-known classical machine learning methods such as KNN, support vector machine (SVM), decision tree (DT), and näive Bayes (NB) to train the classification model [15]. Other research also uses ensemble learning methods such as ensemble voting (EV) to improve the performance of the existing classical machine learning method [16].

We propose a novel classification-integrated MA (CIMA) model method to solve the problem. The MA is to increase the correlation of motion sensor features to human presence, while the classification model is for a smart lighting intelligent control based on these features. We train the proposed classification model with KNN, SVM, DT, and NB. We also use ensemble learning methods such as EV to improve classical machine learning performance. An Internet of things (IoT) system is built on a test-bed environment to retrieve movement data from the PIR sensor. At the end-device layer, the microcontroller used is NodeMCU. We build a Node-Red server on the Raspberry Pi at the Platform layer. It stores the movement data log in a comma-separated value (CSV) file. We use test parameters such as accuracy, precision, recall, and $F1$-score to discover the optimal classification model. In addition, to check the robustness of the model, the cross-validation method is used.

The main contributions of our work are listed below:

Increasing the correlation between movement data and human presence through the MA method. A novel classification model with significant accuracy from state-of-the-art research utilizing MA data from movement data as a feature. An accurate yet low-price solution for human presence-based smart lighting control because of the utilization of motion sensors.

The remainder of this document has the following systematic: Section 2 presents works related to the research undertaken, Section 3 describes methods used in this research, Section 4 gives the results of the tests conducted, Section 5 reports the results and compares them with state-of-the-art studies while highlighting the contributions provided from our work, and finally, Section 6 emphasizes the important findings of this study.

2. Related Works

Several studies have discussed automatic smart lighting control using the PIR sensor. Jin et al. [9] aimed to improve the accuracy of PIR sensors using the time-series-artificial neural network (TS-ANN) method and compared several features such as time, occupied ratio, time steps, and historical occupied state data. The study showed that the proposed method can provide up to 97% accuracy for the intelligent control. Putrada et al. [17] used a hierarchical hidden Markov model (HHMM) to classify five different types of activities from four PIR sensors to control smart lighting in offices. The HHMM model tested is better than the hidden Markov model (HMM), NB, and KNN method and has an accuracy of 87.6%. Ramadhan et al. [5] also used the HHMM method on 14 different activities from five PIR sensors. The accuracy of HHMM was 93%, and the method was superior to HMM. Fakhruddin et al. [10] used activity recognition to detect five activities using four PIR sensors installed in the house using the principal component analysis-k-nearest neighbor (PCA-KNN) method and to get an accuracy of 94%. Each study investigates a different amount of activity and obtains varying performance. There is an opportunity to find a correlation between the number of activities and performance in using PIR sensors for activity recognition on smart lighting. Other factors are also opportunities for investigation.

Furthermore, other studies also conducted smart lighting control but with devices or sensors other than the PIR sensor. Dai et al. [18] used five low-resolution cameras to detect nine activities in a smart lighting environment. The study provided a solution that ensures privacy even when using a camera, while the accuracy is up to 89.6%. Chun et al. [19] used a depth camera to detect four human activities in a room. The proposed method results provided 100% accuracy for the location where people are and 78.3% accuracy for the type of performed activity. Lupion et al. [11] used PIR sensors and also several different sensors including smartwatches and real-time location systems. The research produced 99.26% accuracy in detecting 14 activities using a random forest classification method. Park et al. [12] used a light sensor and actuators such as Switchmate. Switchmate consists of a motor and a position sensor to control and monitor a conventional light switch. The average light utility ratio (LUR) of the research is 67%. The studies mentioned have performance that vary from inadequate to highly adequate results. However, the equipment used is expensive when compared to the PIR sensor, which costs around US$ 1. There is an opportunity to implement an accurate and low-priced solution using CIMA and PIR sensors.

Several previous studies have applied MAs for smoothing and increasing the PC between two variables.

Husnayain et al. [20] used MA to increase the correlation between the incidence of dengue fever with Google search activities for dengue and found that the correlation was very high between the two. Hu et al. [21] utilized MAs to reduce noise in water pH and water temperature data to improve the correlation of the two data with other water quality data to provide better performance in mariculture water quality forecasting. Peng et al. [22] used MA to increase the PC between drought and flood to predict the occurrence of these two disasters in China. Badr et al. [23] showed that the correlation between the mobility ratio and growth rate ratio increased as the MA window size increased but slowly decreased when the window size was too large. Singh et al. [24] used MAs to refine CO_2 sensor readings and improved the correlation of sensor data with respiratory rate and Hjorth activity in a cardiorespiratory assessment. The results of the mentioned studies show that there is an opportunity to apply MA to movement data to increase the PC coefficient for an accurate classification model.

3. Materials and Methods

3.1. Research Methodology. This section discusses the research methodology, from how the test data were collected, to how we obtained the final model. The methodology for developing a classification model to predict human presence is shown in Figure 1.

The PIR sensor is one of the most utilized sensor in smart lighting control [25]. We build an IoT system with PIR sensors to collect human movement data. Labeling is done to each movement as to whether there are people or not at each given moment. The system stores the data in a CSV file for further analysis. The next step is to apply the MA and observe the PC coefficient. Further is to prepare data before conducting classification training with methods KNN, SVM, DT, and NB. The possibility of applying EV to improve the performance of the classification model is analyzed later. The last step is to analyze the most optimum model and perform cross-validation to check for possible overfitting.

3.2. Smart Lighting IoT System. The IoT architecture of the smart lighting system for automatic light control based on human presence is as shown in Figure 2. We chose a living room as a test-bed environment to implement the proposed architecture.

In the proposed IoT architecture, there are three main layers, namely, the end-device layer, platform layer, and application layer [26]. Then, there are additional communication protocols and gateways that connect the three layers. At the end-device layer, the layer directly related to the IoT hardware, the three main devices are PIR sensors, NodeMCU, and relays. The PIR sensor functions to detect human movement [27]. NodeMCU has a system on chip (SoC), ESP8266, which includes a microcontroller and WiFi communication [28]. WiFi is used for communication between the end-device layer and platform layer [29]. The relay is an actuator connected to the LED light [30]. Its function is to turn the LED on and off like a switch controlled via the microcontroller.

We build the platform layer on a Raspberry Pi (Raspi), an open-source mini-personal computer (mini-PC) running with a Raspbian operating system (OS) [31]. We use Node-Red for web service functions. Node-Red is also an open-source web service based on Node.js, which has a special add-on for IoT systems [32]. The Node-Red performs movement sensor data log dumps to a CSV file used for training the classification model. Raspi can also be used to run Python functions [33]. Hence, the classification model running in Python can be executed on this server.

The application layer is concerned with the interaction between the system and the user. Users can use the Python-based graphical user interface (GUI) to set the light status manually or automatically. Especially for testing, the user can also choose to control lights with the novel method or the conventional method, which compares the comfort between the new system and the legacy system. The platform layer links to the application and end-device layers via the Internet and the hypertext transfer protocol (HTTP) application programming interface (API) protocol.

The device is a single set and detects the presence of one person at one location in one room. A chart depicting the placement of devices in a room is shown in Figure 3. The PIR sensor, NodeMCU, and relay are on the ceiling as part of the end-device. The PIR sensor is placed approximately above where humans conduct activities, for example, working. The end devices, especially the relay, are connected to the LED light. The LED light is on the ceiling in the middle of the room. The NodeMCU receives motion sensor data, controls the LED light via relays, and communicates with the IoT Platform via WiFi. A wall-mounted WiFi-4G router connects the WiFi network with the Internet.

The motion detection distance from the sensor is 10 meters forward. In addition, the PIR sensor has a capture range as wide as 110°. Figure 4 shows the coverage area of the PIR sensor when placed on the ceiling. If, for example, the room's height is 2.4 m, with the range described previously, then the coverage area will form a cone with a base diameter of 5 m and a base radius of 2.5 m. Hence, the area of the cone base is approximately 20 m^2. The proposed smart lighting system hypothetically considers a person present if the person is in that mentioned space.

3.3. Moving Average. As the name suggests, MA is a method of averaging on time-series data in which a certain period of data (called data points) is averaged continuously and moves along the data series [34]. The data points are notated as N. Applying the MA results in a smoother data series [35]. Due to this nature, scientists and analysts utilize MAs in cases involving fluctuating data such as financial data, stock predictions, and signal filters [36, 37]. The MA formula for N values is as follows:

$$MA(n) = \frac{1}{N} \sum_{i=n-N+1}^{n} p_i, \tag{1}$$

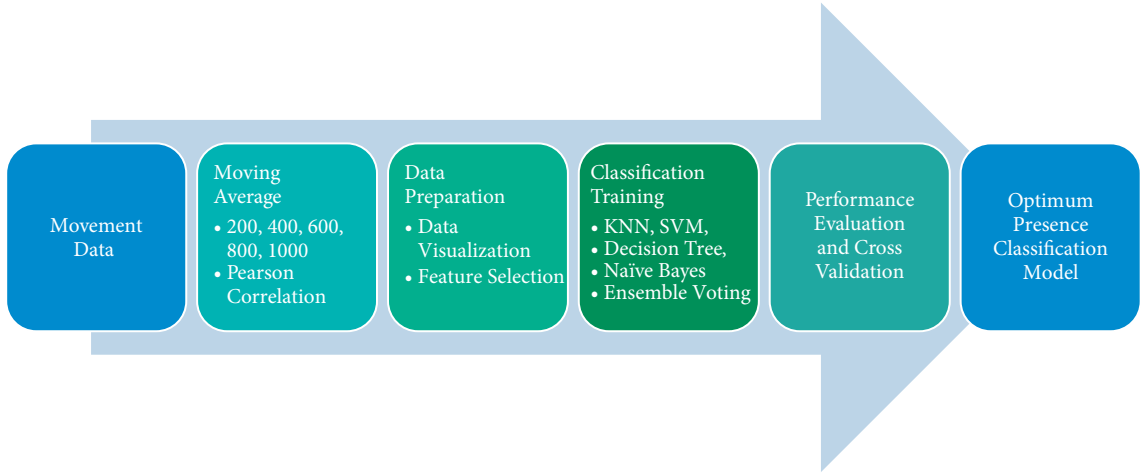


Figure 1: A chart explaining the proposed methodology for developing a classification model for predicting human presence.

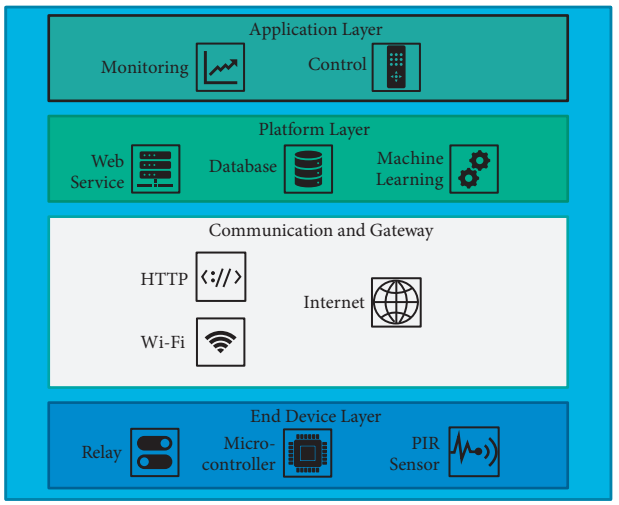


Figure 2: The IoT architecture of a smart lighting system for automatic light control based on human presence.

where p_i is the i^{th} data series in range $n-N+1$ to n and $MA(n)$ is the MA on p_n. The MA for N values and the following n data $(n+1)$ can use the following formula:

$$MA(n+1) = MA(n) + \frac{1}{N}(p_{n+1} - p_{n-N+1}). \tag{2}$$

We also introduce a novel theorem for $MA(n-1)$. The description is given in Theorem 1. This theorem is a low-level solution that simplifies the complexity of our real-time system in calculating the MA. It considers the property of the data structure in use.

Theorem 1. *If $MA(n+1)$ is given by equation (2), then $MA(n-1)$ is given by the following formula:*

$$MA(n-1) = MA(n) + \frac{1}{N}(p_{n-N} - p_n). \tag{3}$$

Proof. Consider a signal p, the $MA(n+1)$ is given by equation (2). Suppose $n = m-1$, substituting n with $m-1$ in equation (2) yields the following formula:

$$MA(m) = MA(m-1) + \frac{1}{N}(p_m - p_{m-N}). \tag{4}$$

Then, the formulas yield

$$\begin{aligned} MA(m) - \frac{1}{N}(p_m - p_{m-N}) &= MA(m-1), \\ MA(m) + \frac{1}{N}(p_{m-N} - p_m) &= MA(m-1). \end{aligned} \tag{5}$$

Moving term $MA(m-1)$ to the left side of the equation and substituting back m with n yield,

$$MA(n-1) = MA(n) + \frac{1}{N}(p_{n-N} - p_n). \tag{6}$$

□

3.4. Classification Models. Assuming our hypothesis on MA is correct, we carry out a comprehensive test to find the optimum classification model in determining attendance based on the novel movement data. The classification methods used are KNN, SVM, DT, and NB. The ensemble learning method can also improve the performance of conventional classification methods. Here we propose EV to combine several classical classification models.

KNN is a type of supervised machine learning that makes decisions based on the closest k training example to a data whose class is unknown [38]. One way to measure the closest distance of data with a training dataset is the Euclidean distance. The formula for calculating the distance in KNN with Euclidean distance is as follows:

$$\text{Distance}(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}, \tag{7}$$

where x is the training dataset, y is the classified data, and n is the number of features in the dataset.

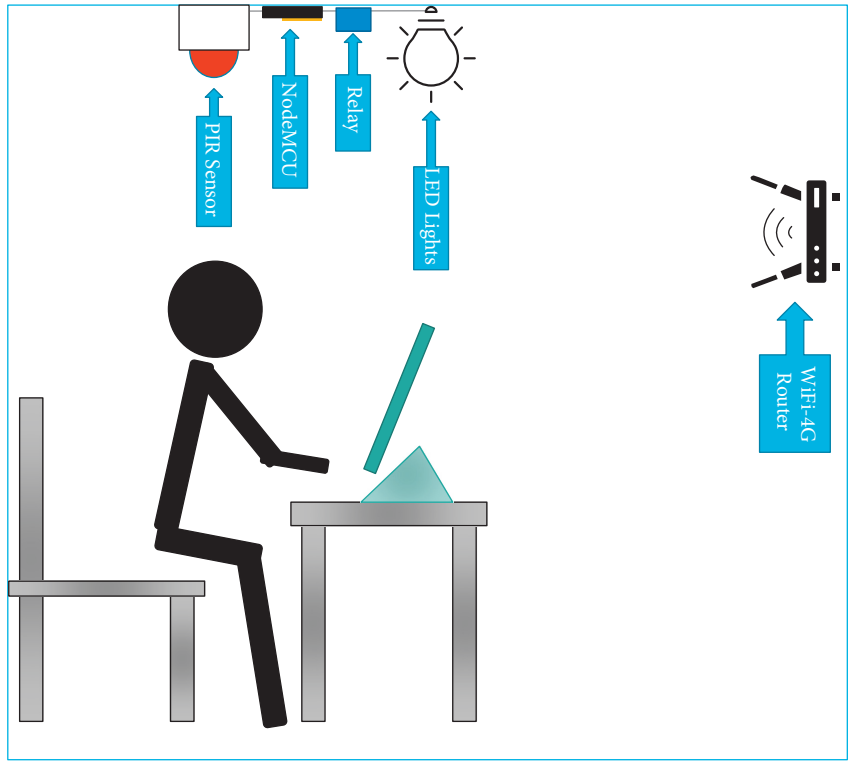


Figure 3: A chart depicting the placement of devices in a room.

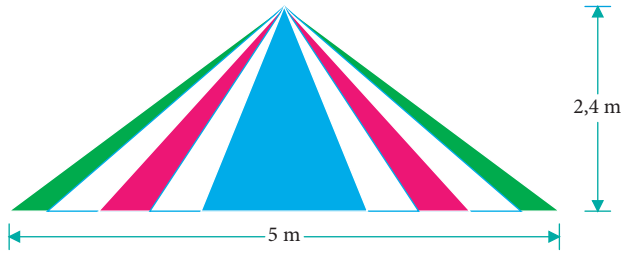


Figure 4: The coverage area of the PIR sensor if placed on the ceiling of the room.

A data structure contains the distance of y with all training examples x. As much as k training example xs closest to y are moved to a new data structure. From the k training example xs, the algorithm chooses the class with the most training example x (calculated with a mode function) as the class of y. Varying the k value influences the KNN model performance. Hence, a further test finds the optimum k value.

SVM is an example of supervised machine learning that uses margins to classify [39]. The classification method is to create a hyperplane to separate the different classes in the dataset [40]. Several kernels determine which hyperplanes can be created, including polynomial, radial basis function (RBF), and sigmoid. Polynomial kernels can use up to some different degrees. Linear kernel is considered a first-degree polynomial kernel. Here is the formula for the SVM polynomial kernel with d-degrees, including the linear kernel,

$$K(x, x') = (x.x' + r)^d, \tag{8}$$

where x and x' are vectors in the input space and r is a free parameter.

The RBF kernel is one of the most used kernel [41]. The kernel's formula of the two vectors x and x' is as follows:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{9}$$

where $\|x - x'\|^2$ calculates the squared Euclidean distance between x and x' and σ is a free parameter [42].

The sigmoid kernel formula for the two vectors x and x' is as follows:

$$K(x, x') = \tanh\left(\gamma x^T x' + c\right), \tag{10}$$

where γ is a free parameter with a value greater than 0 and c is a free parameter with a value less than 0.

If the dataset is linearly separable, then the suitable kernel is a linear kernel. However, if the dataset is non-linearly separable, a kernel that fits between polynomials (several d-degrees are useable), RBF, or sigmoid is the solution.

The SVM classification function is as follows:

$$f(x) = \sum_{i=0}^{n} a_i y_i K(x, x') + b, \tag{11}$$

where a_i is the Lagrange multiplier, y_i is the y value of x_i, and b is the intercept.

The DT is a classification model which is essentially a binary tree, where each branch in the tree is an ordinary if-else decision [43]. However, the if-else decision comes from a training process through several stages [44]. The two most common types of DTs are iterative dichotomiser 3 (ID3), and classification and regression tree (CART) [45]. The main difference between the two is that ID3 can only be used for classification, while CART can be used for classification as well as regression [46]. The CART formation uses a calculation of the Gini index of each feature. The Gini index describes the inequality value of a feature [47]. The lower the Gini index value, the better the feature is used to make decisions. The Gini index formula is as follows:

$$\text{Gini}(p) = 1 - \sum_{i=1}^{J} p_i^2, \tag{12}$$

where p is the feature index, p_i is the fraction of the feature p with the label i, and J is the number of labels present.

If, after the decision, the resulting class is still not uniform, then the process of calculating the Gini index for that branch is repeated for other features. The process is iterative until all branches produce a uniform class or have reached the max depth limit. Max depth is the farthest distance from the root to the leaf. Limiting the max depth value is usually to prevent overfitting.

NB classifies with the concept of the Bayes theorem, which is looking for opportunities from a hypothesis on events that have never happened [48]. NB is an efficient algorithm because each variable can be independent. The following is the formula used for the classification of NB:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}, \tag{13}$$

where x is the data to be classified, c is the hypothetical data of a class, and $P(c|x)$ is the a posteriori probability of the data c against x.

Ensemble learning is a method of combining several learning models where the results are usually better than if only one of its members is used [49]. The downside of ensemble learning is that the algorithm is usually more computationally heavy [50]. EV is a type of ensemble learning in which, by utilizing several models from several different methods, EV selects the answer with the most number of results from each model [51]. EV can exploit the peculiarity of each member's classification model so that the advantages of each model can be seen in the results of the ensemble [52]. In hard EV, the formula used is as follows:

$$\hat{z} = \text{mode}\{X_1(y), X_2(y), \ldots, X_a(y)\}, \tag{14}$$

where $\hat{z}$ is the classification result of the EV, X is each classification model, a is the number of classification models used, and y is the data to be classified.

3.5. Evaluation Metrics. PC measures the linear correlation between two datasets [53]. The usual denotation for PC is the letter r, and the PC formula between data x and data y is as follows:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}}, \tag{15}$$

where n is the number of records in the dataset [54].

The range of the calculated values for the PC formula is -1 to 1. There are several interpretations of the results of the PC. A negative result means that the x and y datasets have a negative correlation, where if the results are positive, the x and y data have a positive correlation. If $0.5 < |r| < 1.0$, then there is moderate to strong correlation between x and y. If $0.0 < |r| < 0.5$, there is no correlation, there is a non-linear correlation, or there is a low correlation between x and y [55].

PC is useful for feature selection in machine learning. Features with a moderate to strong correlation with the label usually pass the selection and continue to the training stage of machine learning. Features that have no correlation or low correlation are eliminated and cannot continue to the training stage [56].

The confusion matrix forms a quadrant for models with binary classification, which only involves two output values. In that quadrant, each row has data with actual positive output and data with actual negative output. Further on, each column has data with predicted positive output and data with predicted negative output. Each cell in the quadrant is an intersection between the sets of each row and each column, resulting in four possible outcomes: True Positive (TP), False Negative (FN), True Negative (TN), or False Positive (FP). The confusion matrix results show a model's predictive ability and strengthen the explanation of its accuracy, precision, recall, and $F1$-score result.

Accuracy is the ability of a model to predict data correctly. The accuracy formula is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{16}$$

Accuracy can only measure the ability of a model to predict the correct data but cannot describe the specific capabilities of a model in making predictions. Therefore, other metrics such as precision, recall, and $F1$-score are used.

Precision shows the ability of a model to sort the negative class from the positive class. The precision formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{17}$$

Recall shows the ability of a model to predict the positive class. In some cases, accuracy is often mistaken for recall, whereas in imbalanced data, recall gives a true picture of the model's ability to predict positive classes. The recall formula is as follows:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{18}$$

$F1$-score is a value that describes a combination of precision and recall capabilities. The $F1$-score is different from the average because the $F1$-score uses the concept of a harmonic average. Even though it combines precision and

recall, the $F1$-score value is usually different from accuracy. The $F1-\text{score}$ formula is as follows:

$$F1-\text{score}=2\cdot\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{19}$$

Sometimes a model can experience overfitting, which is a condition when the model produces good performance on training but poor performance on validation [57]. The characteristic of an overfitting model is that it has high variance and low bias [58]. High complexity is another nature of an overfitting model. The cross-validation method can examine models with high complexity. In K-fold cross-validation, the method divides training data into several random subsamples of the same size. The fold is the term for each subsample, where K is the number of subsamples. After division, the method performs K iterations. It uses one different fold as validation data in each iteration and the rest as train data. In each iteration, accuracy or one other performance metric evaluates the model. At the end of execution, the average accuracy of each iteration becomes the final result of the cross-validation evaluation [59]. The complete process of K-fold cross-validation is given in Algorithm 1.

4. Results

4.1. IoT Implementation and Data Collection. With the IoT architecture as described in Subsection 3.2, we implement the proposed human presence-based smart lighting control. Parts of the implementation are shown in Figure 5. The main parts of the implementation are PIR sensors, NodeMCU, 4G-WiFi router, Raspberry Pi, Google Sheets, LED lighting, and relays. The Google Sheets monitors the sensed movement data. Moreover, the Raspberry Pi saves a CSV file containing movement data.

Data collection begins after the smart lighting system with the PIR sensor is successfully implemented. Movement data is collected with a sampling rate of 10 seconds and collected for seven days in one test area. During that period, the system collected 56852 data records. The data consists of movement data with binary values. A value of 1 means the PIR sensor detects movement. Otherwise, 0 means there is no movement. Each data is labeled manually. The label describes the presence of people in the room. The manual filling is done based on the presence of a subject in the room. The specification of movement data collection is given in Table 1.

A line plot can visualize how sensor data capture human movement and how the data looks compared to the actual human presence in the room. A partial snippet of the movement dataset with attendance labels is shown in Figure 6. The snippet shows data from two days out of seven days of data collection. The line plot explains that the PIR sensor reports 0 even though a subject is present. It is not that the PIR sensor is not accurate enough, but more because, while PIR sensors can only detect movement, one can imagine that subjects are not always moving while they are present. It is conceivable that if a smart lighting system directly uses the PIR sensor results for light control, the lights will turn on and off while people are still present. It results in disturbance to people's comfort.

4.2. Moving Average Application. The intuition is that the application of MA to the movement data results in a curve with a PC coefficient closer to human presence than movement data. Visualization in the form of a line plot can help illustrate this intuition. The line plot of MA results, movement data, and human presence are shown in Figure 7. Movement data of Day 1 goes through a MA with $N = 200$. The plot shows that the MA curve elevates when people are present and approaches 0 when otherwise. However, it does not fully resemble human presence.

The PC evidences the closeness of the MA value to human presence data according to equation (16). We create five MA curves with different N values and observe which curve has the strongest PC coefficient. A matrix showing the PC of movement, five types of MAs, and human presence is shown in Figure 8. Payload is the feature name for movement data. The last row of the matrix shows the PC coefficient of presence with each feature. The highest value is 0.56, which is the MA at $N = 200$. Based on the interpretation, the curve has a moderate positive correlation with human presence. In comparison, the payload correlation is 0.36, which classifies as a low positive correlation.

A line plot can illustrate the growing trend of the PC coefficient based on the number of N values. The line plot is shown in Figure 9. The green line is the growth of the PC based on the increasing N values of the MA, where the red line is the PC of raw movement data. The MA method can increase the PC coefficient of movement features. However, using a data point too large will decrease the correlation. The optimum value is 200.

4.3. Training Classification Models. Because the MA of movement data has a moderate positive correlation with human presence, training a machine learning method with the new feature can hypothetically result in a model with good performance. In the model to be trained, the proposed input features are motion sensor data and some MA curves with different N values. The output class is human presence, with labels 1 for a human being present and 0 for no human present. It means that the type of classification is binary classification. We carry out an exhaustive test to find the optimum classification model for human presence based on movement data. The classification methods used are KNN, SVM, DT, and NB. EV is also applied to improve the performance of some of the mentioned methods.

The training process uses 50% of the dataset, while the testing stage uses the rest. It means there are 23436 training data and 23436 testing data. The dataset is shuffled prior to the data split to prevent uneven distribution. The test metrics are accuracy, precision, recall, and $F1$-score. Six initial features are used including five MA curves with a variation of N values: movement, MA ($N = 200$), MA ($N = 400$), MA ($N = 600$), MA ($N = 800$), and MA ($N = 1000$). The label is human presence. Cross-validation is also applied to

(1) Divide data into K equal folds
(2) **for** k in range $(0, K)$ **do**
(3) $R \leftarrow \text{Fold}_k$ in data
(4) $T \leftarrow$ data/R
(5) Train T
(6) $Acc_k \leftarrow$ evaluate R with trained model
(7) **end for**
(8) $Acc \leftarrow 1/K \sum_{k=1}^{K} Acc_k$

Algorithm 1: K-fold cross-validation.

Figure 5: Parts of the results of implementing the IoT architecture for smart lighting control based on human presence.

Table 1: Movement data collection specification.

Attribute	Value
Sampling rate	10 s
Collecting period	7 days
Collected data	56852 records
Feature	Movement
Feature values	1 (movement detected) 0 (no movement)
Label	Presence
Label values	1 (present) 0 (not present)

test the robustness of each model. A summary of the training specifications is given in Table 2.

In KNN, k describes the number of neighbors involved in calculating the closest distance between test and training data. A test of varying k finds the optimum KNN model. Changes in the value of accuracy, precision, recall, and $F1$-Score to the increase of k in KNN training is shown in Figure 10. The graph shows the comparison of the performance of the KNN model with $k = 1$ to $k = 5$. The values of precision and recall fluctuate, while the $F1$-score and accuracy values have a decreasing trend. Based on these tests, we conclude that $k = 1$ is the exact value for the optimum KNN model.

In SVM, the right type of kernel provides the optimum model. The kernel types compared are the linear kernel, 2^{nd}-degree polynomial, 3^{rd}-degree polynomial, RBF, and sigmoid. A comparison of the performance of the SVM classification model with five kernels is shown in Figure 11. The bar chart compares four performance values: accuracy, precision, recall, and $F1$-score. In all four metrics, sigmoid has the lowest performance. The 2^{nd}-degree polynomial has the highest recall but not the highest $F1$-score. The highest $F1$-score and accuracy go to the 3^{rd}-degree polynomial and RBF. However, the precision value of the 3^{rd}-degree polynomial is lower than RBF. Hence, the RBF kernel provides the optimum SVM model.

Using the right model depends on how to understand the data [60]. In 3.4, it has been explained that the selection of the SVM kernel depends on whether the data is linearly separable or not. In addition, the amount of data and the type of data also affect the selection of the model. A scatter plot matrix helps to understand the data better. The scatter plot matrix is often a tool for understanding high-dimensional data [61]. A visualization of the dataset in the form of a scatter plot matrix is shown in Figure 12. The scatter plot matrix shows that the scatter plot between each feature is not linearly separable. It explains why the linear kernel does not produce an optimum SVM model. Moreover, if the data is non-linearly separable and the RBF kernel is more optimum

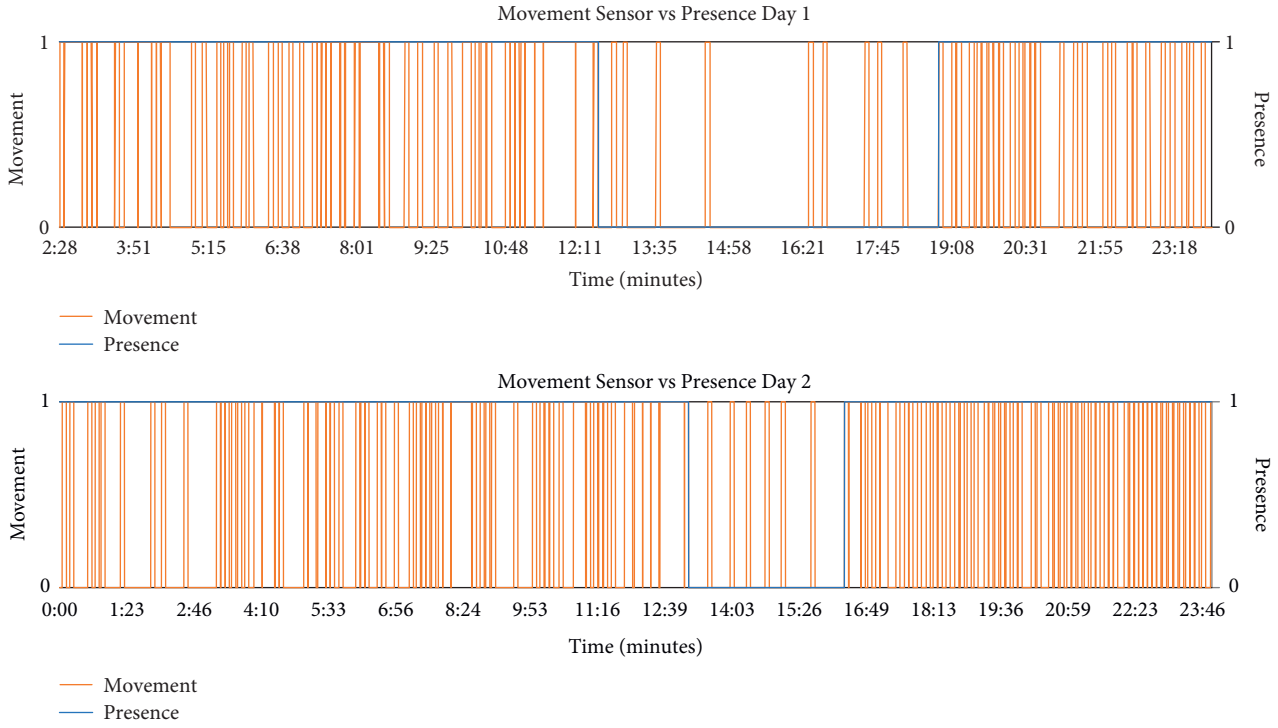


Figure 6: Partial snippet of the movement dataset with attendance labels.

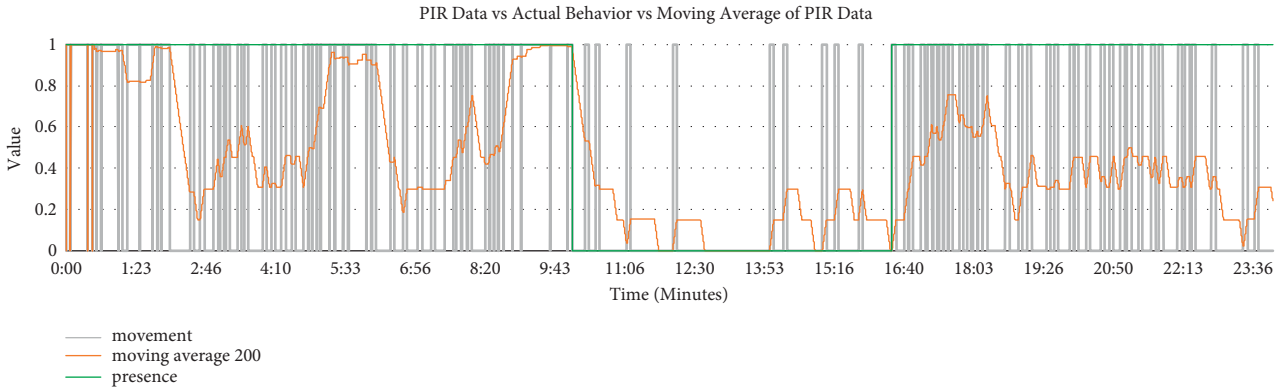


Figure 7: Visualization of the results of applying MA to the movement data and compared with actual human presence.

than other kernels, then the data is radially separable. In two dimensions, the binary class forms a doughnut shape, with one of the classes in the doughnut hole [62].

A too high max depth value in training can result in an overfitting DT model. The symptom of overfitting is that when comparing the performance of the tree with train data and cross-validation, the performance of the train data will continue to increase. In contrast, the cross-validation value will decrease or stagnate. Pruning is a solution to prevent overfitting the model. When using the early stop method in pruning, adding depths to the tree is stopped when the cross-validation value starts to drops [63]. The effect of increasing DT max depth on the accuracy of training data and validation data is shown in Figure 13. The orange line in the graph is the model's accuracy based on the train data, while the blue line is the average accuracy based on cross-validation. After the value of max depth = 12, the accuracy value of the cross-validation value decreases, so max depth = 12 is considered to provide the optimum DT model.

NB is a machine learning method that is more suitable for text-based analysis than classification on sensor data [64, 65]. It is also seen in this case when comparing the confusion matrix of KNN, SVM, DT, and NB. The confusion matrix of the four classifiers is shown in Figure 14. In the comparison, NB has the lowest performance. However, the FN and FP values of SVM, DTs, and NB are worth observing. The FP value of SVM is higher than its FN. However, it is the other way around in NB. They are peculiarities that EV can exploit, so we build it based on the four previous models. The test results complete the confusion matrix comparison. The

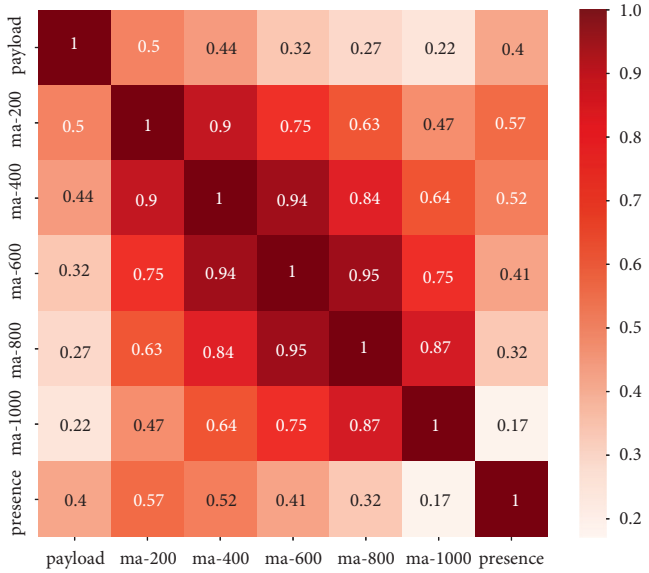


Figure 8: A matrix that displays the PC of movement (payload), five types of MAs, and presence.

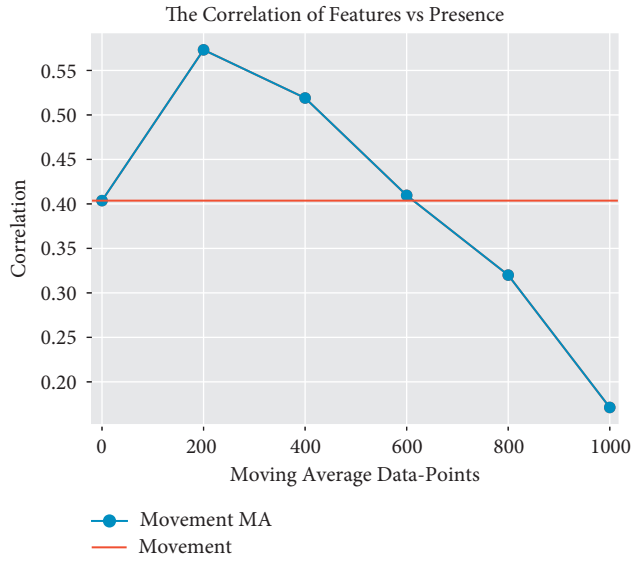


Figure 9: The growth and decline of the MA correlation with the increase in N values.

Table 2: Training specifications.

Attribute	Value
Machine learning methods	KNN, SVM, DT, NB, and EV
Split ratio	50 : 50
Training data	23436 records
Testing data	23436 records
Features	Movement, MA ($N = 200$), MA ($N = 400$), MA ($N = 600$), MA ($N = 800$), and MA ($N = 1000$)
Label	Presence
Data shuffle	On
Performance metrics	Accuracy, precision, recall, $F1$-score, and cross-validation

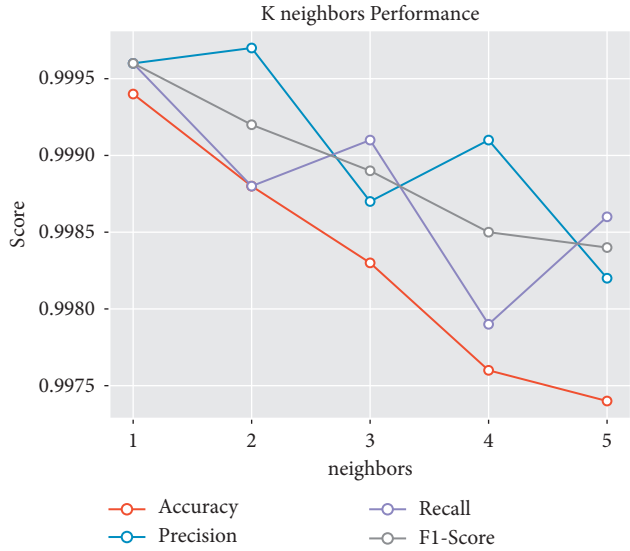


Figure 10: The increase of neighbors on the KNN performance.

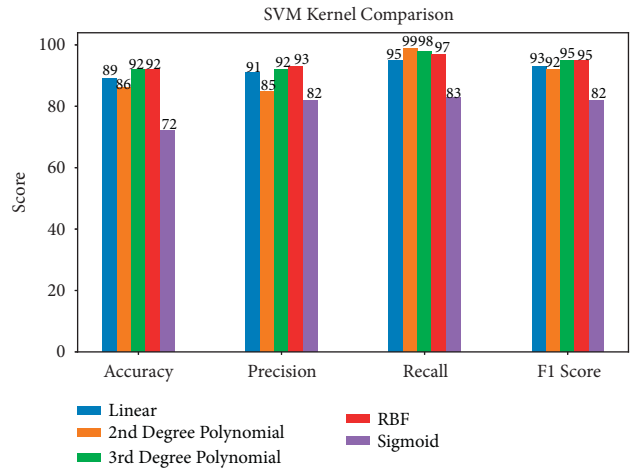


Figure 11: The performance comparison of different kernels on SVM classification.

results show that EV has a better confusion matrix than SVM and NB. In addition, the EV model has the lowest FP compared to SVM, DT, and NB. As a result, EV is a model that optimizes the NB model.

4.4. Performance Evaluation and Cross-Validation. Four optimum classification models can be compared: KNN with $k = 1$, SVM with RBF kernel, DT with max depth = 12, and EV from KNN, SVM, DT, and NB. The performance comparison of the four classifiers is shown in Figure 15. The comparison is in the form of a bar chart. In the bar chart, KNN is the blue bar, SVM is the orange bar, DT is the green bar, and EV is the red bar. Four metrics test the models: accuracy, precision, recall, and *F*1-score. SVM has the lowest performance in all four metrics of the four models. Between the three remaining models, EV is the only model with a recall value below 0.99. KNN excels in all four metrics, even compared to the DT. The optimum classification model for human presence based on movement data is KNN with $k = 1$.

The robustness of each model is also measured. We set SVM aside and only compare the models with the top three best performances from the previous tests, namely, KNN, DT, and EV. *K*-fold cross-validation measures the robustness of each model. The *K* values for testing are 2, 5, and 10 as they are commonly used [58]. Accuracy metric measures each cross-validation iteration. The cross-validation process accuracy comparison of the three models is shown in Figure 16. A box plot visualizes the performance comparison. In addition to the accuracy average, the box plot can also compare the accuracy variance of each case. For each model, the average accuracy trend increases as the number of folds increases. However, for EV specifically, the variance also increases. The EV owns the lowest average accuracy for each *K* value. At $K = 2$, the DT has the highest accuracy variance. For all *K* values, KNN has the lowest variance and the highest average. It concludes that KNN is also the most robust model apart from having the best performance.

The KNN model with $k = 1$ can still be optimized. Not all features will be related to the output class in machine learning. If an irrelevant feature enters the training process, what happens is garbage in, garbage out, and the performance of the model will drop [66]. Hence, at this stage, feature selection is carried out based on the PC value, previously calculated in 4.2. Assuming that increasing the number of uncorrelated features will reduce the performance of the classification model, the following scenarios are made based on the PC value and compared:

(i) 1 feature: MA ($N = 200$).

(ii) 2 features: MA ($N = 200$) and MA ($N = 400$).

(iii) 3 features: MA ($N = 200$), MA ($N = 400$), and MA ($N = 600$).

(iv) 4 features: MA ($N = 200$), MA ($N = 400$), MA ($N = 600$), and MA ($N = 800$).

(v) 5 features: MA ($N = 200$), MA ($N = 400$), MA ($N = 600$), MA ($N = 800$), and MA ($N = 1000$).

(vi) All features: all features included.

The effect of the number of features on the prediction performance of the KNN model is shown in Figure 17. The image is in the form of a line plot. The four metrics compared include accuracy, precision, recall, and *F*1-score. Results show an increasing trend in the addition of the number of features. It proves that although the MA with $N > 200$ has a lower correlation than the MA with $N = 200$, these features are still relevant in classifying human presence. Subsequently, the model with five features and six features has the same performance. The conclusion is that if the raw movement data departs the dataset, the model performance does not decrease, reducing complexity. Hence, the feature selection process result concludes that the KNN model with five features is the optimum KNN model.

For example, two features have a high correlation with the output class. In the understanding of multicollinearity, if

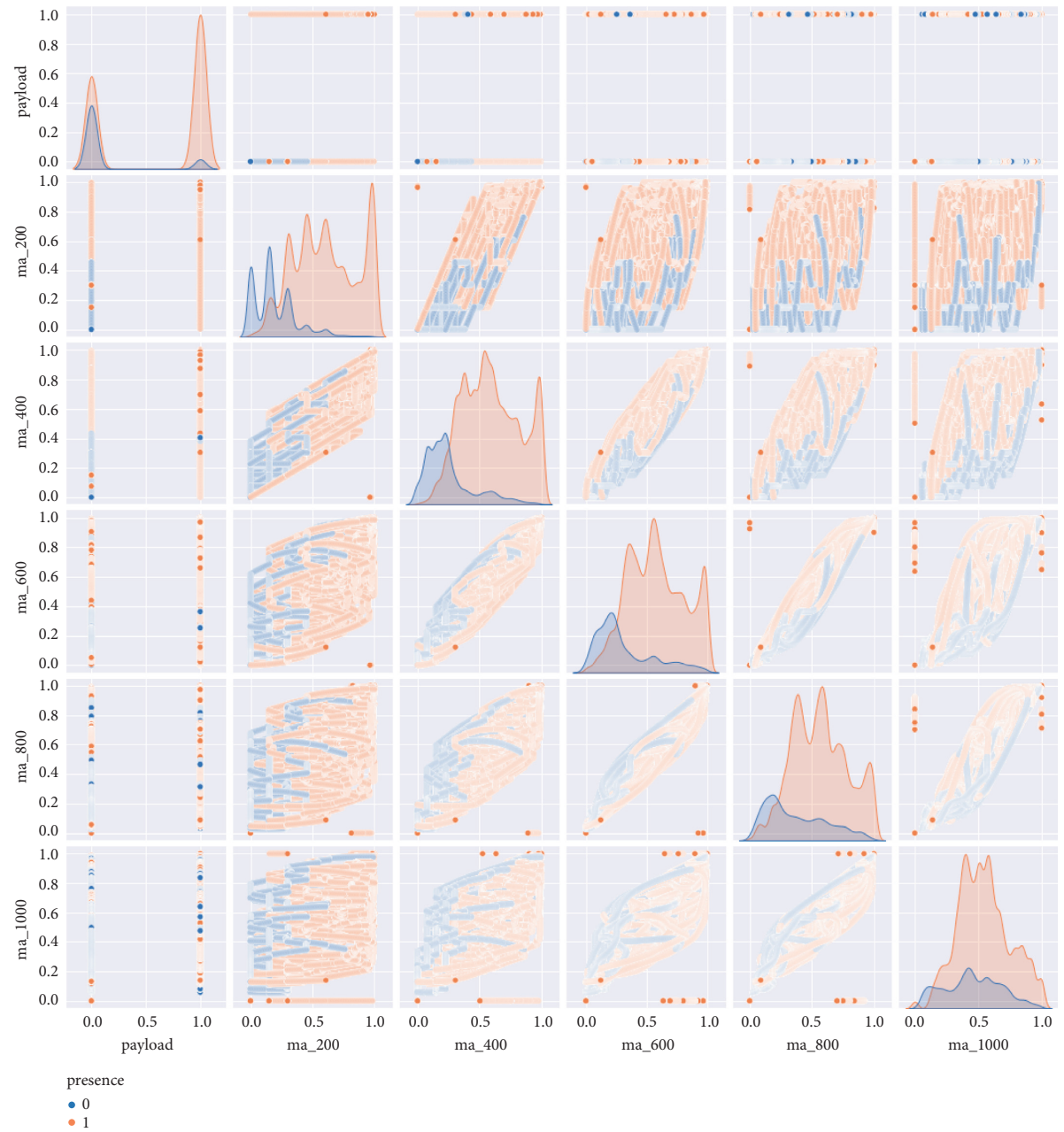


Figure 12: The scatter plot matrix of all features.

the two features have a high correlation and one of them is not excluded, the performance of the model will become poor, especially the linear regression model [67]. For example, revisiting the PC matrix in Figure 8, MA ($N = 600$) is highly correlated to MA ($N = 800$). We investigate this by applying the moving PC to the time-series dataset. A visualization of the application of the moving PC with $N = 6000$ to the dataset can be seen in Figure 18. We take a snapshot of two different cases. The upper part of the image is a situation where there is not much fluctuation in attendance. In this situation, MAs with high N have a high correlation. The bottom part of the image is a situation where there is much fluctuation in human presence. The MA with low N has a high correlation in this situation. It explains why the 5-feature model has the best performance.

We use the test data to measure how directly using PIR sensor movement data to lighting control would perform. We call it the raw method. The significance of the presence classification model (proposed method) on the raw method appears in a side-to-side comparison. The comparison of the two methods is shown in Figure 19. The image is in the form of a bar plot. It shows accuracy, precision, recall, and $F1$-score, where the proposed method is the blue bar, and raw is the orange bar. The two biggest significances are accuracy and recall, 99.7% to 67.8% and 99.8% to 62.6%, respectively.

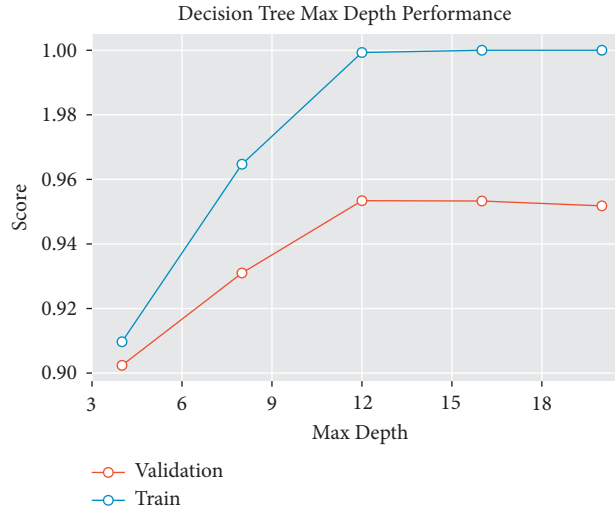


Figure 13: The increase of max depth on the DT accuracy.

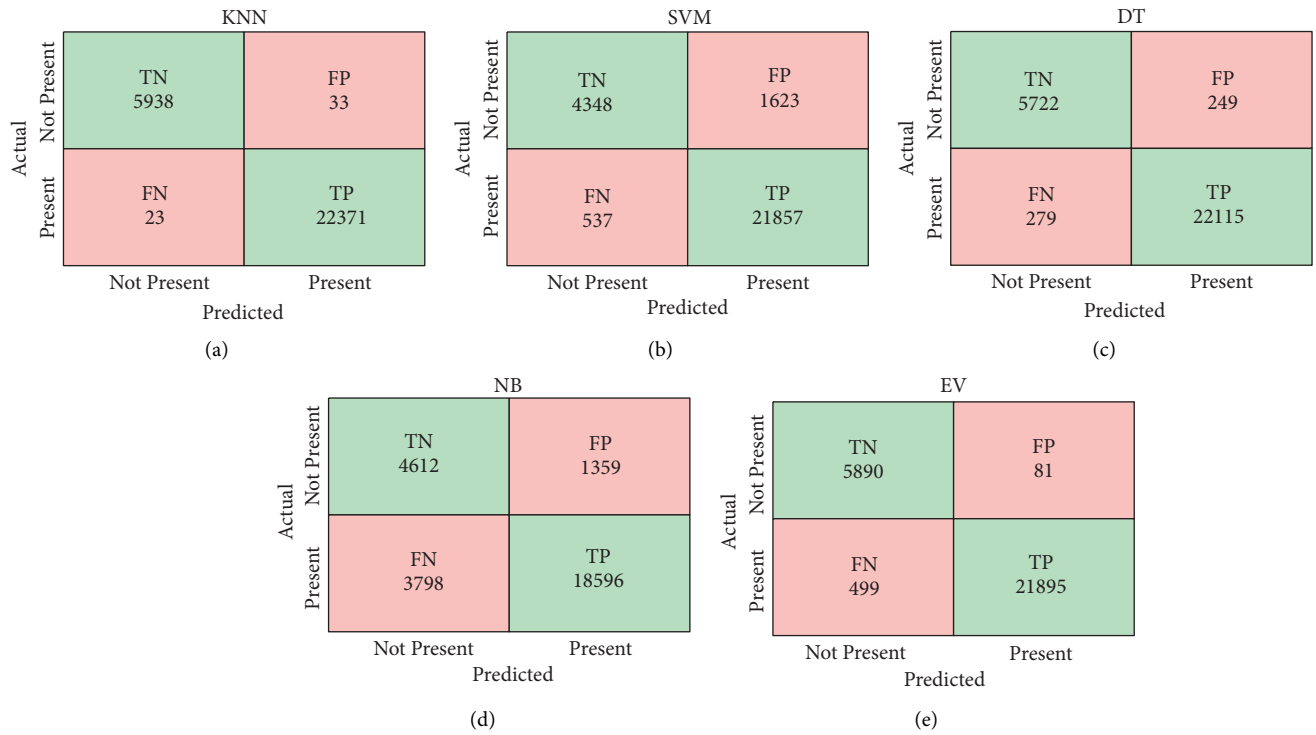


Figure 14: The confusion matrix of all classification models: (a) KNN, (b) SVM, (c) DT, (d) NB, and (e) EV.

Visualization can showcase the performance of the KNN model in predicting human presence. The visualization compares the actual time-series attendance and the predicted time-series attendance. The comparison of the two and also movement data with sensors is shown in Figure 20. The top part of the image is the time-series presence of the PIR sensor measurement results. Then, the middle part of the image is the actual attendance time series. The last part of the image below is the time series of the prediction results of the KNN model. When compared between the movement data from the PIR sensor and the presence data from the KNN model predictions, the latter is more in line with the actual presence data.

5. Discussion

In the test results, the application of MAs to the movement data of the motion sensor results can increase the PC of the features on the actual presence of humans in the room. This is in accordance with existing studies, namely, [20–23] and [24]. The related studies use MA to increase the correlation

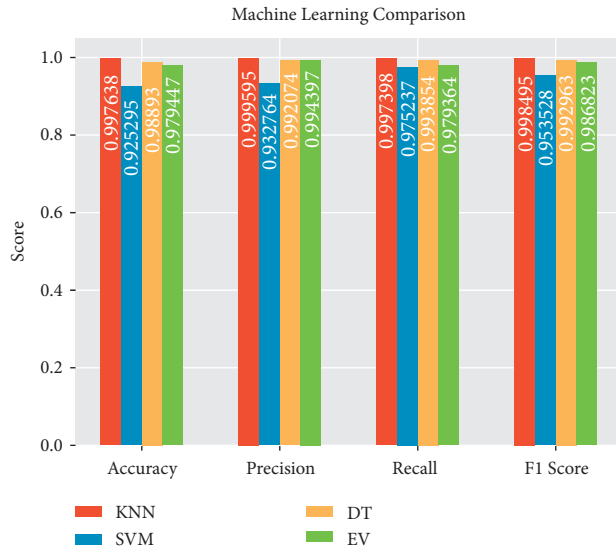


Figure 15: A performance comparison of four classification models in predicting human presence.

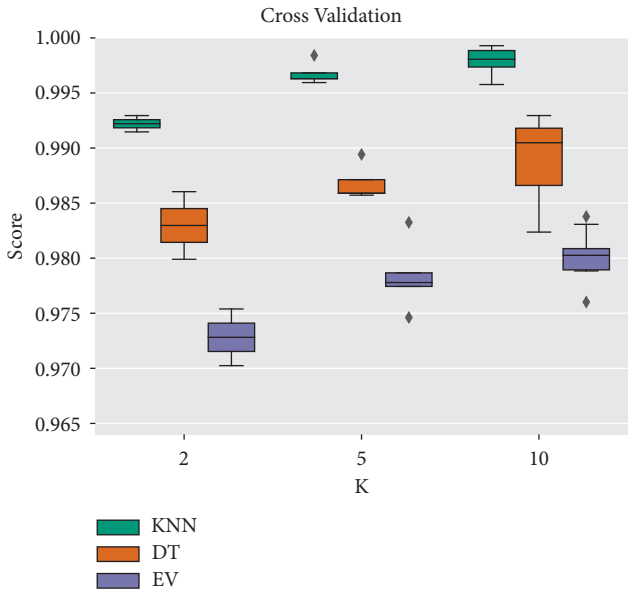


Figure 16: The cross-validation accuracy results on K-folds with $K = 2$, 5, and 10.

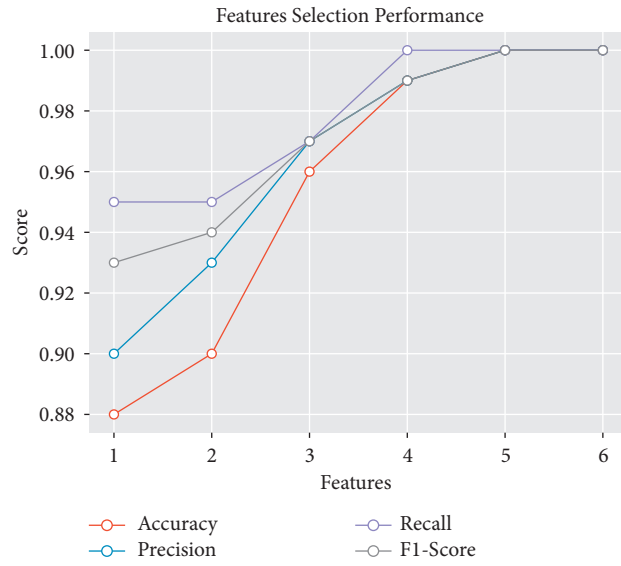


Figure 17: The increase of features on the classification model performance.

of regression and classification features, among others, for noise reduction and forecasting.

The reason why KNN can be better than SVM, DT, NB, and EV is the nature of KNN, which is robust against noisy data [68]. SVM with RBF kernel is indeed good for radially separable data. However, if the data has high variance, then it possibly affects the performance of both SVM and DT in performing data separation.

We make direct comparisons of our proposed method with related studies to emphasize the contribution and novelty of our proposed method. The related studies are human presence-based smart lighting control using other equipment. The comparison is given in Table 3. The superior values of each column are made bold. Compared to the benchmarked studies, our proposed method has the best performance, 99.8%. The research [11] has an approximate result, which is 99.3%. The study uses a concept similar to a MA, namely, a sliding window to calculate several statistical features such as mean, standard deviation, and max. A random forest RF model is an optimum model that applies the sliding window feature in the study. However, it uses several different sensors, some of which are expensive sensors, such as smartwatches and real-time location systems. Studies that also use expensive devices for activity recognition are [27], which uses a depth camera; [18], which uses five monochrome cameras; and [12], which uses Switchmate.

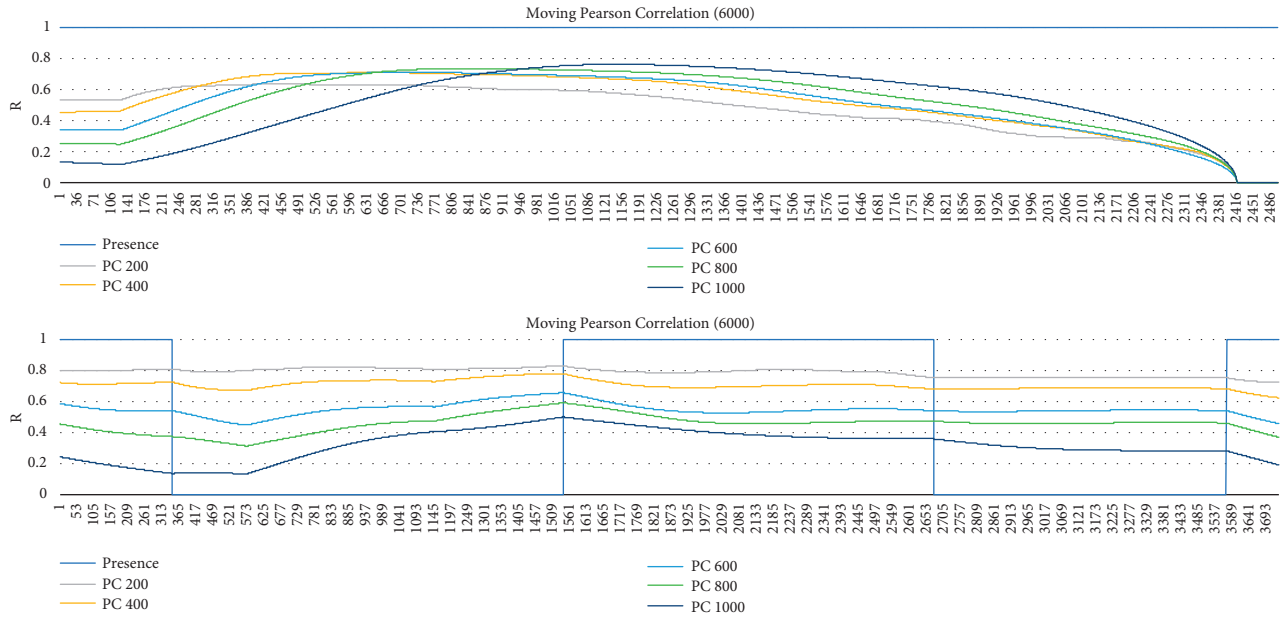


Figure 18: Visualization of the application of the moving PC ($N = 6000$) to the dataset.

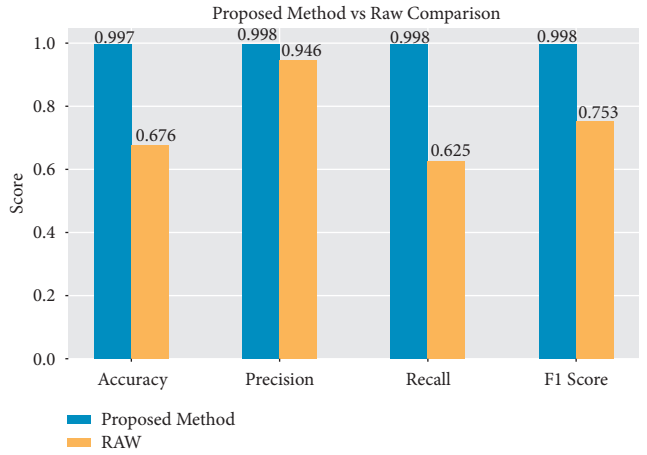


Figure 19: A performance comparison of classification with the KNN model (proposed method) with decision-making directly done on motion sensor (raw method).

The latter research does not use accuracy in calculating performance but uses LUR. LUR is their proposed metric that describes the ratio between the time the lights are on and the time someone is present in the room. We assume that LUR is equivalent to accuracy. Our proposed method is the method with the best performance and a low-cost solution.

Moreover, we also investigate the factors that influence performance in studies regarding PIR sensors in human presence-based smart lighting control. The comparison of these related works is shown in Table 4. Based on our proposed method and [9], it seems that there is a negative relationship between the number of activities and performance. However, [5] that has 14 activities has a better performance than [10, 17] that only have five activities. In addition, our proposed method and [9] are location-based methods. A person's presence is determined based on whether the person is under sensor or not. Meanwhile, [5, 10] and [17] that have significantly lower performance are not location-based. These methods define an activity where the activity is independent of its location. Further research can investigate the performance of a PIR sensor-based activity recognition on determining activities that are location-based.

In future work, the direction of this research is to increase user comfort from smart lighting. Hence, if the automatic light control is carried out based on the presence of people, people will not feel discomfort. For the long term, the research aims to measure user comfort when users use smart lighting that has applied the novel method in this research. The user comfort method proposal is a novel one, which is a quantitative method.

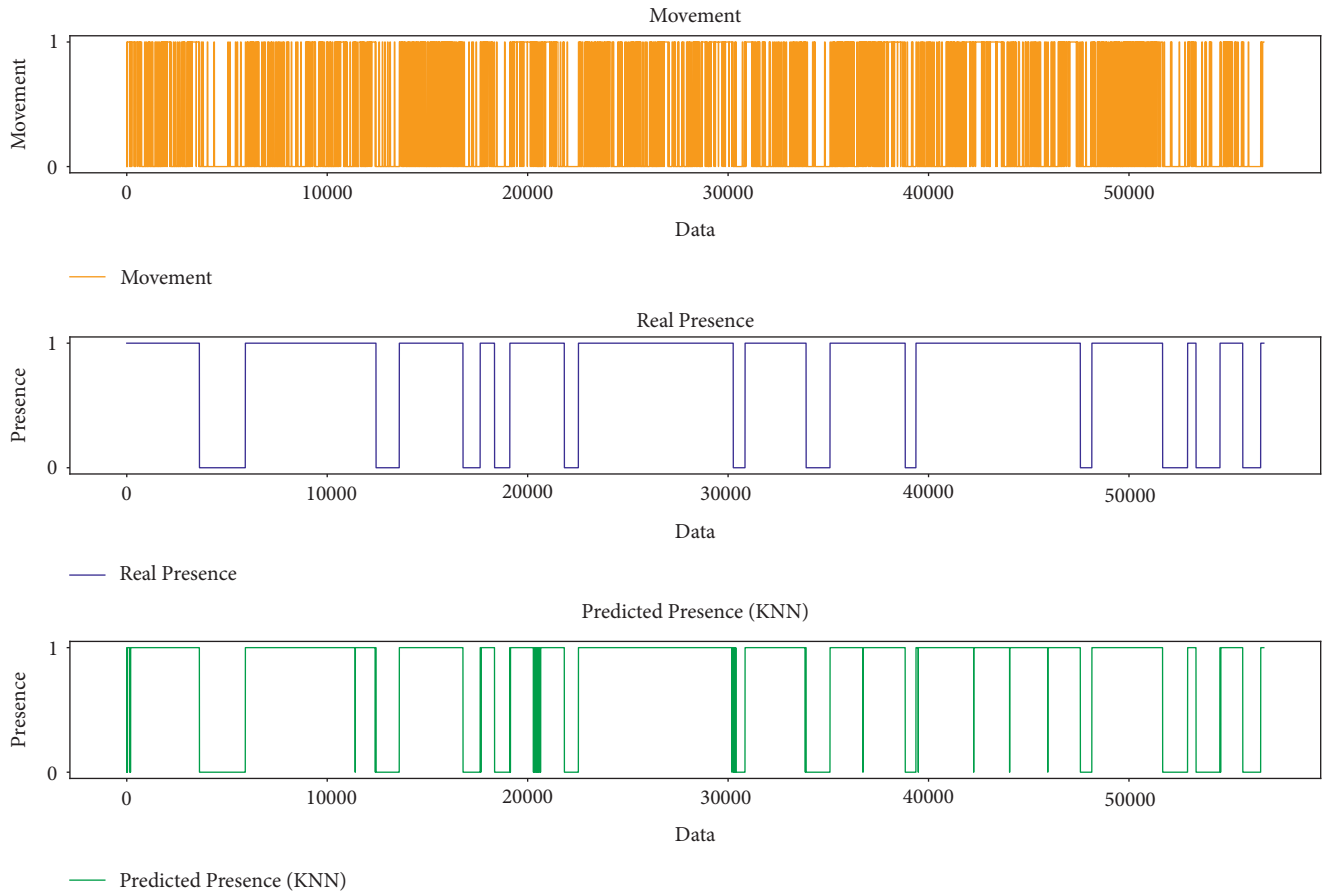


Figure 20: A time-series comparison of the movement data from the PIR sensor, the real presence, and the predicted presence by the KNN classification model.

Table 3: A Comparison of related works on human presence-based smart lighting control.

Reference	Equipment	Cost	Accuracy (%)
Proposed method	PIR sensor	**US$13**	**99.8**
Lupion et al. [11]	PIR sensor, pressure sensor, switch sensor, smartwatches, RTLS	US$65	99.3
Park et al. [12]	Light sensor, switchmate	US$68	67
Dai et al. [18]	Five monochrome cameras	US$300[1]	90.2
Chun et al. [19]	Depth camera	US$36[1]	78.3

[1]Estimated.

Table 4: A comparison of related studies using the PIR sensor for smart lighting control.

Reference	Method	Number of activities	Location-based	Accuracy (%)
Proposed method	CIMA	2	Yes	**99.8**
Ramadhan et al. [5]	HHMM	14	No	93
Jin et al. [9]	TS-ANN	2	Yes	97
Fakhruddin et al. [10]	PCA-KNN	5	No	94
Putrada et al. [17]	HHMM	5	No	87.6

Furthermore, the next future work is to use this novel method to monitor the movement of people in the house. This achievement leads to a novel predictive control of lights based on the user movement. The benefit of this proposal is that automatic light control can occur without the user being aware of it. As an illustration of the case, before people enter the room, the lights are already on. It will furtherly increase user comfort while still maintaining energy efficiency.

6. Conclusions

This paper proposes CIMA, a novel classification-integrated moving average model for smart lighting intelligent control based on human presence. A smart lighting system based on the Internet of things (IoT) applies the proposed method. It uses passive infrared (PIR) sensors, light-emitting diode (LED) lights, relays, NodeMCU, Raspberry Pi, and supporting software. In the PC test, the movement data from the PIR sensor has a correlation of 0.36 to attendance, while the moving average (MA) correlation to human presence can reach 0.56. In exhaustive testing of machine learning classification methods, k-nearest neighbor (KNN) is the model with the best and most robust performance with an accuracy value of 99.8%. It is more accurate than direct light control decisions based on motion sensors with 67.6%. We conclude that our proposed method can increase the correlation value of movement features on attendance. At the same time, an accurate and robust KNN classification model is applicable for human presence-based smart lighting intelligent control.

Data Availability

Data supporting reported results can be found at https://doi.org/10.34820/FK2/8BXAYW.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

All authors contributed equally to this manuscript.

Acknowledgments

The authors would like to thank Telkom University and the Ministry of Education, Culture, Research, and Technology for fully funding this research through the Doctoral Dissertation Research scheme and other supporting funds.

References

[1] M. Soheilian, G. Fischl, and M. Aries, "Smart lighting application for energy saving and user well-being in the residential environment," *Sustainability*, vol. 13, no. 11, p. 6198, 2021.

[2] P. Smallwood, *Lighting, leds and smart lighting market overview*, US Dept. Energy SSL Workshop, Raleigh, NC, USA, 2016.

[3] M. Füchtenhans, E. H. Grosse, and C. H. Glock, "Smart lighting systems: state-of-the-art and potential applications in warehouse order picking," *International Journal of Production Research*, vol. 59, no. 12, pp. 3817–3839, 2021.

[4] O. O. Ordaz-García, M. Ortiz-Lopez, F. J. Quiles-Latorre, J. G. Arceo-Olague, R. Solis-Robles, and F. J. Bellido-Outeirino, "DALI bridge FPGA-based implementation in a wireless sensor Node for IoT street lighting applications," *Electronics*, vol. 9, no. 11, p. 1803, 2020, https://www.mdpi.com/2079-9292/9/11/1803.

[5] R. Nur Ghaniaviyanto, "Aji Gautama Putrada, and Maman Abdurohman. Improving smart lighting with activity recognition using hierarchical hidden Markov model," *Indonesia Journal on Computing (Indo-JC)*, vol. 4, no. 2, pp. 43–54, 2019.

[6] L. Gyu Myoung and J. Yun Kim, "The internet of things—a problem statement," in *Proceedings of the 2010 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), November 2010.

[7] J. Guo, Y. Mu, M. Xiong, Y. Liu, and J. Gu, "Activity feature solving based on tf-idf for activity recognition in smart homes," *Complexity*, vol. 2019, pp. 1–10, 2019.

[8] M. José, E. Irigoyen, and V. M. Becerra, "Intelligent control approaches for modeling and control of complex systems," *Complexity*, vol. 2018, pp. 1-2, 2018.

[9] Y. Jin, D. Yan, X. Zhang, J. An, and M. Han, "A data-driven model predictive control for lighting system based on historical occupancy in an office building: methodology development," *In Building Simulation*, vol. 14, pp. 219–235, 2021.

[10] R. Irsyad Fakhruddin, "Maman Abdurohman, and Aji Gautama Putrada. Improving pir sensor network-based activity recognition with pca and knn," in *Proceedings of the 2021 International Conference On Intelligent Cybernetics Technology & Applications (ICICyTA), Pages 138–143* IEEE, Bandung, Indonesia, December 2021.

[11] M. Lupión, P. M. Ortigosa, Q. Javier Medina, J. Medina-Quero, and J. F. Sanjuan, "Dolars, a distributed on-line activity recognition system by means of heterogeneous sensors in real-life deployments—a case study in the smart lab of the university of almería," *Sensors*, vol. 21, no. 2, p. 405, 2021.

[12] J. Y. Park, T. Dougherty, H. Fritz, and Z. Nagy, "LightLearn: an adaptive and occupant centered controller for lighting based on reinforcement learning," *Building and Environment*, vol. 147, pp. 397–414, 2019.

[13] J. Lei, X. Wang, Y. Zhang, L. Zhu, and L. Zhang, "Policy and law assessment of covid-19 based on smooth transition autoregressive model," *Complexity*, vol. 2021, Article ID 6659117, 13 pages, 2021.

[14] L. Borzì, S. Fornara, F. Amato, G. Olmo, C. A. Artusi, and L. Lopiano, "Smartphone-based evaluation of postural stability in Parkinson's disease patients during quiet stance," *Electronics*, vol. 9, no. 6, p. 919, 2020, https://www.mdpi.com/2079-9292/9/6/919.

[15] H. Layth Rafea, "Four classification methods naïve bayesian, support vector machine, k-nearest neighbors and random forest are tested for credit card fraud detection," *Master's thesis, Altınbaş Üniversitesi*, 2018.

[16] K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," in *U-healthcare Monitoring Systems,* vol. 1, pp. 179–196, Elsevier, 2019.

[17] P. Aji Gautama, "Nur Ghaniaviyanto Ramadhan, and MA Makky. An evaluation of activity recognition with hierarchical hidden Markov model and other methods for smart lighting in office buildings," *ICIC International*, vol. 16, 2022.

[18] Ji Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, "Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 68–76, Boston, 2015.

[19] S. Y. Chun, C.-S. Lee, and J.-S. Jang, "Real-time smart lighting control using human motion tracking from depth camera," *Journal of Real-Time Image Processing*, vol. 10, no. 4, pp. 805–820, 2015.

[20] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between google trends on dengue fever and national surveillance

report in Indonesia," *Global Health Action*, vol. 12, no. 1, Article ID 1552652, 2019.

[21] Z. Hu, Y. Zhang, Y. Zhao et al., "A water quality prediction method based on the deep lstm network considering correlation in smart mariculture," *Sensors*, vol. 19, no. 6, p. 1420, 2019.

[22] Yu Peng, S. Long, J. Ma, J. Song, and Z. Liu, "Temporal-spatial variability in correlations of drought and flood during recent 500 years in inner Mongolia, China," *Science of the Total Environment*, vol. 633, pp. 484–491, 2018.

[23] H. S. Badr, E. Dong, M. M. Squire et al., "Association between mobility patterns and covid-19 transmission in the USA: a mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1247–1254, 2020.

[24] O. P. Singh, T. A. Howe, and M. B. Malarvili, "Real-time human respiration carbon dioxide measurement device for cardiorespiratory assessment," *Journal of Breath Research*, vol. 12, no. 2, 2018.

[25] P. Aji Gautama, "Maman Abdurohman, Doan Perdana, and Hilal Hudan Nuha. Machine learning methods in smart lighting towards achieving user comfort: a survey," *IEEE Access*, vol. 10, 2022.

[26] P. Aji Gautama and A. Maman, "Anomaly detection on an iot-based vaccine storage refrigerator temperature monitoring system," in *Proceedings of the 2021 International Conference On Intelligent Cybernetics Technology & Applications (ICICyTA)*, pp. 75–80, IEEE, Bandung, Indonesia, December 2021.

[27] S. Chun and C.-S. Lee, "Applications of human motion tracking: smart lighting control," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 387–392, Portland, 2013.

[28] S. Fuada, T. Adiono, and L. Siregar, "Internet-of-things for smart street lighting system using esp8266 on mesh network,".

[29] T. Montanaro, I. Sergi, A. Motroni et al., "An iot-aware smart system exploiting the electromagnetic behavior of uhf-rfid tags to improve worker safety in outdoor environments," *Electronics*, vol. 11, no. 5, p. 717, 2022, https://www.mdpi.com/2079-9292/11/5/717.

[30] P. Kumar, P. Rai, and H. B. Yadav, "Smart lighting and switching using Internet of Things," in *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pages 536–539*, IEEE, Noida, India, January 2021, https://ieeexplore.ieee.org/document/9377078/.

[31] E. Juškevičius, "Smart home lighting system using iot technologies," Bachelor Thesis, South Eastern Finland University of Applied Science, Kouvola, 2021.

[32] J. Purmaissur, A. Seeam, S. Guness, and X. Bellekens, "Augmented reality intelligent lighting smart spaces," in *Proceedings of the 2019 Conference On Next Generation Computing Applications (NextComp)*, pp. 1–5, IEEE, Mauritius, September 2019.

[33] P. Aji Gautama and P. Doan, "Improving thermal Camera Performance in Fever Detection during Covid-19 Protocol with Random forest Classification," in *Proceedings of the 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, pp. 1–6, IEEE, Bali, Indonesia, October 2021.

[34] G. S. Smrithy, R. Balakrishnan, and N. Sivakumar, "Anomaly detection using dynamic sliding window in wireless body area networks," *Data Science and Big Data Analytics*, Springer, pp. 99–108, Singapore.

[35] G. Kechyn, L. Yu, Y. Zang, and S. Kechyn, "Sales forecasting using wavenet within the framework of the kaggle competition," 2018, https://arxiv.org/abs/1803.04037.

[36] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 599–606, 2020.

[37] T. Duong Truong, "Nguyen Ba Hoang Quan, and Nopadon Maneetien. Implementation of Moving Average Filter on Stm32f4 for Vibration Sensor Application," in *Proceedings of the 2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 627–631, IEEE, Ho Chi Minh City, Vietnam, November 2018.

[38] F. Ghassani, M. Abdurohman, and A. G. Putrada, "Prediction of Smarthphone Charging Using K-Nearest Neighbor Machine Learning," in *Proceedings of the 2018 Third International Conference on Informatics and Computing (ICIC)*, pp. 1–4, IEEE, Palembang, Indonesia, October 2018.

[39] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803–855, 2019.

[40] G. A. Rattá, J. Vega, A. Murari, and J. Efda, "Improved feature selection based on genetic algorithms for real time disruption prediction on jet," *Fusion Engineering and Design*, vol. 87, no. 9, pp. 1670–1678, 2012.

[41] M. Gelfusa, A. Murari, M. Lungaroni et al., "A support vector machine approach to the automatic identification of fluorescence spectra emitted by biological agents. In Optics and Photonics for Counterterrorism," *Crime Fighting, and Defence XII*, vol. 9995, 2016.

[42] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in *Proceedings of the 2019 International Conference On Intelligent Sustainable Systems (ICISS)*, pp. 24–28, IEEE, Palladam, India, February 2019.

[43] A. Taufiqurrahman, A. Gautama Putrada, and F. Dawani, "Decision tree regression with adaboost ensemble learning for water temperature forecasting in aquaponic ecosystem," in *Proceedings of the 2020 6th International Conference On Interactive Digital Media (ICIDM)*, pp. 1–5, IEEE, Bandung, Indonesia, December 2020.

[44] A. N. Iman, A. G. Putrada, S. Prabowo, and D. Perdana, "Peningkatan ktmrpp," *Jurnal Elektro dan Telekomunikasi Terapan*, vol. 8, no. 1, pp. 978–985, 2021.

[45] K. Yusuf, M. Abdurohman, and A. G. Putrada, "Increasing Passive Rfid-Based Smart Shopping Cart Performance Using Decision Tree," in *Proceedings of the 2019 5th International Conference on Computing Engineering and Design (ICCED)*, pp. 1–5, IEEE, Singapore, April 2019.

[46] H. Bagaskara, A. Gautama Putrada, and E. Ariyanto, "Proximity and dynamic device pairing based authentication for iot end devices with decision tree method," in *Proceedings of the 2020 6th International Conference On Interactive Digital Media (ICIDM)*, pp. 1–5, IEEE, Bandung, Indonesia, December 2020.

[47] T. Daniya, M. Geetha, and K. Suresh Kumar, "Classification and regression trees with gini index," *Advances in Mathematics: Scientific Journal*, vol. 9, no. 10, pp. 8237–8247, 2020.

[48] D. Berrar, "Bayes' theorem and naive bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics," *Elsevier Science Publisher: amsterdam, The Netherlands*, vol. 403, 2018.

[49] O. Sagi and L. Rokach, "Ensemble learning: a survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Article ID e1249, 2018.

[50] F. T. Breiner, M. P. Nobis, A. Bergamini, and A. Guisan, "Optimizing ensembles of small models for predicting the distribution of species with few occurrences," *Methods in Ecology and Evolution*, vol. 9, no. 4, pp. 802–808, 2018.

[51] S. Mani, S. Kumari, A. Jain, and P. Kumar, "Spam review detection using ensemble machine learning," in *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 198–209, Springer, Cham, July 2018.

[52] B. H. Al-Zadid Sultan and T. Tanpia, "An ensemble hard voting model for cardiovascular disease prediction," in *Proceedings of the 2020 2nd International Conference On Sustainable Technologies For Industry 4.0 (STI)*IEEE, Dhaka, Bangladesh, December 2020.

[53] O. Esezi Isaac and C. A. Eric, "Test for significance of pearson's correlation coefficient," *International Journal of Innovative Mathematics, Statistics & Energy Policies*, vol. 6, no. 1, pp. 11–23, 2018.

[54] S. Muhammad Bagus, "Aji Gautama Putrada, and Maman Abdurohman. Evaluation of face detection and recognition methods in smart mirror implementation," in *Proceedings of Sixth International Congress on Information and Communication Technology*, pp. 449–457, Springer, Singapore, September 2022.

[55] Y. Demir, N. Ö. Atar, Ü. Güzelküçük, K. Aydemir, and E. Yaşar, "The use of and satisfaction with prosthesis and quality of life in patients with combat related lower limb amputation, experience of a tertiary referral amputee clinic in Turkey," *Age*, vol. 61, no. 1, pp. 6–10, 2019.

[56] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo, "Daily activity feature selection in smart homes based on pearson correlation coefficient," *Neural Processing Letters*, vol. 51, no. 2, pp. 1771–1787, 2020.

[57] M. Belkin, D. J. Hsu, and P. Mitra, "Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[58] B. Ghojogh and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial," 2019, https://arxiv.org/abs/1905.12787.

[59] F. Khan, A. Urooj, K. Ullah, B. Alnssyan, and Z. Almaspoor, "A comparison of autometrics and penalization techniques under various error distributions: evidence from Monte Carlo simulation," *Complexity*, vol. 2021, Article ID 9223763, 8 pages, 2021.

[60] Y. Gil, J. Honaker, S. Gupta et al., "Towards human-guided machine learning," in *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 614–624, California, Marina del Ray, March 2019.

[61] K. Asai, T. Fukusato, and T. Igarashi, "An interactive tool for feature analysis of outliers in multi-dimensional data," Research Gate, 2018.

[62] G. Lovell, "Unified model combining the boundary conditions encountered at the input, the carrier optical fiber and the output in spatially multiplexed optical communication channels," *PhD thesis, Florida institute of Technology*, Melbourne, 2019.

[63] T.-Yi Chen, Y.-H. Chang, M.-C. Yang, and H.-W. Chen, "How to cultivate a green decision tree without loss of accuracy," in *Proceedings of the ACM/IEEE international symposium on low power electronics and design*, pp. 1–6, Massachusetts, Boston, August 2020.

[64] S. He, Y. He, and M. Li, "Classification of illegal activities on the dark web," in *Proceedings of the 2019 2nd international conference on information science and systems*, pp. 73–78, Tokyo, Japan, March 2019.

[65] O. Cherqi, G. Mezzour, M. Ghogho, and M. E. Koutbi, "Analysis of Hacking Related Trade in the Darkweb," in *Proceedings of the 2018 IEEE international conference on intelligence and security informatics (ISI)*, pp. 79–84, IEEE, Miami, FL, USA, November 2018.

[66] E. Victor, J. Staartjes, M. Kernbach et al., "Foundations of feature selection in clinical prediction modeling," in *Proceedings of the Machine Learning in Clinical Neuroscience*, pp. 51–57, Springer, Cham, 2022.

[67] B. Shi, B. Meng, H. Yang, J. Wang, and W. Shi, "A novel approach for reducing attributes and its application to small enterprise financing ability evaluation," *Complexity*, vol. 2018, Article ID 1032643, 17 pages, 2018.

[68] S. Alfin Pratama, "Kusuma Ayu Laksitowening, and Ibnu Asror. Time series prediction on college graduation using knn algorithm," in *Proceedings of the 2020 8th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–4, IEEE, Yogyakarta, Indonesia, June 2020.

Hindawi
Complexity
Volume 2022, Article ID 8670154, 13 pages
https://doi.org/10.1155/2022/8670154

Research Article

Simultaneous Process Mining of Process Events and Operator Actions for Alarm Management

László Bántay,[1] **Gyula Dörgö**,[1] **Ferenc Tandari**,[2] **and János Abonyi**[1]

[1]*ELKH-PE Complex Systems Monitoring Research Group, University of Pannonia, Veszprém H-8200, Hungary*
[2]*MOL Danube Refinery, Százhalombatta H-2443, Hungary*

Correspondence should be addressed to János Abonyi; janos@abonyilab.com

Received 1 June 2022; Accepted 1 August 2022; Published 19 September 2022

Academic Editor: Andrea Murari

Alarm management is an important task to ensure the safety of industrial process technologies. A well-designed alarm system can reduce the workload of operators parallel with the support of the production, which is in line with the approach of Industry 5.0. Using Process Mining tools to explore the operator-related event scenarios requires a goal-oriented log file format that contains the start and the end of the alarms along with the triggered operator actions. The key contribution of the work is that a method is presented that transforms the historical event data of control systems into goal-oriented log files used as inputs of process mining algorithms. The applicability of the proposed process mining-based method is presented concerning the analysis of a hydrofluoric acid alkylation plant. The detailed application examples illustrate how the extracted process models can be interpreted and utilized. The results confirm that applying the tools of process mining in alarm management requires a goal-oriented log-file design.

1. Introduction

The motivation of the present work is to develop a methodology for the process mining-based analysis of alarm and event-log databases to increase process safety and reduce the workload of the operators. As a result, we will be able to understand the chain of events that trigger an operator action, as well as explore the effects of different operator action strategies; from another point of view, to gain the models of processes, leading potentially to malfunctions or safety incidents. With the help of this knowledge, a better designed and more effective alarm management [1] and operator training system can be developed.

The Industry 5.0 approach considers the wellbeing of the workers in productivity and efficiency improvement projects more. State-of-the-art industrial production systems contain complex process control solutions. The amount of recorded signals and process variables makes it difficult to have a clear view of the relationships between the different process elements; the control of the processes can be a demanding task for the workers. To lower the workload of the operators, a good understanding of the process element relationships is needed to predict the probable event scenarios that can be the basis of a decision-supporting system. The work of the operators can be reduced and supported in other ways as well. Generally speaking, predictable alarms do not contain useful information for the operators. Therefore, automation solutions should handle them before they occur, or completely useless alarms should be suppressed before an announcement is made [2]. Future alarm sequences can also be predicted using historical knowledge of the process [3]. The exploration of operational strategies holds promising opportunities for automation as well: the sequences of alarms and the corresponding operator actions can be determined besides the best operational practices [4]. The analysis of the announced alarms can also facilitate root cause analysis [5]. A wide range of data-based solutions have been applied to reduce the operator workload, e.g., the conventional techniques of deadbands [6], delay timers [7], or filtering [8]. Advanced alarm management solutions aim to define more informative features for the operators by identifying redundant and co-occurring alarms. Two main approaches are

known. First, correlation analysis-based techniques are widespread, where the aim is to find frequently co-occurring alarms over a short period of time, which can be considered redundant [9]. Second, the frequently occurring longer operational patterns can be considered to be the symptoms of the same malfunction and can be revealed by frequent pattern mining algorithms [10], as well as applied in terms of alarm prediction and suppression [2]. It is also advantageous to apply highly efficient data-driven solutions, like deep learning [11] or decision tree-based classifiers [12]. To gain understandable models that can be directly used in alarm management is challenging. To explore complex event chains or comparable models of operator action strategies, determining the correlation of event pairs, deep learning methods with hard-to-understand results are not satisfying. For compact and comprehensible models, we need to apply process mining.

Process mining is a collection of techniques that support the understanding of processes based on event logs [13]. Process mining algorithms are applied in various fields, without the aim of completeness: in healthcare to improve the operational efficiency of processes [14], in business management to analyze the processes and reveal their bottlenecks [15], for the automatized analysis of financial statements during audit processes [16], or the support of e-learning are among the applications as well [17]. Moreover, a recent and highly promising application field is the identification of repetitive processes using process mining to discover potential processes for robotic process automation [18]. For a recent collection of research fields and application areas in process mining, please refer to the work of Garcia et al. [19]. These process mining algorithms are tailored to goals for the purpose of discovering process flows, where the events of the different processes are organized into traces. A trace is a collection of events that are considered to be related in some way. The definition of the traces is essential for the purpose of process discovery, especially in the case of alarm management, where the connection between alarms and operator actions requires accurate trace generation. Therefore, the structures of alarm and event logs are unsatisfactory in terms of process mining as they lack this very important component, that is, trace indicators. A process mining-based approach was applied to evaluate the behavior of plants and support the rationalization of alarms. Its basic concepts are presented in Reference [20], while the applicability of process mining algorithms for the determination of alarm performance metrics and the exploration of process behavior are presented in Reference [21], which can be considered to be the motivation of the present study.

Although these studies usually focus on different aspects of alarm management, e.g., operator actions [4] or alarm rationalization [21], they lack the general formalization of the problem and the goal-oriented definition of the input database for different purposes of analysis. Given the lack of a comprehensive study on the applicability of the techniques of process mining with regard to large-scale industrial alarm and event log databases, the contribution of the present work is the following:

(i) We discuss the goal-oriented tasks and the related definition of the events, as well as further characteristics of the events and resources. Traces must be defined to provide a suitable structure of alarms and the related operator actions, resulting in the most appropriate input dataset for the applied mining algorithms.

(ii) The effectiveness and applicability of the proposed methodology are presented with regard to the analysis of the large-scale alarm and event-log database of an industrial plant. We have gained understandable and comprehensive information that is useful not only in alarm management and operator training systems but also in the process mining of other industrial sectors.

According to the contributions, the core applicability of the proposed methodology is not narrowed down to the industrial alarm systems but can be transferred to other fields of applications as well, where temporal events are present and their follow-up or cause and effect type of relationship is to be analyzed (similarly to the back and forth relationship of alarms and operator actions). In the case of alarm systems, the type of process control system, let it be a distributed control system (DCS), supervisory control and data acquisition system (SCADA), or any other type of system, is irrelevant, as long as the provided alarm (or event) data are timely and accurate. Using the proposed methodology, the alarm evolution paths can be identified, the triggering alarms of operator actions can be revealed, and recommendations for the reduction of the probability of hazardous situations can be provided.

The structure of the work is the following. In Section Materials and Methods, we provide a brief overview of industrial log files (structure and content), introduce the goal-oriented definition of the traces, identify the necessary rules to generate traces, and discuss the process mining tools necessary to achieve our goals. In Section Results and Discussion, we examine the previously defined process mining tasks, the preparation of the log file, the time distribution analysis of the events (alarms and operator actions), the alarm spillover analysis between the production units, and the discovery of the connection between alarms and operator actions. In Section Conclusion, some concluding remarks are provided, experiences are discussed, and possible future research directions are identified.

2. Materials and Methods

This section introduces the basis of process discovery (log files), the event-clustering method (trace generation rules), and the goal-oriented selection of process exploration tools. As in the case of any project-like activity, an execution plan to achieve our predefined goals must be drawn up. In terms of the present case study, a schematic summary is provided in Figure 1.

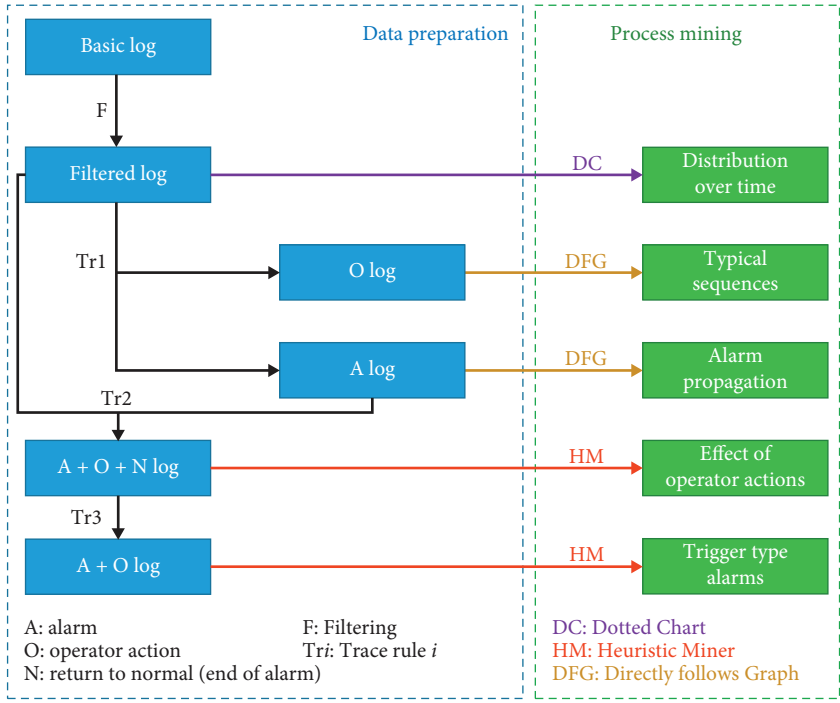


Figure 1: The concept of process mining-based alarm management. After preparing the data, we can perform our process discovery tasks on our sub-logs that were generated by applying the suggested trace rules.

2.1. The Structure of the Alarm and Event Log Databases. Discrete events, e.g., alarms and warnings, are recorded in the process control unit of almost every production site when a variable exceeds its associated threshold. An industrial alarm and event-log database is usually composed of these alarms and warnings, operator actions, system messages, as well as any further timestamped information logged by the control system. For the purpose of further mathematical formulation, every event can be regarded as a state of the technology (denoted by s) represented by $<\text{pv}, a>$ data pairs. pv indicates the name of the process variable, while a represents the attribute that indicates the state occurring on the given variable, e.g., high or low alarm in the case of an alarm announcement or increase or decrease (open or close) in the case of operator action. An *event* is the occurrence of a given state, which can be an event of temporal duration such as alarm messages or point-like in time such as operator actions. For example, the description of an alarm and an operator action can be mathematically represented as $<$column inlet temperature, high alarm$>$ and $<$cooling water inlet valve, open$>$, respectively. The events are represented by $<s, st, et>$ triplets, where s denotes the state that occurs between st starting and et ending times. For an event with a point-like temporal characteristic, even though the st and et times are regarded as equal to maintain a uniform mathematical formulation, it must be decided whether to keep both or utilize only the st starting time for further processing. This question is addressed later when the XES file is considered.

The files of industrial alarm and event logs are usually composed of timestamped events of alarms, operator actions, system messages, and any further temporal information. Additional information can help us to determine which sensor generates the alarm in the process and the priority of that alarm, e.g., the location of the event, that is, in which part of the plant/organization it occurred. Each event that occurs is labeled with a tag name. Every alarm and operator action can be considered to be the state of the process. The different alarms in the event log have starting and end times. The starting time is when the alarm in the process was raised, and the end time is when the process returned to its proper operational zone. The operator actions are usually considered to be point-like events, i.e., their starting and end times coincide. A series of events can be defined as a trace. Let L denote a set of events; $\sigma \in L$ stands for an event trace, that is, a sequence of events; $T \subseteq L$ represents an event log, i.e., a set of event traces. Besides information on the occurrence of an event, the log files usually contain other information as previously discussed, e.g., the location of the sensor that raised the alarm in the process, which is possibly categorized into units or production units, etc. This work demonstrates the importance of task-specific trace definitions. The methodology of trace segmentation will be discussed later.

Our goal is to identify basic patterns in the chain of alarms to focus on frequent sequences that can help us compile a prediction model of the alarms. In Table 1, an example of an industrial alarm and event-log can be seen. The column labeled "Tag" has been added to support the analysis and is a summarized representation of the sensor's name. The first part is the name of the tag, the second and third are those of the unit and production unit, respectively, while the last one is the type of event (A: alarm, O: operator action, N: return to normal). Different "sub-logs" are

Table 1: An example of an industrial alarm and event log file.

ID	Tag name	Event	Start time	End time	Unit	Prod. unit	Tag
1	321	A	2018.05.01 00:01:01	2018.05.01 00:03:08	3	1	321_3_1_A
2	632	A	2018.05.01 00:01:03	2018.05.01 00:10:06	4	5	632_4_5_A
3	421	A	2018.05.01 00:01:10	2018.05.01 00:05:10	2	5	421_2_5_A
4	312	A	2018.05.01 00:01:30	2018.05.01 00:03:08	3	1	312_3_1_A
5	321	O	2018.05.01 00:02:10	2018.05.01 00:02:10	3	1	321_3_1_O
7	321	N	2018.05.01 00:03:08	2018.05.01 00:03:08	3	1	321_3_1_N
8	421	O	2018.05.01 00:04:01	2018.05.01 00:04:01	2	5	421_2_5_O
10	632	N	2018.05.01 00:10:06	2018.05.01 00:10:06	4	5	632_4_5_N

Table 2: Task oriented event types in sub-logs and suggested tools to process (A: alarm, O: operator action, N: return to normal, PM: process mining).

Object	Task	Event types	Tool
Basic log	Data preparation for PM	A, N, O	Filtering and XES generator
Alarms	Distribution over time	A	Dotted chart
Alarms	Typical processes	A	Directly-follows graph
Alarms	Spillover among units	A	Directly-follows graph/Heuristic miner
Alarms	Trigger type events	A, O	Heuristic miner
Operator actions	Distribution over time	O	Dotted chart
Operator actions	Typical processes	O	Directly-follows graph
Operator actions	Effect of operator actions	A, N, O	Heuristic miner

necessary to examine various mining tasks. The required types of events are summarized in Table 2. The object of the analysis indicates what information is of interest with regard to the analysis provided in the task column. The event types indicate the types of events available for analysis. Finally, the suggested tools to process are mentioned. The three types of objects are the following:

(i) Basic log: all three types of events are needed, after a filtering/cleaning step, the log file has to be put into a standardized format. This format is XES (Section 2.2).

(ii) Alarms: dotted chart, directly-follows graph, and heuristic miner are proper tools to analyze the different aspects of alarm propagation. By adding operator actions to the traces, their triggering alarms can be identified.

(iii) Operator actions: the tools and tasks are more or less the same as in the case of alarms. The most complex task is to explore the effect of the operator actions, this needs all three types of events, as alarms are the trigger events and return to normal events are the consequences of operator actions.

There are two kinds of tools to process the data, the preparation type and the ones responsible for the Process Mining part itself. The filtering/cleaning step is a standard data processing task; it can be tailored and automatized by using the Python programming language, as well as its transformation to XES standard. The three Process Mining tools are also available in Python. This way, it is quite easy to develop an integrated solution tailored for the actual purpose.

2.1.1. Dotted Chart Analysis. The most transparent method by which to visualize the event log is the dotted chart analysis. In these charts, a dot represents a single event in the log with two orthogonal dimensions, namely, time and component types. Component types like instance, originator, task, event type, or data elements are shown on the vertical axis. Time is measured on the horizontal axis of the chart. Many measures related to events can be determined, such as the average number of events occurring over a certain time period, the maximum number of events in that time period, the maximum and minimum time interval between events, etc. Time can be presented factually or relatively. The relative time can be used to abstract the log file. For every component type, the first event is positioned at time 0 and subsequent events are placed relative to the time of occurrence of the first event. Moreover, the shape and color of the dots can be changed depending on the examined event attributes, adding dimensions to our chart.

From a chart like this, a lot of useful information can be obtained, e.g., where the alarms occur more frequently or which production units are affected more by alarm events.

2.1.2. Directly-Follows Graph. On a directly-follows graph, an edge is represented between two nodes when at least one trace where the target event follows the source event is present. In a nutshell, this method by which a DFG is obtained ([22]):

(i) defines 3 parameters, namely, τ_{Var}, τ_{act}, τ_{df}

(ii) removes cases with a frequency lower than τ_{Var} from the log

(iii) removes events with a frequency lower than τ_{act} from the filtered log and adds a node for each of the remaining activities

(iv) connects the nodes where 2 activities follow on from each other at least τ_{df} times

On these graphs, two metrics can be represented, namely, frequency (the number of times the target event follows on from the source event) and performance (the average time elapsed between the source and target events), the decision depends on the type of required information.

2.1.3. Heuristic Miner Algorithm. The heuristic miner algorithm explores the control-flow perspective of the process model. The log is analyzed for the presence of causal dependencies. If an event is always followed by another event, a dependency relationship probably exists between these events. The log should be analyzed for these causal dependencies. The advantage compared to α-miner is that the heuristic miner algorithm considers frequencies and can handle skipping activities [13]. Several parameters can be adjusted, e.g., minimum activity count. Events that occur under this threshold are not shown on the nets, which can be a Heuristic net or a Petri net. Another parameter is the minimum DFG occurrences, which is the minimum number of occurrences of an edge to be considered. This attribute shows that the heuristic miner is based on DFG. Heuristic mining requires a clear starting and end event, assuming that every activity is located on a path from the starting activity to the end activity. As is the case in DFG, frequency and performance parameters can be entered on the net.

2.2. Goal-Oriented Definition of the Traces. The input of process mining tools is a log file in a standard format; in this work, we have chosen the XES standard. XES is an XML-based standard for event logs. Its purpose is to provide a generally acknowledged format for exchanging event log data between tools, applications, and domains. The main reason for choosing the XES model is the support it provides for traces. As mentioned above, it is very important that traces are well defined. Usually, in industrial log files, the events are sequenced independently of one another, unlike in the XES standard. The majority of process mining algorithms take into consideration traces in addition to events.

For the effective mining of the alarm and event-log files, a definition of the time window applied for the segmentation of the alarm and event log files into traces that is dependent on the purpose is required.

As previously mentioned, since traces play a significant role in process mining methods, the task-dependent determination of trace-defining time windows is required. As the main event of the alarm log file is the appearance of the alarm event itself (that is, its starting time), the basis for the trace generation is the following:

Let α and β denote two consecutive events in log L. Let $T(\alpha)$ and $T(\beta)$ stand for the times of occurrence of events α and β, respectively, and σ represents the spillover constant. If $T(\beta) - T(\alpha) < \sigma$, then α and β are located on the same trace. However, if $T(\beta) - T(\alpha) \geq \sigma$, then β is placed in the following trace. The spillover constant can be tuned based on the knowledge about the dynamics of the system. Obviously, it can be calculated with the help of the experience of the operators and identification of data driven dynamical models (Figure 2).

There are three areas to discuss the rules concerning trace generation, the analysis of alarms, operator actions, and their relations.

2.2.1. Analysis of Alarms. The first concept to explore in an alarm management system is the spillover effect of the alarms. These sequences of alarms are caused primarily by the decline in product streams, as well as the spread of pressure anomalies or attributes (temperature and concentration) connected to technology streams. According to this, probable propagation times are related to the sojourn times of equipment, the length of pipelines, and the logic of the control system. Hazard and operability analysis (HAZOP) provides options to explore this malfunction propagation, in addition, more and more attention is being paid to dynamic HAZOP. As its automated use is challenging [23] and these methods are very resource-demanding (require expert engineers), it would be beneficial to explore these potential relationships automatically from the log files. In this case, it is practical to use a rule of thumb to define the possible propagation times. This rule can be determined by analyzing the time of occurrence of consecutive events originating from different production units.

2.2.2. Analysis of Operator Actions. A similar analysis can also be performed on operator actions. Despite being a complex troubleshooting process that can last for hours, our primary goal is to identify the sequences of correlated operator actions (similar to parent-child type alarms). One way of achieving this is to define a time window lasting between 10 and 60 seconds (based on the cognitive model of operators and the attributes of the existing system). A new series of actions is identified if the time gap between two consecutive actions exceeds this time window. Another way is to regard alarm acknowledgements as the end of intervention activities (if this type of event is found in our log file). This way, they generate the groups of action series.

2.2.3. Connection between Alarms and Operator Actions. The most complex task is to analyze operator actions with regard to the alarms that trigger them and qualify the efficacy of the interventions.

To determine which operator actions trigger alarm events, operator actions should be placed into our existing alarm traces (which commence after the starting time of the trace and finish before the end time of the trace or the starting time of the following trace).

If the aim is to explore the effect of the operator actions, the end times of the alarms should be put into the aforementioned traces, as the Return to Normal pair of alarm events.

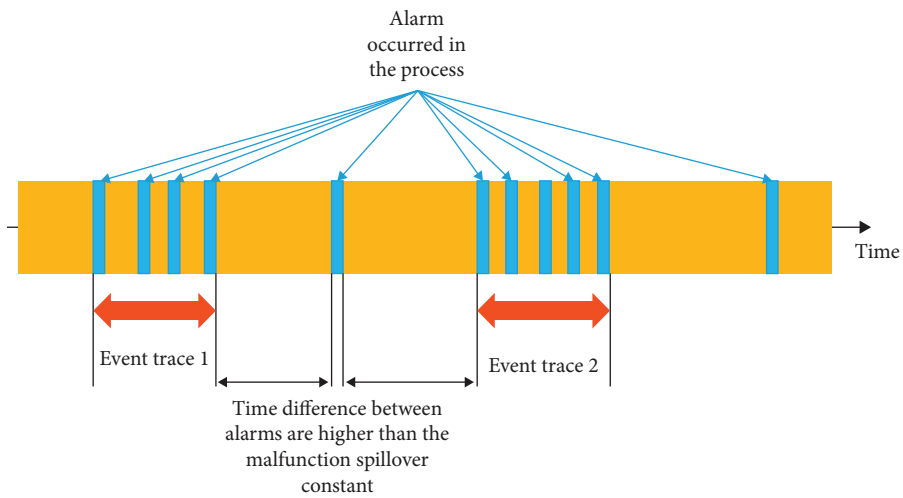


Figure 2: The segmentation of an event log database into event traces of potential propagation of error.

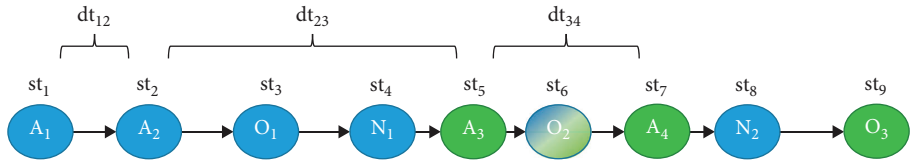


Figure 3: st_i denotes the start time of the event, dt_{ij} indicates the time difference between alarm start times, and dt_w indicates the time window and T_i the trace. If $dt_{12}, dt_{34} < dt_w$ and $dt_{23} > dt_w$ and $\forall st$: $st_i < st_{i+1}$, then $A_1, A_2, O_1, N_1, O_2, N_2 \in T_1$ and $A_3, O_2, A_4, O_3 \in T_2$. It is worth to note that operator action O_2 belongs to two traces.

(1) Rules of Generating Traces. In the previous list, trace generating methods were identified; now, the formalization of these is provided. At the start, we consider one trace, which contains all events, and with the following methods we will split it step by step.

$A_{j,n}$, $O_{k,n}$, and $N_{p,n}$ are the jth alarm, kth operator action, and pth return to normal event of the nth trace (σ_n), respectively, where $n \in \{1, \ldots, |\sigma_i|\}$, $j \in \{1, \ldots, |A_i|\}$, k and $p \in \{1, \ldots, |N_i|\}$. tw_1 is the time window constant for alarms, tw_2 is the time window constant for operator actions, and $t(A_{j,n})$, $t(O_{k,n})$ and $t(T_{n,n})$ are the timestamps of the related events. The explanation of the rules and their mathematical description is as follows.

Trace rule 1: $t(A_{j+1,n}) - t(A_{j,n}) > tw_1, A_{j+1,n} \longrightarrow A_{1,n+1}, A_{j,n} \longrightarrow A_{|A_i|,n}$: if the difference between the timestamps of two consecutive alarms is greater than tw_1, then the alarm with the higher index will be the first event of the next trace and the one with the lower index will be the last event of the actual trace. This rule can be applied to operator actions as well. These traces provide the input to gain the distribution of events over time, the identification of typical event sequences and the spillover of the alarms among production units.

Trace rule 2: From traces made by Trace rule 1, we generate the return to normal events from the end timestamps of the alarms. This means that the number of return to normal events will be equal to the number of alarm events ($|A_i| = |N_i|$) and traces will lap over each other, as an alarm of a trace can end later, than the start time of the next trace's first alarm. We put an operator action into the trace, if its timestamp is later than the first alarm of the trace and sooner than the last return to normal event of the trace ($t(O_{1,n}) > t(A_{1,n})$, $t(O_{|O_i|,n}) < t(N_{|N_i|,n})$). A visualized example is shown in Figure 3. From these traces, the effect of operator actions can be gained. To identify trigger type alarms, return to normal events have to be removed from the traces made by Trace rule 2. Obviously, they should not be excluded, but process mining a log without unnecessary events results in a more clear process model.

2.3. Process Mining-Based Alarm Management Solutions. In this section, the theoretical background of the applied analysis methods is presented.

Different algorithms of process mining can help us to identify patterns within the swarm of data placed in the log files. Given the need to find a solution to this common problem, different process mining techniques and several software products to evaluate the data mining tasks can be used. In this work, instead of using the well-known and usual Process Mining tools (like ProM [24] or EMiT [25]), we have used an open-source Python programming language-based solution, PM4Py. This library provides a wide range of process mining tools and since it is based on Python, which is a great tool for manipulating huge data sources (like industrial log files), PM4Py is an excellent option to develop semi- or fully automated process discovery methods from

Table 3: Tasks and process mining tools (A: alarm, O: operator action).

Task	Tool
Distribution over time (A and O)	Dotted chart
Typical sequences (A and O)	Directly-follows graph
Spillover among units (A)	Directly-follows graph/Heuristic miner
Trigger type alarms	Heuristic miner
Effect of operator actions	Heuristic miner

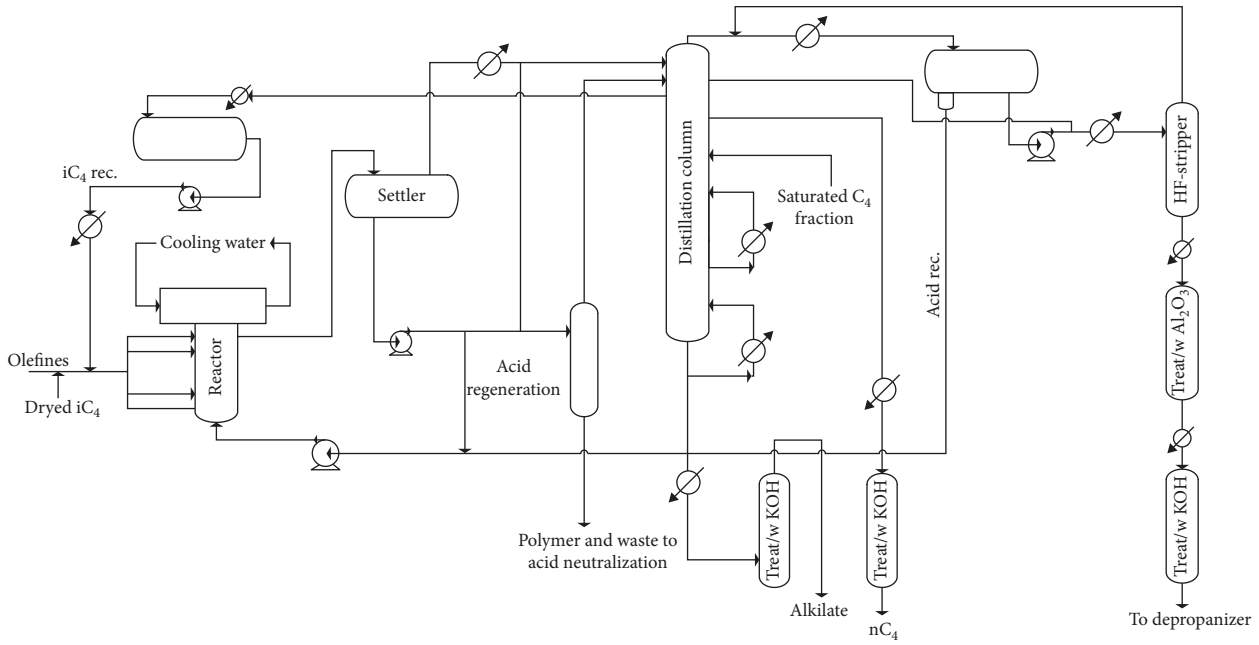


Figure 4: Schematic diagram of the plant in the case study.

scratch on one platform. The tasks and tools to be used are summarized in Table 3.

3. Results and Discussion

In order to show the industrial applicability of the formerly introduced process mining method, the alarm management system of an industrial hydrofluoric acid alkylation plant will be analyzed. The process flow diagram of the plant can be seen in Figure 4. The plant consists of four production units and more than 400 tags, the distributed control system is a Honeywell product. The logic of the tag names is $X,\ldots,X\_YY\_Z$, where $X\ldots X$ is the identifier of the tag, YY is the identifier of the production unit (where the tag is located), and Z is the type of the event. The identifiers of the production units are the following: CH: isostripper, propane-depleting and propane handling unit; U1: utility streams; AC: reactor and acid generating unit; FD: raw material and drying unit; 02: "virtual" unit, collection of sensors, that cannot be assigned to a specific unit. There are three types of events, namely, alarm (denoted with A), operator action (denoted with O), and return to normal (denoted with N).

Even though an alarm rationalization was performed on the plant, so the events in the log files are all considered to be relevant by the operating personnel, the log files need to be processed carefully, as remaining problems can be present, which can cause issues while undertaking the process discovery tasks.

3.1. Preparation of the Log File of the Alarm System. The log file of the aforementioned plant contains a lot of data, in excess of 200,000 events over a time period of four months. As previously discussed, the log file must be filtered to ensure only relevant and valuable data for the purpose of process mining is retained. The minimum number of attributes for process mining is three: an identifier of the event, at least one timestamp of the event (start or complete), and an identifier of the trace. Although not mandatory, it is useful to have additional attributes, for example, the name of the resource that triggered the event, the name of the organizational group in which the resource is located, and in the case of temporal-type events, the counterpart of the timestamp (start or complete) and the type of the event.

Another aspect that must be taken under consideration is what type of events to keep. Ten event types are present in the analyzed log file, namely, alarm, return to normal, acknowledge, operator action, system, operator message, suppress, shelved, unshelved, process event. According to our goals, first, it was decided to retain three types of events, that is, alarm, return to normal, and operator action. Alarm

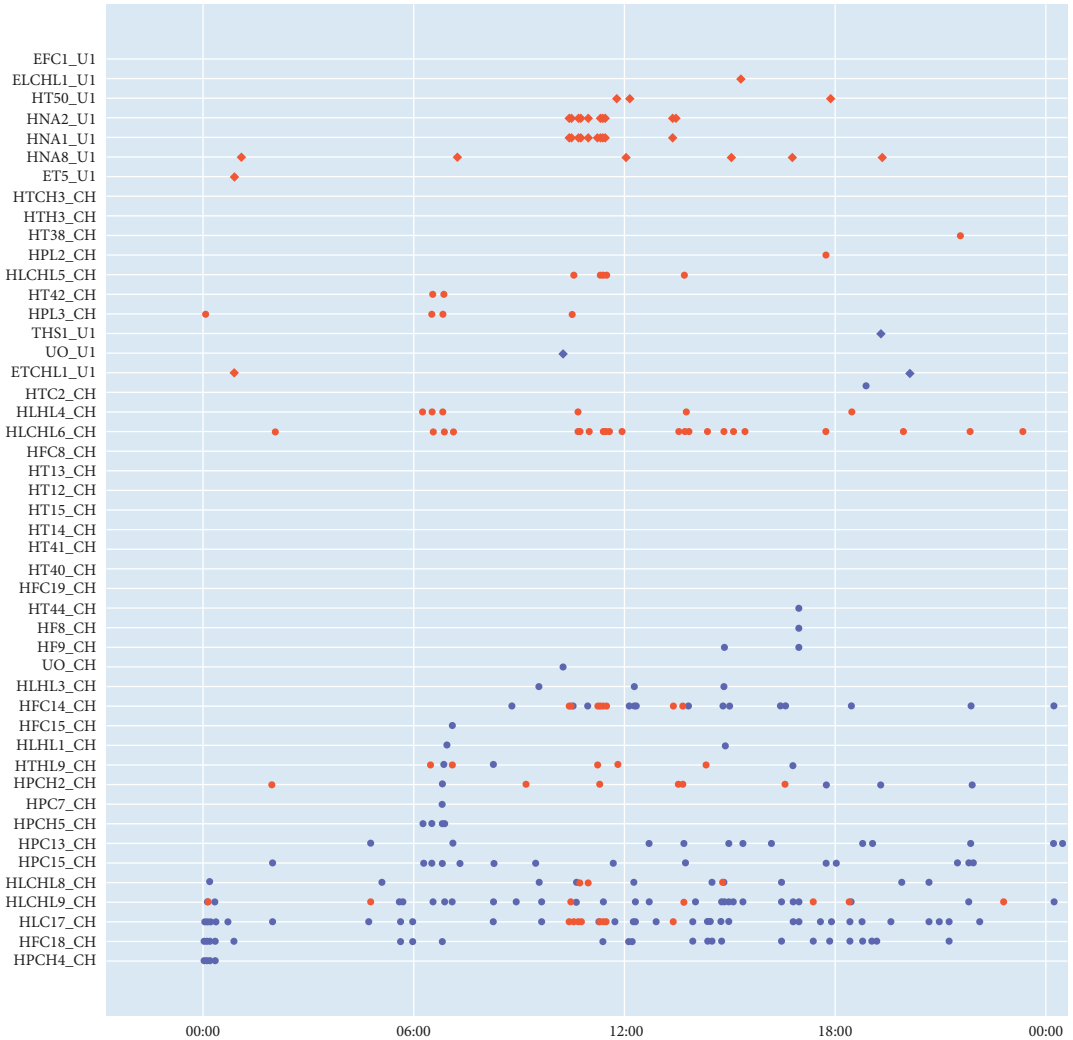


Figure 5: Distribution of alarms and operator actions over time (red: alarm, blue: operator action). The *x* axis shows the time and the *y* axis shows the name of the tags.

and operator action will be used to explore their number of occurrences and the sequences in which they occur; moreover, all three will be used to determine the causal relations between the alarms and operator actions. As a result of filtering out types of events, approximately 80,000 events remained.

As soon as the sub-logs with the needed types of events are obtained, traces must be generated. Trace window constants have to be defined for every sub-log, the object and the contained event types of these sub logs are collected in Table 2. The value of these constants depends on the actual system, literature data, and industrial experiences.

3.2. Distribution over Time and Typical Event Chains. To visualize the distribution of the events over time, a dotted chart is an excellent tool to use. Formerly, time distribution analysis has been identified as a task for both alarms and operator actions. As it can be interesting to compare the time of occurrence of alarms with the time of occurrence of operator actions, both have been visualized on one dotted chart. Figure 5 shows a one-day-long time window of production units CH and U1; the colors represent the types of events. It can be seen that although tags are present where both Alarm and OperatorAction events occur, many can be identified where this is not the case. It is not inevitable that interventions are made at the same place where the alarm occurs. For example, only alarms are located in production unit U1, in production unit CH, more operator actions are present than alarms. It can be supposed that the alarms in U1 trigger alarms in CH which in turn trigger the operator actions.

By briefly examining the time distribution statistics, it can be seen that either extremely long-lasting and near-to-zero second long alarms are present. These extreme values can have a biased effect on our process mining results. These events contain little or no timely information for the operators were filtered out from the log file. Only alarms that lasted between 5 and 28,800 seconds (8 hours) were retained. Of course, the upper and lower limits should be considered based on the given system and task.

Now that our log file has been "normalized," the next step is to generate the traces. If the needed trace windows are to be determined, the statistics regarding the time difference

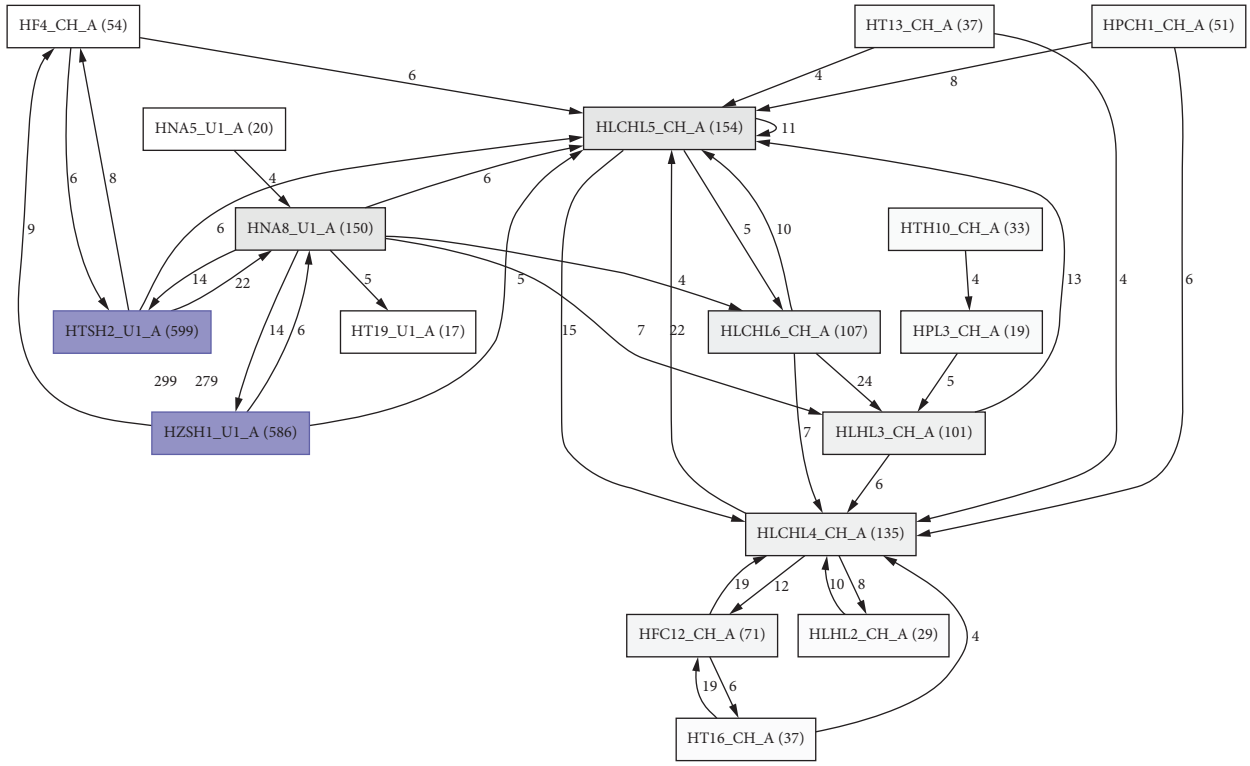


Figure 6: Alarm series (part of the DFG). CH: isostripper, propane-depleting, and propane handling unit; U1: utility streams.

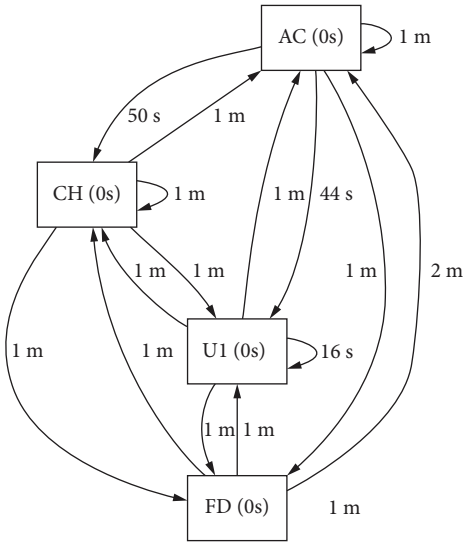


Figure 7: Alarm propagation times between units. CH: isostripper, propane-depleting, and propane handling unit; U1: utility streams; AC: reactor and acid generating unit; FD: raw material and drying unit.

between the starting times must be examined. According to the statistics concerning the alarm starting times, for exploring the typical alarm chains, the trace window was chosen to be 220 seconds (the median value). To ensure at least two events are found in a trace (which is the minimum to be considered in a chain), the traces consisting of one event must be removed.

The first tool that can be used to explore the typical alarm chains is the directly-follows graph (DFG). The frequency of both the events and nodes can be put on the graph. Figure 6 shows a portion of the DFG of the alarms, as the original graph is too large to be presented in full here.

Suppose the typical alarm propagation time between the units is sought, a DFG can be used once again, where the average elapsed time between two alarms occurring in different units is added, as is presented in Figure 7:

Although the average times are more or less identical, the frequency at which the alarms occur in unit U1 is much higher than in the other units. Zero seconds can be seen in the boxes of the units because the events were regarded as point-like, so they are dimensionless in terms of time.

Even though a DFG can provide a good overview of typical sequences, a different tool must be used to explore processes. One option is the heuristic miner algorithm to obtain a so-called heuristic net, which is one form of visualizing the typical processes. As frequent event chains are to be explored, the parameter minimum activity count was used and set at 100 and 500, as presented in Figure 8. On a heuristic net, the green ellipsis represents the start and the orange one represents the end of the process.

Obviously, the biggest proportion of the alarm events occurs in unit U1, which has a high rate of interaction with units CH and AC. Although our assumption that alarms in U1 trigger alarms in CH is proven, the opposite can also occur. This shows the advantage of a heuristic net over a dotted chart or a DFG.

To explore the typical series of operator actions, the formerly presented DFG is used. Figure 9 shows the

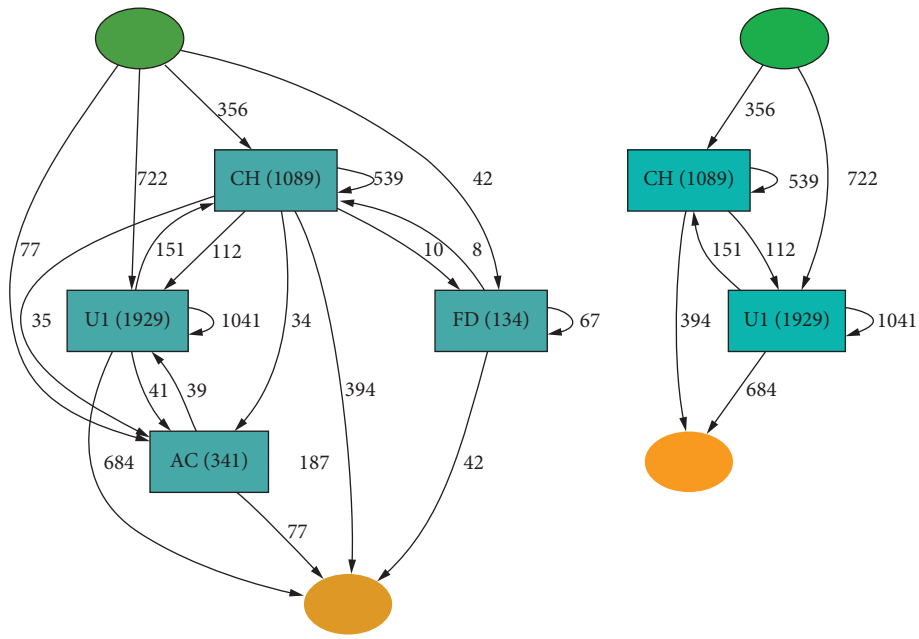


Figure 8: Heuristic nets of alarm propagation between production units. CH: isostripper, propane-depleting, and propane handling unit; U1: utility streams; AC: reactor and acid generating unit; FD: raw material and drying unit. (a) Minimum activity count = 100. (b) Minimum activity count = 500.

frequency of two consecutive operator actions. It can be seen that the most operator actions are located in unit CH, as was presumed from Figure 5. Nodes denoted in a darker color with thicker edges represent more frequent events and transitions. From this graph, the group of tags where the most of the operator actions occur can be identified, along with information about the frequent series of operator actions.

3.3. Correlation between Alarms and Operator Actions. To understand the connection between operator actions and alarms, two different questions may have to be answered:

(1) Which alarm triggers an operator action?

(2) What is the effect of the operator actions?

To answer these two questions, different log file and trace generation rules are required. The first requires the event types alarm and operator action, while the second requires return to normal events, which can be a result of an operator action. The trace generation rules were determined in Section 2.2. First, our formerly generated alarm traces are taken and return to normal events are generated from the end timestamps of the alarms. Subsequently, operator actions are placed into our traces with timestamps between the first and last events in the trace (practically speaking, the first alarm and the last return to normal events). By considering this method, trigger-type alarms and the effect of operator actions can be handled in one task as the log for trigger-type alarms would also be the one for exploring the effect of operator actions in the absence of the return to normal events.

Figure 10 shows the explored processes (minimum DFG was set at 55). By closely examining the net, two main types of processes (in addition to a third one) can be identified. For the purpose of better readability, starting and end points of the alarm sequences have been highlighted with colored boxes (same color for each pair). The boxes with dashed lines denote those processes where no operator action occurred between the starting and the end of an alarm sequence (process type 1). The ones denoted with solid lines mark processes containing operator action (process type 2). The third type, denoted by dotted lines belongs to both, as this alarm (HNA8_U1_A) can end either with or without an operator action (process type 3). Furthermore, an area which is worth examining closely is the green ellipse, as this part definitely appears to be a typical process (this "group" can also be seen in Figure 9). From this net, which tag is relevant in which kind of process discovery task can be determined, moreover, our log file can be filtered further and the mining conducted again.

The HLCHL9_CH_A alarm indicates problems with the liquid level at the HF stripper bottom. It is well visible that the operators frequently apply the HFC17_CH_O action in this situation, which modifies the bottom inlet of the HF stripper. Similarly, they apply the HFC18_CH_O action in this situation, which controls the steam inlet of the re-boiler of the stripper. In the case of the events marked by black brackets, the HPCH2_CH_A alarm indicates problems with the depropanizer pressure, while the action applied in this situation, HFC15_CH_O, controls the blow off of the technology. The HTSH5_U1_A and HZSHP1_U1_A alarms (dark blue and brown dashed brackets) almost always co-occur. As these alarms both related to the problem of the

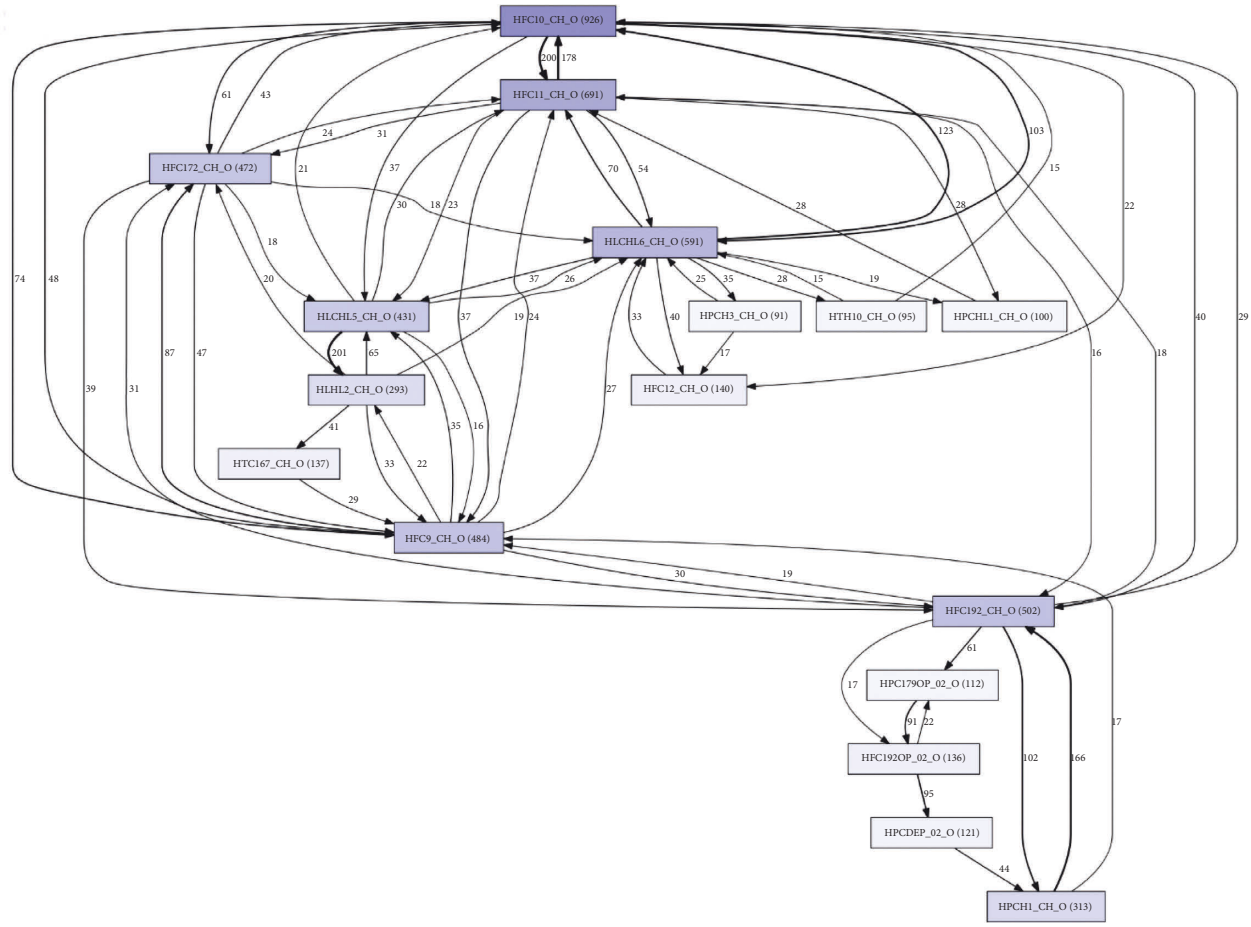


Figure 9: Series of operator actions. CH: isostripper, propane-depleting, and propane handling unit; 02: virtual unit.

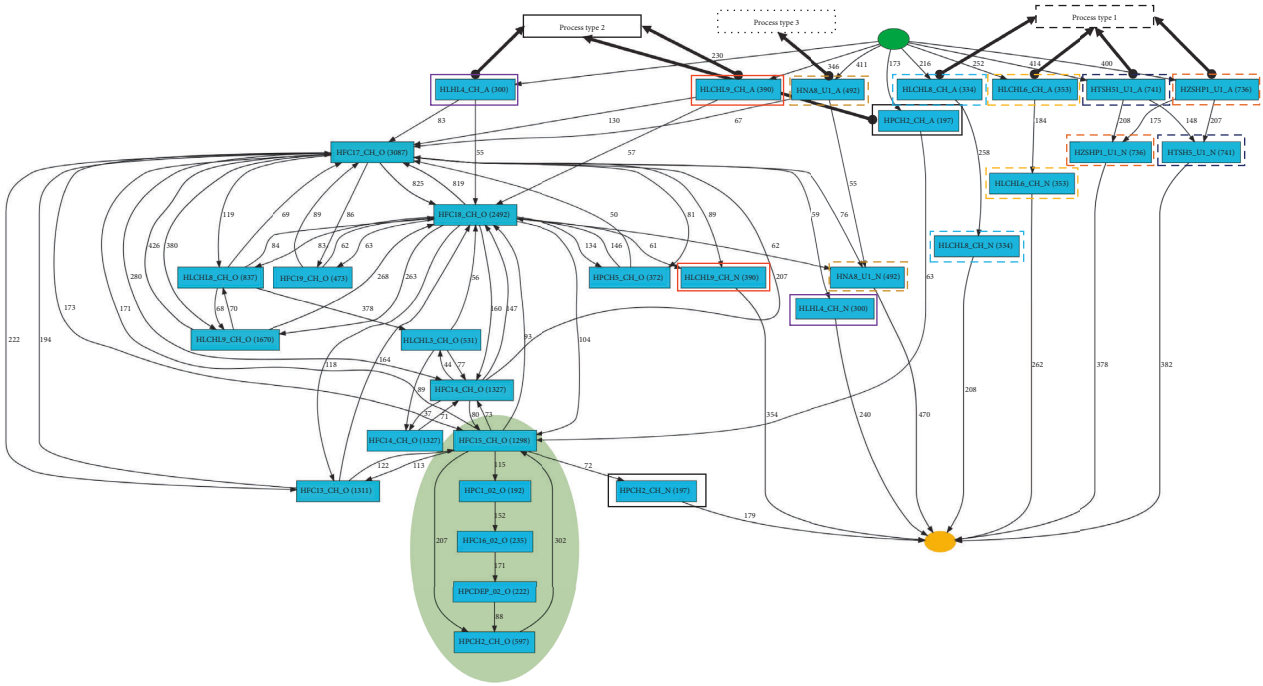


Figure 10: Alarm (A)-operator action (O)-return to normal (N) sequences. CH: isostripper, propane-depleting, and propane handling unit; U1: utility streams; 02: virtual unit.

same pump, they tend to be redundant and their definition should be revised by the process experts.

4. Conclusion

As industrial technologies are becoming more and more complex, the work of operators required to ensure the safe and optimal operation is getting increasingly challenging. With the introduction of the Industry 5.0 approach in progress, there is an emerging need for solutions to balance the productivity and efficiency with the life quality (work and home) of the workers, as well with the affection of the industry on society. One way to achieve this is to design alarm management systems based on tools that were developed to answer the challenges of the 4th Industrial Revolution (Industry 4.0). A properly built alarm management system can lower the workload of the operators significantly and reduce the probability of hazardous situations, affecting the environment (including civilians). The primary goal of this paper was to explore useful information from the historical data of industrial alarm management systems that can support the reduction of the operator workload and learn optimal operating strategies.

The paper proposed a process mining-based method to discover the fundamental relations between alarms and the related operator actions. Standard process mining techniques are not suitable for the analysis of historical process data of similar type. The paper demonstrated the benefits of the goal-oriented design of the log files that allows the extraction of information available to more effective alarm management and operator training.

The method was applied in the alarm management rationalization project of an industrial hydrofluoric acid alkylation plant. The project demonstrated that with the help of process mining, alarm signals could be rationalized; therefore, the work of the operators will become safer, as well as more effective, and last but not least, the workload has been decreased.

The discovered process models are easy to understand and provide some kind of improved digital visualization of alarms and operator actions. Lee et.al. summarized the applicability of digital twins in Industry 4.0-driven process safety management [26]. Our process mining-based method is also suitable to support some of the improvement actions collected in that study as a complementary tool. These improvement actions, related to alarms are: generate alarm signatures that can be useful in abnormal situation management, identify critical operator interventions, improve procedural risk assessments, and reduce the time and risk of errors during traditional risk assessment processes. They can support operator action-related tasks as well, namely, processing of procedures and operator actions: enhancing work design and operator performance, and representation and assessment of people and procedure related performance deviations and failures.

The gained information can be used to improve control systems, get a better insight into plant failure and behavior (process hazard analysis), review process safety incidents (incident investigation), and conduct what-if scenarios to understand how the plant may continue to run during unplanned maintenance or examining specific abnormal operation scenarios in more depth. It also gives the ability to assess initiating causes systematically and in an automated way based on historical data, due to the many failure modes of equipment that exist in a process plant. This is considered a key attribute to reduce resource intensive analysis.

Traditional hazard analysis processes have some shortcomings. A data science-based approach can address some of them, for example when there is a lack of

(i) depth of analysis

(ii) follow through to final consequences

(iii) completeness in identifying initiating causes and scenarios

The outcome of this study proved that the process mining-based analysis of events, along with the goal-oriented design of log files, should be added to the digital toolkit of process safety management.

Data Availability

The log data used to support the findings of this study have not been made available because of confidentiality reasons.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been implemented by the TKP2021-NVA-10 project with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development, and Innovation Fund, financed under the 2021 Thematic Excellence Programme funding scheme.

References

[1] E. Equipment, M. U. Association, E. Equipment, and M. U. A. Staff, *Alarm Systems: A Guide to Design, Management and Procurement*, EEMUA publication, London, England, 2015.

[2] G. Dorgo and J. Abonyi, "Sequence mining based alarm suppression," *IEEE Access*, vol. 6, Article ID 15379, 2018.

[3] G. Dorgo, P. Pigler, M. Haragovics, and J. Abonyi, "Learning operation strategies from alarm management systems by temporal pattern mining and deep learning," in *Proceedings of the 28th European Symposium on Computer Aided Process Engineering (ESCAPE28)*, A. Friedl, J. Klemes, S. Radl, P. S. Varbanov, and T. Wallek, Eds., pp. 1003–1008, Graz, Austria, June 2018.

[4] W. Hu, A. W. Al-Dabbagh, T. Chen, and S. L. Shah, "Process Discovery of Operator Actions in Response to Univariate Alarms," *IFAC-*, vol. 49, no. 7, pp. 1026–1031, 2016.

[5] D. S. Kim, H. Shinbo, and H. Yokota, "An alarm correlation algorithm for network management based on root cause analysis," in *Proceedings of the 13th International Conference on Advanced Communication Technology (ICACT2011)*, pp. 1233–1238, Gangwon-Do, South Korea, February 2011.

[6] N. A. Adnan, I. Izadi, and T. Chen, "On expected detection delays for alarm systems with deadbands and delay-timers," *Journal of Process Control*, vol. 21, no. 9, pp. 1318–1331, 2011.
[7] H. Zang, F. Yang, and D. Huang, "Design and analysis of improved alarm delay-timers," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 669–674, 2015.
[8] I. Izadi, S. L. Shah, D. S. Shook, S. R. Kondaveeti, and T. Chen, "A framework for optimal design of alarm systems," *IFAC Proceedings Volumes*, vol. 42, no. 8, pp. 651–656, 2009.
[9] W. Hu, J. Wang, and T. Chen, "A new method to detect and quantify correlated alarms with occurrence delays," *Computers & Chemical Engineering*, vol. 80, pp. 189–198, 2015.
[10] S. Lai and T. Chen, "A method for pattern mining in multiple alarm flood sequences," *Chemical Engineering Research and Design*, vol. 117, pp. 831–839, 2017.
[11] G. Dorgo, P. Pigler, and J. Abonyi, "Understanding the importance of process alarms based on the analysis of deep recurrent neural networks trained for fault isolation," *Journal of Chemometrics*, vol. 32, no. 4, Article ID e3006, 2018.
[12] G. Dorgo, A. Palazoglu, and J. Abonyi, "Decision trees for informative process alarm definition and alarm-based fault classification," *Process Safety and Environmental Protection*, vol. 149, pp. 312–324, 2021.
[13] W. van der Aalst, *Data Science in Action*, pp. 3–23, Springer, Berlin, Heidelberg, 2016.
[14] J. Theis, W. L. Galanter, A. D. Boyd, and H. Darabi, "Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 388–399, 2022.
[15] P. Zerbino, A. Stefanini, and D. Aloini, "Process science in action: a literature review on process mining in business management," *Technological Forecasting and Social Change*, vol. 172, Article ID 121021, 2021.
[16] M. Werner, M. Wiese, and A. Maas, "Embedding process mining into financial statement audits," *International Journal of Accounting Information Systems*, vol. 41, Article ID 100514, 2021.
[17] R. Cerezo, A. Bogarin, M. Esteban, and C. Romero, "Process mining for self-regulated learning assessment in e-learning," *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 74–88, 2020.
[18] V. Leno, A. Polyvyanyy, M. Dumas, M. La Rosa, and F. M. Maggi, "Robotic process mining: vision and challenges," *Business & Information Systems Engineering*, vol. 63, no. 3, pp. 301–314, 2021.
[19] C. D. S. Garcia, A. Meincheim, E. R. Faria Junior et al., "Process mining techniques and applications – a systematic mapping study," *Expert Systems with Applications*, vol. 133, pp. 260–295, 2019.
[20] R. E. Kondo, E. de F R Loures, and E. A. P. Santos, "Process mining for alarm rationalization and fault patterns identification," in *Proceedings of the 2012 IEEE 17th International Conference on Emerging Technologies Factory Automation (ETFA 2012)*, pp. 1–4, 2012.
[21] R. E. Kondo, E. de Freitas Rocha Loures, E. A. Portela Santos, and C. M. P. Braga, "Alarm rationalization based on process mining techniques," *Advanced Materials Research*, vol. 1061-1062, pp. 1258–1265, 2014.
[22] W. M. P. W. D. Aalst, "A practitioner's guide to process mining: limitations of the directly-follows graph," *Procedia Computer Science*, vol. 164, pp. 321–328, 2019.
[23] J. R. Taylor, "Automated HAZOP revisited," *Process Safety and Environmental Protection*, vol. 111, pp. 635–651, 2017.
[24] H. M. W. Verbeek, J. C. A. M. Buijs, B. v. Dongen, and W. M. P. V. D. Aalst, *XES, XESame, and ProM*, vol. 6, 2011.
[25] B. F. V. Dongen and W. M. P. V. D. Aalst, "EMiT: a process mining tool," in *Applications and Theory of Petri Nets 2004*, J. Cortadella and W. Reisig, Eds., pp. 454–463, Springer, Berlin, Heidelberg, 2004.
[26] J. Lee, I. Cameron, and M. Hassall, "Improving process safety: what roles for Digitalization and Industry 4.0?" *Process Safety and Environmental Protection*, vol. 132, pp. 325–339, 2019.

Hindawi
Complexity
Volume 2022, Article ID 9518303, 12 pages
https://doi.org/10.1155/2022/9518303

WILEY | Hindawi

Research Article

Complexity: Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022 on the Development of Information Theoretic Model Selection Criteria for the Analysis of Experimental Data

Andrea Murari,[1] Michele Lungaroni,[2] Riccardo Rossi,[2] Luca Spolladore,[2] and Michela Gelfusa[2]

[1]*Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, Padova 35127, Italy*
[2]*University of Rome "Tor Vergata", Department of Industrial Engineering, via del Politecnico 1, Roma, Italy*

Correspondence should be addressed to Michele Lungaroni; michele.lungaroni@uniroma2.it

Received 21 January 2022; Accepted 20 July 2022; Published 24 August 2022

Academic Editor: M. De Aguiar

It can be argued that the identification of sound mathematical models is the ultimate goal of any scientific endeavour. On the other hand, particularly in the investigation of complex systems and nonlinear phenomena, discriminating between alternative models can be a very challenging task. Quite sophisticated model selection criteria are available but their deployment in practice can be problematic. In this work, the Akaike Information Criterion is reformulated with the help of purely information theoretic quantities, namely, the Gibbs-Shannon entropy and the Mutual Information. Systematic numerical tests have proven the improved performances of the proposed upgrades, including increased robustness against noise and the presence of outliers. The same modifications can be implemented to rewrite also Bayesian statistical criteria, such as the Schwartz indicator, in terms of information-theoretic quantities, proving the generality of the approach and the validity of the underlying assumptions.

1. Introduction to Nonfrequentist Model Selection Criteria

The promised land of modern scientific enterprises is often the formulation of robust and generally applicable mathematical models [1, 2]. The ultimate validation of any model resides in the comparison with the results of experiments or observations. In the last decades, enormous quantities of data have become available in many fields of science and engineering. The statistical inference has therefore progressively moved to centre stage. The older frequentist techniques, based on traditional significance level criteria, have been complemented by a series of Bayesian and information-theoretic criteria, in many respects more suited to managing large amounts of information.

One of the most popular model selection criteria (MSC) is the Akaike Information Criterion (AIC) [3]. The AIC can be derived from the Kullback–Leibler divergence and can be interpreted as the loss of information associated with the adoption of a model different from the exact one, generating the data. The basic idea underlying the AIC criterion resides indeed in the consideration that the less information a model loses, the higher its quality. The theoretical derivation of the AIC gives the unbiased form of the criterion [4].

$$AIC = -2\ln(L) + 2k, \tag{1}$$

where L is the likelihood of the data given the model and k is the number of estimated parameters in the model. The AIC is a metric that is minimised by the best model as a compromise between the goodness of fit (the first term) and complexity (the second term).

The general formulation of the AIC is not always easy to apply in practice as can be appreciated by a simple inspection of (1). First, in many instances, it can be impossible to

reliably calculate the likelihood. Moreover, it is well known that the number of parameters is a poor quantifier of a model complexity and it is not inherently an information-theoretic indicator. The more practical expression of the AIC, very often the one used in practice, is even more distant from its original information theoretic origin, as discussed in the next section.

The first quantity, proposed to improve the AIC, is the Gibbs–Shannon entropy H

$$H = -\sum_i p_i \log\ p_i. \tag{2}$$

The higher the value of H, the higher the uniformity of the corresponding probability distribution function (whose values are indicated with p_i). The Gibbs- Shannon entropy can improve significantly the quantification of the model complexity, as discussed in detail in Section 2.2.

The second quantity, used in the rest of the work, is the mutual information, MI.

$$MI = -\sum_x \sum_y p_{xy} \ln\left(\frac{P_{xy}}{P_x P_y}\right), \tag{3}$$

where $P_{x,y}$ is the joint pdf of the random variables X and Y. Mutual Information can play a fundamental role in determining the goodness of fit of the models, as discussed in Section 2.1.

With regard to the organization of the paper, the next section introduces the rationale and details of the proposed information-theoretic upgrades of the Akaike Information Criterion. Section 3 is devoted to a simple but challenging didactic case, meant to illustrate the effects of the modifications with an easy-to-grasp example. The family of functions and the types of noise statistics, implemented to perform a series of systematic tests, are summarised in Section 4. The results of the aforementioned tests are exemplified in Section 5 with the help of some representative cases. The extension of the approach to the Bayesian Selection criterion is covered in Section 6 before the conclusions and lines of future developments are discussed in the final section of the paper.

2. Model Selection Formulated in terms of Information Theoretic Quantities

Among the many indicators, for identifying the "best model" among a set of candidates, the Akaike Information Criterion AIC can be conceived originally as a pure information theoretic criterion. Unfortunately, the original formulation of the AIC criterion is typically problematic to implement in practice, particularly in applications involving complex systems and nonlinear phenomena. Both terms in the AIC present significant issues [5–7]. To bypass the practical difficulties of calculating the likelihood, the strong assumption that the data are identically distributed and independently sampled from a normal distribution is the most commonly invoked. If this traditionally called iid hypothesis is valid, it can be demonstrated that the AIC can be written (up to an additive immaterial constant depending only on the number of entries in the database) as follows:

$$AIC = n \cdot\ \ln(MSE) + 2k. \tag{4}$$

In (4), formally derived in [4], the Mean Squared Error (MSE) is calculated in terms of the residuals, the differences between the data, and the estimates of the models; in its turn n indicates the number of entries in the database.

(4) is certainly the most widely used form of AIC. On the other hand, as can be easily appreciated by inspection, the criterion is now expressed in terms of quantities, which are not information theoretic anymore. Moreover, all the statistical information content, originally in the likelihood, is reduced to the mere MSE of the residuals. The first obvious question, which comes to mind, is whether some additional statistical information about the distribution of the residuals could be taken into account, to improve the discriminatory capability of the criterion. The practical relevance of this issue is quite significant also because, in many applications, the assumptions behind (4) are clearly violated. In real life, indeed, the statistics of the noise can have a non-Gaussian distribution, memory effects can be important, and a significant number of outliers can be unavoidable. How to improve the model selection criteria in this respect is the subject of Section 2.1.

The second term in (4) is also problematic because it is well known that the number of parameters is a quite poor indicator of the complexity of a model. More sophisticated quantifiers exist, such as the VC dimension [8] and the Rademacher dimension [9], but they are often impossible to calculate for most practical functions. An alternative information theoretic and computationally simple way to calculate a model complexity is the subject of Section 2.2.

2.1. Expressing the Goodness of Fit in terms of Mutual Information. The main idea informing one of the AIC upgrades, proposed in this work, is based on the observation that the better a model, the more similar the residuals to the noise affecting the measurements. In the case of a perfect model, the residuals should present exactly the same distribution as the noise. Assuming that the noise is not correlated with the measurements, absolutely legitimate in most practical applications, this consideration can be quantified mathematically by calculating the mutual information between the model predictions and the residuals, MI_{MRes}.

$$MI_{MRes} = MI\left(y_{mod}, y_{res}\right). \tag{5}$$

The AIC can therefore be rewritten as follows:

$$AIC_{MI} = 2k + n\ \ln\left(MSE\left(1 + MI_{MRes}\right)\right). \tag{6}$$

Conceptually, (6) is to be preferred to (4) for various reasons. First, it formulates the criterion in terms of an information theoretic quantity, the mutual information. Moreover, it retains much more statistical information about the model and the residuals. At the same time, MI_{MRes} takes into account also nonlinear correlations and does not make any "a priori" assumption about the statistics of the

noise or the presence of outliers. Consequently, as shown by numerical tests, AIC_{MI} is a much more general and sensitive model selection criterion than the original AIC.

2.2. Expressing the Complexity in terms of the Shannon Entropy. The other weakness in the original definition of AIC is certainly the quantification of complexity. Indeed, the simple number of parameters in a model is a very poor indicator of its flexibility and in particular of its potential to overfit (see Section 3). A possible alternative relies on the traditional idea that complexity is the middle ground between randomness and determinism. According to this view, complete randomness and perfect determinism are considered less complex than a combination of the two. This approach to complexity has a long pedigree and can be traced back to the interpretation of information as uncertainty, the concept at the basis of information theory [10]. A possible way of expressing this idea in mathematical terms is the following complexity measure C[X]:

$$C[X] = H^{\alpha}[X]D^{\beta}[X], \tag{7}$$

where H is the usual Shannon entropy and D is the distance from a uniform distribution.

$$D[X] = \sum_{1}^{N}\sum_{1}^{N}\left(p_i - \frac{1}{n}\right), \tag{8}$$

where with the usual notation, n is the number of entries in the database. The distance D reduces the estimated complexity of models, whose predictions are uniform. The entropy reduces the estimated complexity of models, whose outputs are concentrated on a few well-defined values. Conceptually, the implementation of this quantification of complexity is quite simple. The pdf of the model predictions can be inserted in (7) to obtain a simple indicator, implementing the aforementioned information theoretic interpretation of complexity.

The most delicate aspect of (7) is the choice of the exponents α and β because they contribute significantly to determining the trade-off between entropy and distance. To this end, the increments of the model predictions have been calculated as follows:

$$\text{Model}_{diff} = \left(y_{\text{model},i+1} - y_{\text{model},i}\right). \tag{9}$$

The moving averages (Mov), of the mean and standard deviation of the squared increments, are good indicators of the flexibility of a model and therefore of its potential to overfit. The normalized versions of these quantities are defined in

$$MF_1 = \frac{\sum MovST\ D\left(\text{Model}_{di\ ff}\right)^2}{n}$$
$$MF_2 = \frac{\sum Mov\text{MEAN}\left(\text{Model}_{di\ ff}\right)^2}{n}. \tag{10}$$

The ratio of the two averages calculated in (10) is

$$MF = \sqrt{\frac{MF_1}{MF_2}}. \tag{11}$$

The parameter MF increases for functions, which have stronger variations in the domain of interest and can therefore be considered more complex. Indeed, these more nervous functions would have a higher potential of overfitting the data, following the noise. This is the interpretation of the quantity MF, which is used to determine the exponents α and β.

$$\alpha = 1 + MF; \beta = 1 - MF. \tag{12}$$

Finally, the proposed final versions of the AIC expressed only in terms of the mentioned information theoretic quantities read

$$\begin{aligned} AIC_{MICx} &= n\ \ln[MSE(1 + MI)] + n\left(\ln C_x\right) \\ &= n\left(\ln\left[C_x MSE(1 + MI)\right]\right). \end{aligned} \tag{13}$$

3. A Didactic Example to Illustrate the Main Characteristics of AIC_{MICX}

To illustrate the potential and the meaning of the proposed upgrades of the AIC, an academic but challenging example, already discussed in detail in the literature [11], is described in this section. To this end, it is assumed that the actual data is generated with a polynomial function depending on 5 parameters.

$$y_{ref} = 10^{-6}x^5 - 8, 10^{-3}x^3 + 3, 10^{-2}x^2 + x - 10. \tag{14}$$

The equations, considered as possible candidate models for the data generated with (14), are reported in Table 1.

A comment about the sinusoidal functions is in place. These functions can be tuned to fit perfectly the data generated with (14) by increasing their frequency. This fact can be appreciated by inspection of the first two plots of Figure 1. If there is any noise added to the data, the sinusoidal functions, given their higher flexibility, can fit the data even better than the original equation generating it.

On the other hand, they depend only on two parameters, their amplitude and frequency. Therefore, the traditional version of the AIC would tend to prefer a well-adjusted sinusoidal model (because it would achieve lower values of both terms of the indicator). The proposed version AIC_{MICx}, on the contrary, manages to properly identify the right model, as shown in Figure 2. The plots report the differences between the AIC and AIC_{MICx} of the candidate models and the reference, the equation used to generate the data.

When these differences are positive, the reference model is the preferred one; the negative cases indicate that the criteria would have selected the wrong model. From the plots of Figure 2, it appears quite clearly that the traditional AIC would have preferred the sinusoids (particularly model 1) for various numbers of entries, whereas the AIC_{MICx} always identifies the reference model as the right one. This is achieved by taking into account the distributions of the

Table 1: The four candidate models to fit the data generated by (14).

#	Models
1	$17\ \sin(210x)$
2	$17\ \sin(209.5x)$
3	$-0.08x^2 + 1.47x - 10.38$
4	$0.75x - 10$

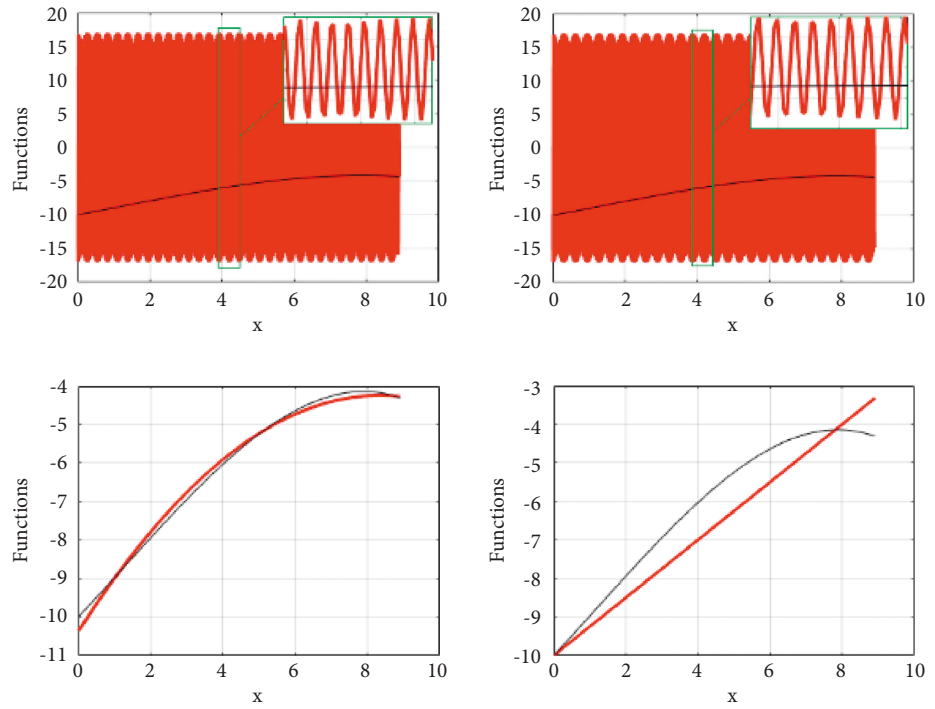


Figure 1: Black: the original data generated with (14). Red: the models of Table 1. From top left to bottom right models from 1 to 4.

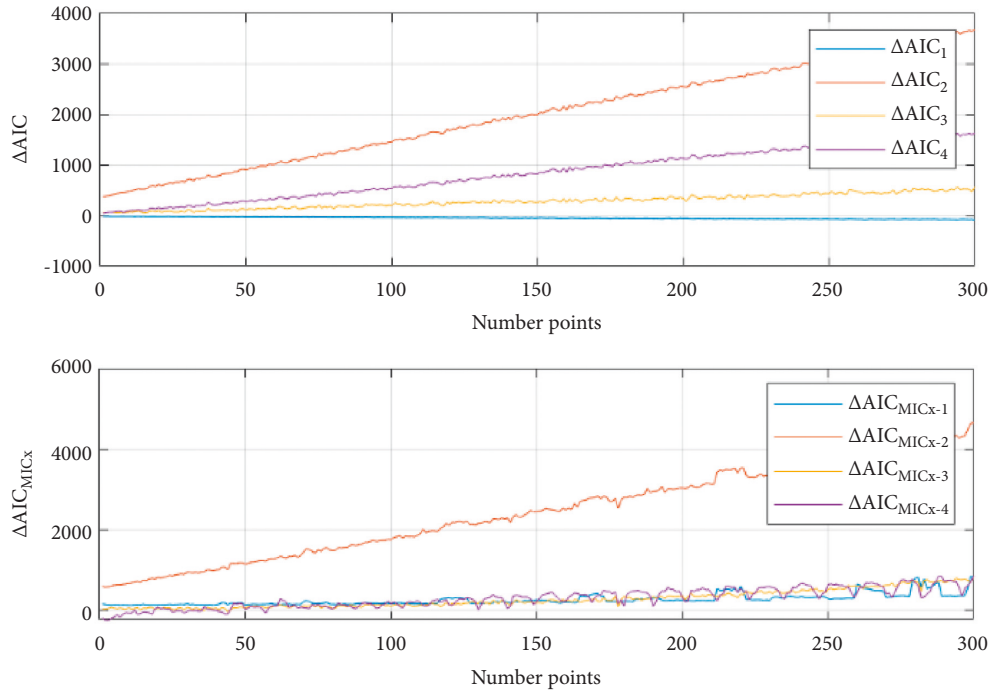


Figure 2: Comparison of the discriminating power of the traditional AIC and the proposed AIC_{MICx}. On the (y) axis the difference between the indicators for the various models and the reference one used to generate the data, is reported. On the (x) axis a scan in the number of entries.

residuals and by better estimating the complexity of the models. The details about the comparison, between the traditional AIC and the new version proposed in this paper, are fully documented in Appendix A for the specific example reported in this section.

4. The Main Functional Classes and Noise Statistics for Practical Applications

To assess the performance of the alternative AIC model selection criterion proposed in Section 2, a series of systematic numerical tests have been performed. The analysis is focussed mainly on four classes of models that cover the most widely used in practice. They are the classes of polynomials, power laws, power laws multiplied by a squashing term, and exponential functions. In the rest of the paper, only the results for bidimensional functions (of the form $z = f(x, y)$) are discussed, because they are susceptible of clear visualization, which helps illustrating the properties of the criterion. The extension to a larger number of variables is straightforward and does not pose any conceptual difficulty. Therefore, the considerations and conclusions reported have to be assumed valid also in higher dimensions. For the reader's convenience, the mathematical form of the aforementioned models is reported in the left column of Table 2.

Significant attention has been devoted to noise statistics. Three of the most relevant distribution functions have been tested: Gaussian, uniform, and multi-Gaussian [12]. Again for the reader's convenience, the mathematical formulation of these types of noise is summarised in the right column of Table 2, together with the parameter values valid for the runs reported in the rest of the paper. Since in practice very often the presence of outliers in the data cannot be excluded, the robustness of the proposed upgrade of the AIC in this respect has also been verified. This has been achieved by randomly adding to the synthetic data values sampled from a Gaussian distribution of small variance but nonzero mean (see the entry called Asymmetric noise in Table 2 for a precise mathematical definition).

Table 2: The main families of functions tested and the statistics of the additive noise.

Families of functions	Additive noise applied
Polynomials $y = a_0 x^{b_0} + a_1 x^{b_1} + a_n x^{b_n}$	*Uniform Noise* $\mu = \pm 10$ until ± 50
Power Laws $y = a_0 x^{b_0}, x^{b_1}, x^{b_n}$	*Traditional Gaussian Noise* $\mu = 0$ range of $\sigma = \pm 10$ until ± 50
Power Laws with Squashing term $y = a_0 x^{b_0}, x^{b_1}, x^{b_2} \frac{1}{1+exp(-a_n x^{b_n})}$	*Multi-Gaussian Noise* $\mu_i = 0$ range of $\sigma_i = \pm 10$ until $\pm 50 \forall i = 1..n$
Exponentials $y = a_0 x^{b_0} \exp(a_n x^{b_n})$	*Asymmetric Noise* $\mathcal{N}_1$: $\mu_1 = 0$ and $\sigma_1 = 10$; $\mathcal{N}_2$: $\mu_2 \neq 0$ and $\sigma_2 = 30$; with $\mu_2 = 2(\sigma_1 + \sigma_2)/100, \overline{f(x)}$ Ratio between $\mathcal{N}_1, \mathcal{N}_2 \Rightarrow 0.75$ until 0.95

5. Representative Results of Numerical Tests

As mentioned, a systematic series of tests with synthetic data has been performed to assess the competitive advantage of the proposed version of the AIC. All the combinations of cases summarised in Section 4 have been investigated. The new version $\mathrm{AIC}_{\mathrm{MICx}}$ has always proved to have better discriminatory capabilities than the traditional AIC. In practice, this means that $\mathrm{AIC}_{\mathrm{MICx}}$ at least provides better separation between the right model (the one used to generate the data) and its wrong competitors. This has proved to occur for any type of function, noise statistics, and levels of outliers. In general, the more severe the conditions, the higher the level of noise or outliers, and the better the $\mathrm{AIC}_{\mathrm{MICx}}$ performance compared to the traditional AIC. In some cases, as the one already discussed in Section 3, only the $\mathrm{AIC}_{\mathrm{MICx}}$ can converge on the right model.

In the rest of this section, some relevant examples of the performed tests are reported. They have to be considered absolutely representative of the vast majority of systematic investigations performed.

In the first case discussed in the following, the model generating the data consists of a power law multiplied by a squashing term. The importance and popularity of power laws are difficult to overstate. Self-similarity can result in many quantities presenting a power law trend. Power laws are also particularly important for the investigation of scalings. On the other hand, power law monomials can be too rigid and the multiplication by a squashing factor can provide some additional flexibility. The function implemented to generate the synthetic data is reported in the last row of Table 3. The other rows of the same table report the alternative models. The synthetic data generated with the reference model of Table 3 is shown in Figure 3, together with the functions constituting the alternative models. Two different levels of Gaussian additive noise are shown; corresponding to a standard deviation of 15% and 30% of the synthetic data averaged amplitude. As can be derived by simple inspection of the plots, $\mathrm{AIC}_{\mathrm{MICx}}$ not only increases the separation between the models, compared to the traditional AIC, but it also allows identifying the equation generating the data. Indeed whereas, for some numbers of entries and 30% of added noise, the AIC of the candidate models can be lower than the reference one, the $\mathrm{AIC}_{\mathrm{MICx}}$ always identifies the model generating the data as the best; this can be seen by noticing that the values of the $\mathrm{AIC}_{\mathrm{MICx}}$ differences, with respect to the best model, are always positive.

The discriminatory power of $\mathrm{AIC}_{\mathrm{MICx}}$ is even higher in the case of high noise. This fact is exemplified by the following example, in which the generating model belongs to the class of exponential functions. The alternative models are reported in Table 4, whose last row reports the equation used to generate the data. In addition to Gaussian noise, with a standard deviation of 30% and 60% of the synthetic data averaged amplitude, some concentrated high noise has also been added, according to the relations specified in the last row of Table 2. The better performance of $\mathrm{AIC}_{\mathrm{MICx}}$ compared to the traditional AIC can be easily recognised by

Table 3: Power law plus a squashing term.

#	Models	k
1	$1.6810^4 \sin(x_1/x_2^{4.18})$	4
2	$3x_2 \exp(-x_3^{9.48})$	4
3	$17.87(x_1/x_2^{0.45})^{0.47}$	4
4	$3.5x_1^{0.4}x_2^{0.8}$	3
ref	$\mathbf{2x_1^{0.6}x_2^{1.1}1/1 + \exp(-2x_3^{1.5})}$	**6**

The value is shown in bold because it is the reference model.

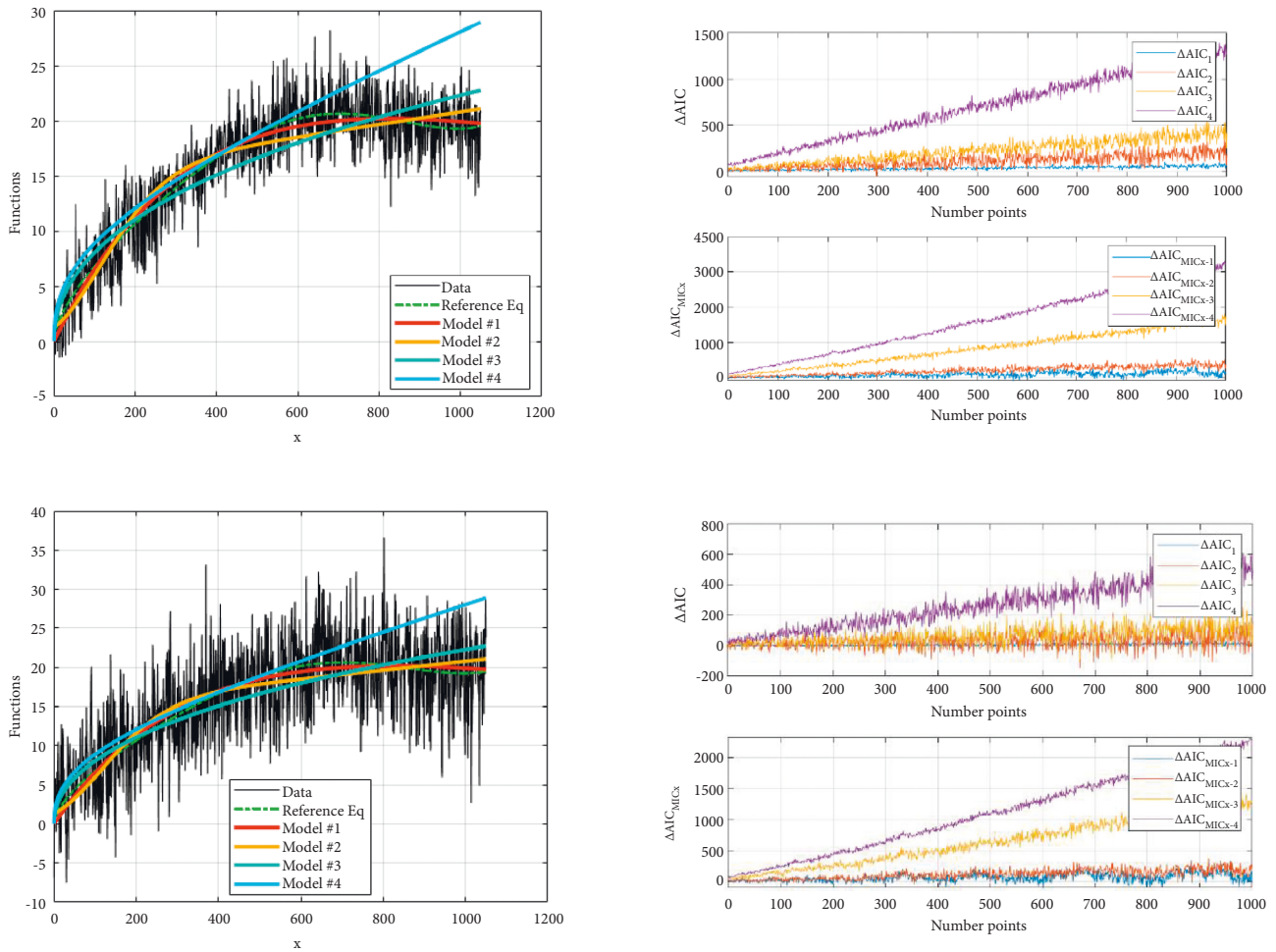


Figure 3: Model selection performances for two levels of additive noise: 15% top and 30% bottom. For each level of noise, the top plots show in black the synthetic data generated with the reference equation of Table 2. The coloured curves are the various candidate models and in dashed point green is the reference one. The bottom plots are the comparison of AIC and AIC_{MICx} results in terms of the difference with respect to the exact reference model.

Table 4: Power law plus a squashing term.

#	Models	k
1	$0.4x^{0.2}\exp(x)$	4
2	$0.8\ \exp(x) - 0.4\,x^2)$	5
3	$3x^2/1 + \exp(-0.1x)$	5
4	$0.5x^3 + 2x$	4
ref	$\mathbf{0.6x\ exp(x^{0.6})}$	**4**

The value is shown in bold because it is the reference model.

inspection of the plots in Figure 4. Indeed, the separation between the alternative models and the right one is much larger for the AIC_{MICx} than for the traditional AIC (the reader should please consider also the different scales of the plots in Figure 4).

6. Extension to Bayesian Model Selection

It is worth noting that the same modifications proposed for the AIC can be applied also to the Bayesian information criterion (BIC) [13]. BIC is based on Bayesian theory and has been designed to maximize the posterior probability of a model given the data. BIC is again a cost function and therefore it is also an indicator to be minimised. The BIC's most general form is

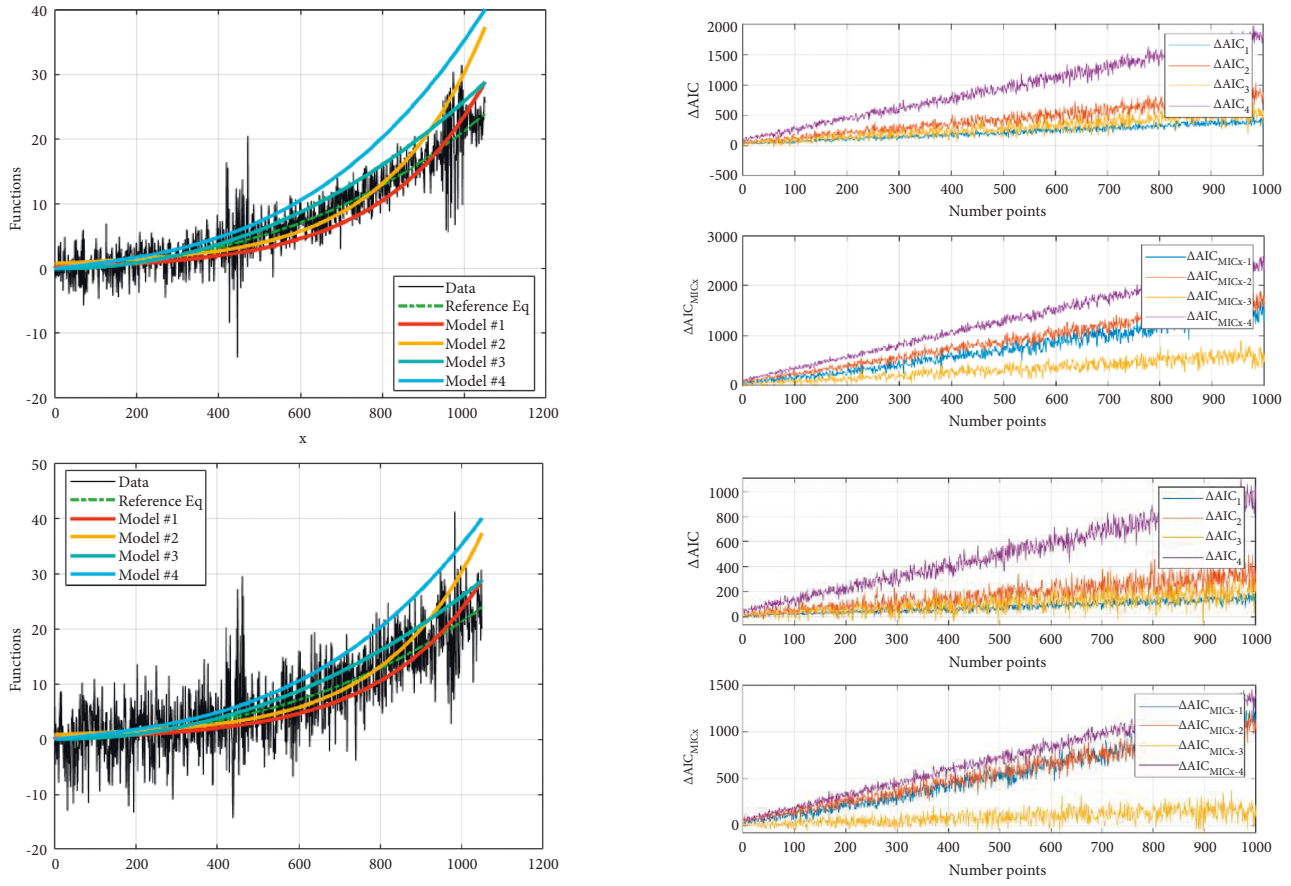


FIGURE 4: Model selection performances for two levels of additive noise: 30% top and 60% bottom. For each level of noise, the top plots show in black the synthetic data generated with the reference equation of Table 4. The coloured curves are the various candidate models and in dashed point green is the reference one. The bottom plots are the comparison of AIC and AIC_{MICx} results in terms of the difference with respect to the exact reference model.

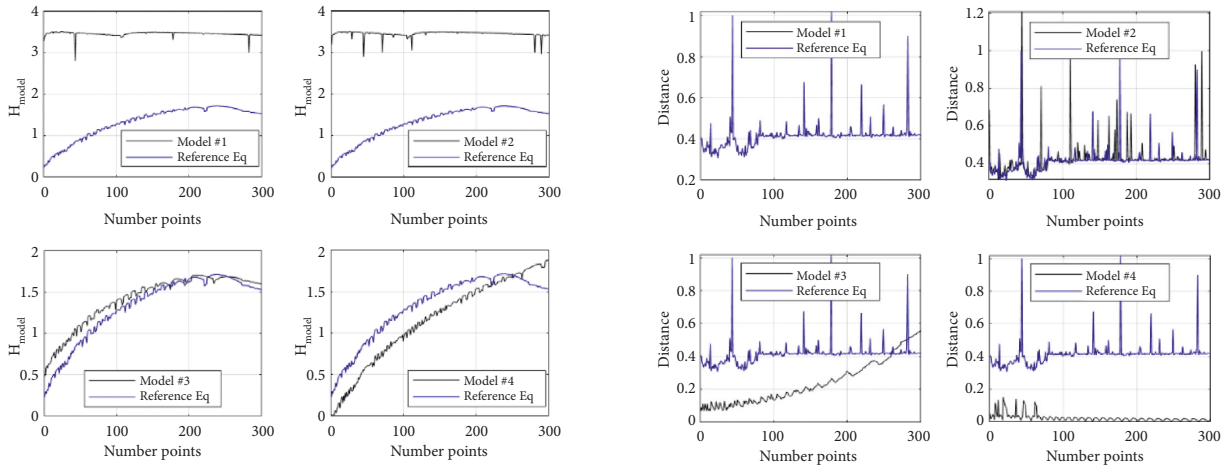


FIGURE 5: Plots of the entropy H and distance D for the models of Table 1 in Section 3.

$$BIC = -2\ln(L) + k\ln(n), \tag{15}$$

where again L is the likelihood of the data given the model, k is the number of estimated parameters in the model, and n is the number of entries in the database. BIC has the same structural form as the AIC and is affected by the same difficulties in practical applications, in particular the challenges posed by the calculation of the likelihood and the quantification of the model complexity.

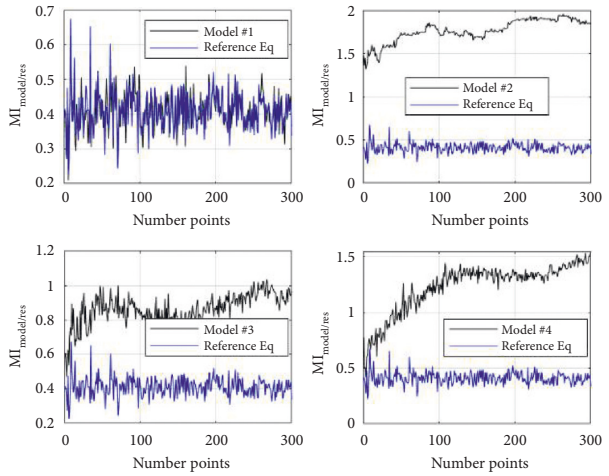


Figure 6: Plots of the mutual information between the models and the residuals for the models of Table 1 in Section 3.

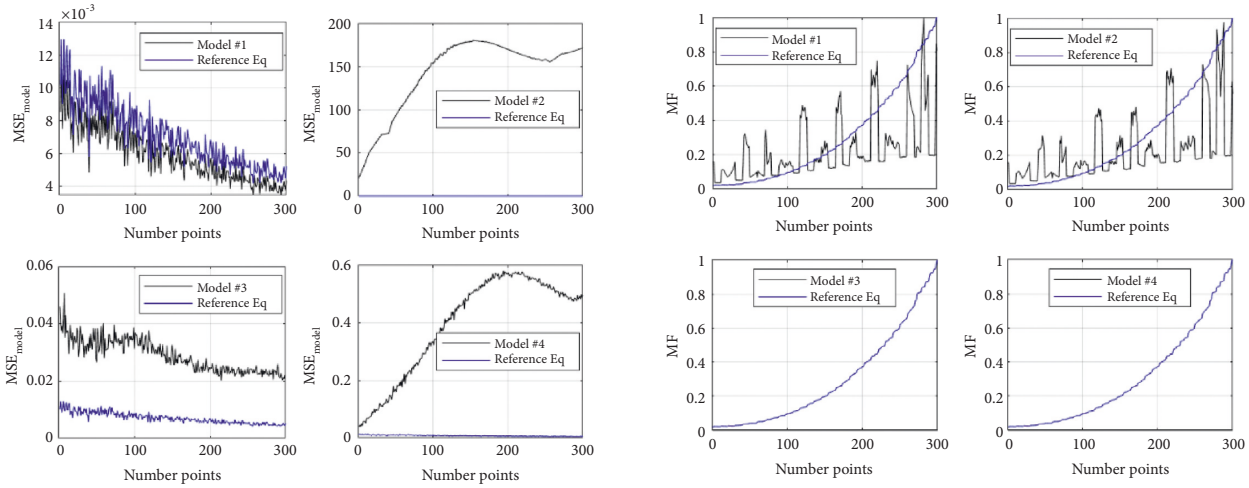


Figure 7: Plots of MSE and MF for the models of Table 1 in Section 3.

Assumptions, similar to the ones leading to (4), allow expressing the BIC criterion as follows:

$$BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n). \tag{16}$$

Even if the conceptual origins of BIC are different, the proposed changes have the same effects, namely, they improve BIC's discriminatory power by including more statistical information about the residuals and by better quantifying the models' complexity. In full analogy to (13), the final upgraded version of the BIC criterion is

$$BIC_{MICx} = n \ln[MSE(1 + MI)] + C_x \ln(n). \tag{17}$$

The tests of the AIC have been performed also for the BIC and they produce basically the same results. The discriminatory capability of BIC_{MICx} is clearly superior to the original version of the indicator, as can be seen in the plots of Appendix B. Of course, given the fact that BIC is based on Bayesian statistics, the argument that the implemented upgrades improve the coherence, with information-theoretic definitions and assumptions, cannot be made. On the other hand, the fact that the proposed modifications improve the quality also of a Bayesian type of selection criterion increases the confidence in the validity of the ideas, which have led to them.

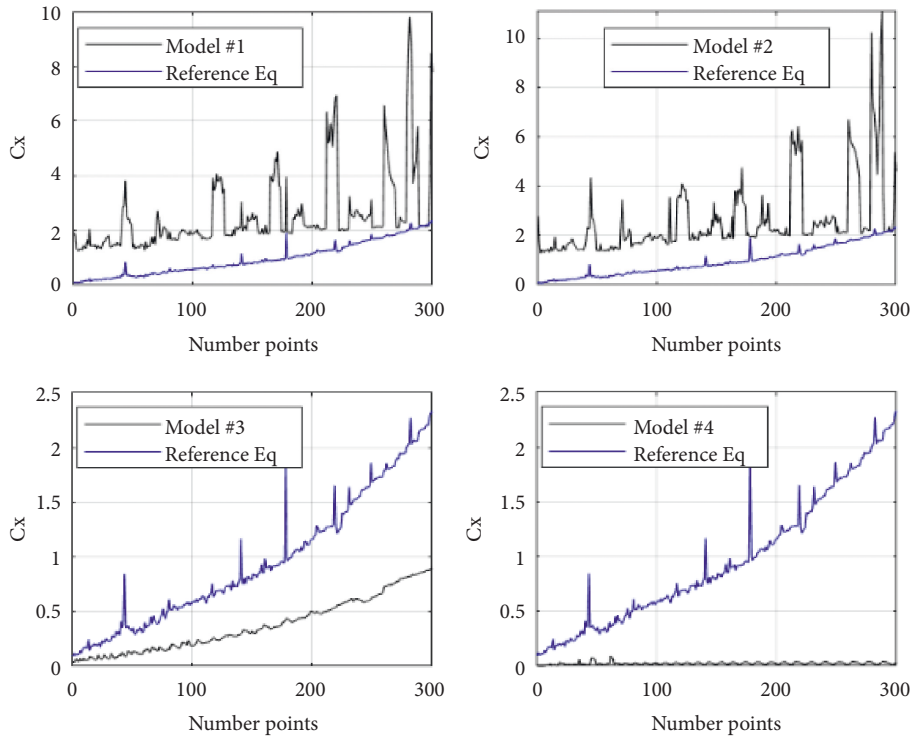


Figure 8: Plots of complexity Cx for the models of Table 1 in Section 3.

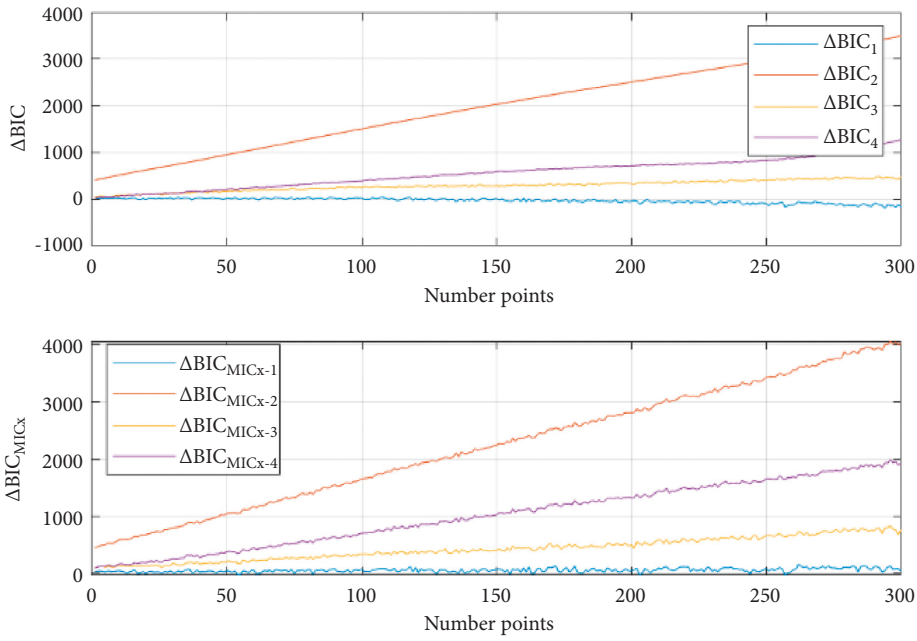


Figure 9: Comparison of the traditional BIC with the new BIC_{MICx} vs. the number of points for the models of Table 1 in Section 3. The plots show the indicator difference between the candidate models and the reference one; therefore negative values indicate that the corresponding indicator would have reached the wrong conclusion about the model to select.

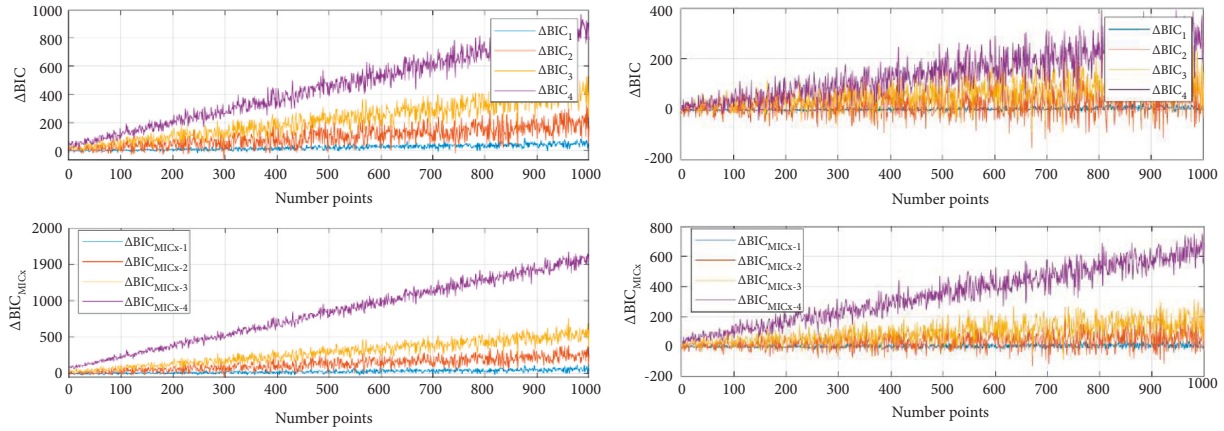


Figure 10: Top two plots: comparison of BIC and BIC_{MICx} for the case of a power law monomial multiplied by a squashing for 15% of Gaussian noise. Bottom: comparison of BIC and BIC_{MICx} for the case of a power law monomial multiplied by a squashing for 30% of Gaussian noise.

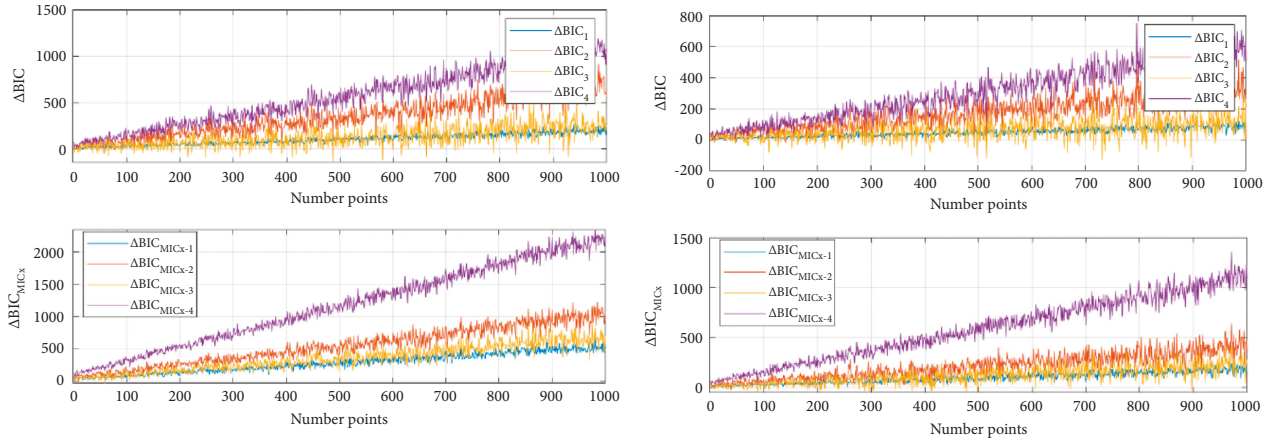


Figure 11: Top two plots: comparison of BIC and BIC_{MICx} for the case of an exponential function for 30% of Gaussian noise plus outliers. Bottom: comparison of BIC and BIC_{MICx} for the case of an exponential function for 60% of Gaussian noise plus outliers.

7. Conclusions

The Akaike Information Criterion was conceived to minimise the out-of-sample error and it is based on information theory. Statistical models are indeed developed to represent the process that generated the data, and the AIC estimates the relative amount of information lost by a given model. On this basis, it is assumed that the better a model, the less information it loses. Unfortunately, the deployment of AIC is problematic because its practical versions are affected by significant limitations. Indeed the most widely used version of AIC is valid under the assumptions that the data are affected by Gaussian, zero-sum additive noise. These hypotheses have to be accepted because, in most practical applications, it is often very difficult, if not impossible, to compute the likelihood of the data given the model. If the processes generating the data do not verify these assumptions, the traditional versions of the AIC can become poorly effective or even misleading.

On the other hand, other information theoretic quantities can be implemented to improve the discrimination potential of the criterion. In particular, the mutual information between the model estimates and the residuals can help reward the goodness of fit. The entropy in its turn can be used to quantify the model complexity. With these upgrades, the proposed version of the AIC has always proved to have much better convergence properties than the traditional version in all respects, including robustness against noise and zero-sum outliers. This has occurred in all the numerical tests performed, some of which consist of very challenging selection tasks, given the fact that some candidate models assume values very similar to the right one in the range covered by the data. The proposed improvements have an equally positive impact on the other criteria of the AIC family, such as TIC and AICc [4]. The extension of the same concepts to the Bayesian information criterion proves the soundness of the basic rationale behind the proposed modifications. The good performance in presence of non-normal noise distributions is particularly encouraging because model assessment in such situations has not yet received a lot of attention in the literature. Indeed, only a few publications have addressed the fact that many existing model selection criteria such as the BIC and Cp may not be suitable for generalized linear model regression, in which the

conditional mean and variance of the response are dependent [14]. Synergies with other formulations of the complexity term would also be very interesting from the methodological point of view [15].

Given the quite positive results obtained with synthetic data, proving their better discriminatory capability, the proposed new versions of the selection criteria are expected to become useful in various fields. They are already being deployed for the investigation of complex systems, ranging from high-temperature plasmas [16–23] to remote sensing of the atmosphere and radar [24–26]. Another promising application seems to be in support of the regularization of recent tomographic inversion methods [27–29]. In these fields, Dimensional Analysis (DA) is a methodology widely used to identify key variables based on physical dimensions. Even if it has been granted some attention recently, in most literature DA is treated as merely a preprocessing tool, creating various statistical problems [30]. The upgrades of the criteria proposed in this work could hopefully help in devising an appropriate statistical methodology that integrates DA and model selection.

Appendix

A. Calculation of the AICMICx and BICMICx Quantities of Section 3

The Figures 5–Figure 9 in this Appendix document all the quantities required to calculate AIC_{MICx} and BIC_{MICx} for the didactic case of Section 3, involving polynomial and sinusoidal models.

B. Performance Details of the BIC and BICMICx Quantities of Section 5

This Appendix documents the performance of BIC_{MICx} for the numerical cases described in Section 5: power laws multiplied by a squashing term and exponentials. Figures 10 and 11 show the comparison of BIC and $BIC_{MICx.}$

Data Availability

The Matlab scripts and data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

AM conceived this research; ML participated in the design of the code and interpretation of the results; ML and RR performed the validation of the analysis; AM and MG wrote the paper and participated in the revisions of it. MG provided the funding and supervised the project. All authors read and approved the final manuscript.

References

[1] F. Bailly and G. Longo, *Mathematics and the*, Natural Sciences" Imperial College Press, London, 2011.

[2] B. D'Espagnat, *On Physics and Philosophy*, Princeton University Press, Oxford, 2002.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] P. B. Kenneth and D. R. Anderson, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, Springer, Berlin, 2nd ed edition, 2002.

[5] G. Claeskens, "Statistical model choice" (PDF)," *Annual Review of Statistics and Its Application*, vol. 3, no. 1, pp. 233–256, 2016.

[6] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non-statisticians: A Step-By-Step Approach*, Wiley, Hoboken, 2009.

[7] A. Murari, E. Peluso, F. Cianfrani, P. Gaudio, and M. Lungaroni, "On the use of entropy to improve model selection criteria," *Entropy*, vol. 21, no. 4, p. 394, 2019.

[8] B. K. Natarajan, "On Learning sets and functions," *Machine Learning*, vol. 4, no. 1, pp. 67–97, 1989.

[9] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[10] M. Efatmaneshnik and M. J. Ryan, "A general framework for measuring system complexity," *Complexity*, vol. 21, no. 1, pp. 533–546, 2016.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, 2000.

[12] S. Bonamente, *Statistics and Analysis of Scientific Data"*, Graduate Texts in Physics) Springer Scince+Bussiness Media LLC, Berlin, 2017.

[13] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] X. Shen, H. C. Huang, and J. Ye, "Adaptive model selection and assessment for exponential family distributions," *Technometrics*, vol. 46, no. 3, pp. 306–317, 2004.

[15] A. Murari, R. Riccardo, and C. Teddy, "Alternative Definitions of Complexity for Practical Applications of Model Selection Criteria," *Defining and quantifying complexity*, vol. 2021, Article ID 8887171, 8 pages, 2021.

[16] A. Murari, I. Lupelli, P. Gaudio, M. Gelfusa, and J. Vega, "A statistical methodology to derive the scaling law for the H-mode power threshold using a large multi-machine database," *Nuclear Fusion*, vol. 52, no. 6, Article ID 063016, 2012.

[17] A. Murari, I. Lupelli, M. Gelfusa, and P. Gaudio, "Non-power law scaling for access to the H-mode in tokamaks via symbolic regression," *Nuclear Fusion*, vol. 53, no. 4, Article ID 043001, 2013.

[18] A. Murari, F. Pisano, J. Vega et al., "Extensive statistical analysis of ELMs on JET with a carbon wall," *Plasma Physics and Controlled Fusion*, vol. 56, no. 11, Article ID 114007, 2014.

[19] A. Murari, E. Peluso, M. Gelfusa, I. Lupelli, M. Lungaroni, and P. Gaudio, "Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form," *Plasma Physics and Controlled Fusion*, vol. 57, no. 1, Article ID 014008, 2015.

[20] A. Murari, E. Peluso, M. Lungaroni, M. Gelfusa, and P. Gaudio, "Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities," *Nuclear Fusion*, vol. 56, no. 2, Article ID 026005, 2015.

[21] A. Murari, M. Lungaroni, E. Peluso et al., "Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET," *Nuclear Fusion*, vol. 58, no. 5, Article ID 056002, 2018.

[22] F. P. Orsitto, A. Boboc, P. Gaudio et al., "Mutual interaction of Faraday rotation and Cotton-Mouton phase shift in JET polarimetric measurements," *Review of Scientific Instruments*, vol. 81, no. 10, Article ID 10D533, 2010.

[23] F. Romanelli and R. Kamendje, "Overview of JET results," *Nuclear Fusion*, vol. 49, no. 10, Article ID 104006, 2009.

[24] P. Gaudio, "New frontiers of forest fire protection: a portable laser system," *FfED WSEAS Transactions on Environment and Development*, vol. 9, no. 3, pp. 195–205, 2013.

[25] P. Gaudio, M. Gelfusa, A. Malizia, and M. Richetta, "Design and development of a compact Lidar/Dial system for aerial surveillance of urban areas," in *Proceedings of the SPIE - The International Society for Optical Engineering*, vol. 8894, Rome, Italy, September 2013.

[26] F. Xin, B. Wang, L. Shumin, X. Song, and W. Chi Hsu, "Adaptive radar waveform design based on weighted MI and the difference of two mutual information metrics," *Complexity*, vol. 2021, Article ID 8947450, 18 pages, 2021.

[27] T. Craciunescu, G. Bonheure, V. Kiptily, A. Murari, I. Tiseanu, and V. Zoita, "A comparison of four reconstruction methods for JET neutron and gamma tomography," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 605, no. 3, pp. 374–383, 2009.

[28] T. Craciunescu and A. Murari, "Geodesic distance on Gaussian manifolds for the robust identification of chaotic systems," *Nonlinear Dynamics*, vol. 86, no. 1, pp. 677–693, 2016.

[29] T. Craciunescu, E. Peluso, A. Murari, and M. Gelfusa, "Maximum likelihood bolometric tomography for the determination of the uncertainties in the radiation emission on JET TOKAMAK," *Review of Scientific Instruments*, vol. 89, no. 5, Article ID 053504, 2018.

[30] W. Shen and D. K. J. Lin, "A conjugate model for dimensional analysis," *Technometrics*, vol. 60, no. 1, pp. 79–89, 2017.

Hindawi
Complexity
Volume 2022, Article ID 9415863, 11 pages
https://doi.org/10.1155/2022/9415863

Research Article

Extreme Gradient Boosting Algorithm for Predicting Shear Strengths of Rockfill Materials

Mahmood Ahmad,[1,2] **Ramez A. Al-Mansob**,[1] **Kazem Reza Kashyzadeh**,[3] **Suraparb Keawsawasvong**,[4] **Mohanad Muayad Sabri Sabri**,[5] **Irfan Jamil**,[6] **and Arnold C. Alguno**[7]

[1]*Department of Civil Engineering, Faculty of Engineering, International Islamic University Malaysia, Jalan Gombak, Selangor 50728, Malaysia*
[2]*Department of Civil Engineering, University of Engineering and Technology Peshawar (Bannu Campus), Bannu 28100, Pakistan*
[3]*Department of Transport, Academy of Engineering, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Street, Moscow 117198, Russia*
[4]*Department of Civil Engineering, Thammasat School of Engineering, Thammasat University, Pathumthani 12120, Thailand*
[5]*Peter the Great St. Petersburg Polytechnic University, Saint Petersburg 195251, Russia*
[6]*Department of Civil Engineering, University of Engineering and Technology Peshawar, Peshawar 25000, Pakistan*
[7]*Department of Physics, Mindanao State University-Iligan Institute of Technology, Iligan City 9200, Philippines*

Correspondence should be addressed to Mahmood Ahmad; ahmadm@uetpeshawar.edu.pk

Received 3 June 2022; Revised 19 July 2022; Accepted 20 July 2022; Published 24 August 2022

Academic Editor: Andrea Murari

For the safe and economical construction of embankment dams, the mechanical behaviour of the rockfill materials used in the dam's shell must be analyzed. The characterization of rockfill materials with specified shear strength is difficult and expensive due to the presence of particles greater than 500 mm in diameter. This work investigates the feasibility of using an extreme gradient boosting (XGBoost) computing paradigm to estimate the shear strength of rockfill materials. To train and validate the proposed XGBoost model, a total of 165 databases obtained from the literature are chosen. The XGBoost model was compared against support vector machine (SVM), adaptive boosting (AdaBoost), random forest (RF), and K-nearest neighbor (KNN) models described in the literature. XGBoost beats SVM, RF, AdaBoost, and KNN models in terms of performance evaluation metrics such as coefficient of determination (R^2), Nash–Sutcliffe coefficient (NSE), and error in the root mean square ratio (RMSE) to the standard deviation of the measured data (RSR). The results demonstrated that the XGBoost model has the highest prediction performance with ($R^2 = 0.9707$, NSE = 0.9701, and RSR = 0.1729), followed by the SVM model with ($R^2 = 0.9655$, NSE = 0.9639, and RSR = 0.1899), RF ($R^2 = 0.9545$, NSE = 0.9542, and RSR = 0.2140), the AdaBoost model with ($R^2 = 0.9390$, NSE = 0.9388, and RSR = 0.2474) and the KNN model with ($R^2 = 0.6233$, NSE = 0.6180, and RSR = 0.6181). A sensitivity analysis has been conducted to ascertain the impact of each investigated input parameter. This study demonstrates that the established XGBoost model for estimating the shear strength of rockfill materials is reliable.

1. Introduction

Rockfill materials (RFM) are commonly used in the construction of high embankment dams in order to harness natural water resources. RFM is comprised of gravels, cobbles, and boulders obtained by blasting rock quarries or natural riverbeds. Material from riverbeds is rounded to subrounded, and material from quarries is angular to subangular. Mineral composition, particle size, shape, gradation, individual particle strength, void content, relative density (RD), and particle surface roughness all influence the behaviour of these RFMs used in the construction of rockfill

dams. Therefore, it is essential to comprehend and characterise the behaviour of these materials for the study and safe construction of rockfill dams.

In engineering practice, the particle size of rockfill materials ranges from 400 to 600 millimetres and can exceed 1000 millimetres. Due to the constraints of laboratory testing equipment, rockfill materials that exceed the maximum permissible particle size must be scaled. To determine the mechanical properties of rockfill materials on-site, analog simulation is used in laboratory testing to build test specimens with the same internal structure as the prototype rockfill materials, thus determining the engineering characteristics of the prototype rockfill materials. Several research studies have investigated the behaviour of the RFM such as Abbas et al. [1], Gupta [2], Venkatachalam [3], Marsal [4], Mirachi [5], and Honkanadavar and Sharma [6] and carried out laboratory experiments on different RFMs, and it was revealed that their stress-strain behaviour is dependent on the stress level, but nonlinear and inelastic. They also reported that the angle of internal friction increases as the maximum particle size of riverbed RFM increases, while the opposite trend is true for quarry RFM. Frossard et al. [7] proposed a rational approach for estimating RFM shear strength based on size effects; Honkanadavar and Gupta [8] developed a power law for the relationship between the shear strength parameter and various riverbed RFM index features due to the difficulty of conducting large-scale strength testing and defining the mechanical behaviour of RFMs. Numerous methodologies have been developed to anticipate the behaviour of such soils. Large particle size RFM cannot be tested under laboratory circumstances as maximum large-scale shear tests are time-consuming and complicated, and it is hard to predict the nonlinear shear strength function without an analytical method (particle size 1200 mm) [8].

Over the last ten years, a newly developed approach based on machine learning (ML) algorithms has been widely applied to solve real-world problems, particularly civil engineering. Numerous practical problems have been effectively addressed using ML techniques, paving the way for many promising opportunities in civil engineering and other fields such as environmental [9] and geotechnical [10–15] including prediction of RFM shear strength [16–18]. In this context, the artificial neural network (ANN) approach is utilized by Kaunda [16] for estimating RFM shear strength. Cubist and random forest regression techniques are used by Zhou et al. [17], and they found that both models are accurate for RFM shear strength estimations than ANN and traditional regression models. Ahmad et al. [18] used support vector machine (SVM), random forest (RF), AdaBoost, and K-nearest neighbor (KNN) algorithms to estimate the shear strength of RFM and concluded that the SVM model achieved a better prediction performance compared to the RF, AdaBoost, and KNN models. This field, however, is currently being investigated. The article aims to provide the following contributions in the research field:

(i) To evaluate the predictive capacity of the XGBoost algorithm for the shear strength of RFM

(ii) To compare the proposed model to the reference models used in the published literature

(iii) Conduct sensitivity analysis to assess the influence of each input parameter on the RFM's shear strength

The structure of the paper is as follows: The theory of extreme gradient boosting is explained in Section 2. Data collection and correlation analysis are presented in Section 3. Section 4 explains the performance measurement employed. Section 5 presents the obtained results and a discussion of them. Finally, conclusions based on the achieved results are provided.

2. Extreme Gradient Boosting (XGBoost)

Chen and Guestrin [19] proposed the sophisticated supervised technique extreme gradient boosting (XGBoost) under the gradient boosting framework which has received widespread recognition in Kaggle machine learning contests due to its advantages of high efficiency and considerable flexibility. XGBoost's loss function adds a regularization term to the objective function, which helps to smoothen the final learning weights and avoid over-fitting [19]. It also optimizes the loss function using first and second-order gradient statistics. XGBoost also supports row and column sampling to address this issue in addition to providing regular terms to prevent over-fitting. As a result of the parallel and distributed computation, faster model exploration is possible.

The following is a description of the XGBoost algorithm [20]: given a dataset with n examples and m features $D = \{(x_i, y_i)\}$ $(|D| = n, x_i \in R^m, y_i \in R)$,$K$ additive functions will be used to predict the output values of a tree ensemble model as follows:

$$\widehat{y}i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{1}$$

where F is the regression trees space. It is calculated as

$$F = \left\{f(x) = \omega_q(x)\right\}\left(q: R^m \longrightarrow T, \omega_q \in R^T\right), \tag{2}$$

where q represents for the structure of each tree, T represents for the number of leaves in the tree, and f_k is a function that corresponds to an independent tree structure q and leaf weights ω. To reduce errors of ensemble trees, the objective function is found in the XGBoost model:

$$L^{(t)} = \sum_{i=1}^{n} l\left(\left(y_i, \widehat{y}_j^{(t-1)}\right) + f_t(x_i)\right) + \Omega(f_k), \tag{3}$$

where l is a differentiable convex objective function to calculate the error between predicted and measured values; y_i and $\widehat{y}_i$ are regulated and predicted values, respectively; t shows the repetitions in order to minimize the errors; and Ω is the complexity penalized with the regression tree functions:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \quad (4)$$

ω is the vector of the score for the blades, and γ the minimal loss required for the further isolation of a blade node. λ is the regularization function. In addition, γ and λ are parameters which are able to control the complexity of the tree, and the regularization term helps to avoid overfitting by smoothening the final learnt weights. Taylor expansion is applied to the objective function in order to further simplify it as

$$F = \sum_{i=1}^{m}\left[f_t(x_i)g_i + \frac{1}{2}(f_t(x_i))^2 h_i\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2, \quad (5)$$

where g_i and h_i are the first and second derivatives obtained on the loss function, respectively. More detailed explanations of the XGBoost algorithm can be found in Chen and Guestrin's [19] research paper.

3. Dataset Collection and Correlation Analysis

In this study, a database of 165 samples of RFM shear strength reports was collected from Kaunda [16] and is presented in Appendix A and Table A1 in supplementary file. All input parameters that might influence the shear strength results of RFM were considered. The included parameters are D_{10}, D_{30}, D_{60}, and $D_{90,}$ corresponding to the 10%, 30%, 60%, and 90% sieve sizes passing, respectively. C_c and C_u refer to the curvature uniformity coefficients (C_c), respectively; FM and GM describe fineness modulus and gradation modulus, respectively; R represents International Society of Rock Mechanics (ISRM) hardness rating; $UCS_{min,}$ and UCS_{max} (MPa) signify the uniaxial compression strengths boundaries (MPa); and γ represents the dry unit weight (kN/m^3), while σ_n is the normal stress (MPa). The considered output is the shear strength of RFM (MPa) (denoted as τ (MPa)). The summary of the database statistics is presented in Table 1, which includes the boundary and standard deviation values of all parameters used in this study.

Correlation (ρ) was used to verify the intensity of correlation between different parameters (see Figure 1). For a given pair of random variables (m, n), the following equation for ρ is used:

$$\rho(m,n) = \frac{\mathrm{cov}(m,n)}{\sigma_m\sigma_n}, \quad (6)$$

where cov denotes covariance, σ_m denotes the standard deviation of m, and σ_n denotes the standard deviation of n. $|\rho| > 0.8$ represents a strong correlation between m and n, values between 0.3 and 0.8 represents a moderate relationship, and $|\rho| > 0.30$ represents a weak relationship [21]. As per Song et al. [22], correlation is considered as "strong" if $|\rho| > 0.8$. In the order of strong to weak, the relationships between input and output parameters are represented in Figure 1. Consequently, no factors from the estimation model's τ were deleted. The correlation coefficient has a maximum absolute value of 0.97, as shown in Figure 1.

4. Evaluation and Prediction

To evaluate the predictive capacity of the XGBoost algorithm, we compared it with some other machine learning methods developed in literature using performance measures.

4.1. Compared Machine Learning (ML) Methods. The XGBoost model was compared with other prediction methods such as support vector machine, adaptive boosting, random forest, and K-nearest neighbor proposed in literature. A brief description of each technique is presented. For a more in-depth discussion, the reader is referred to the relevant references.

4.1.1. Support Vector Machine (SVM). The Support Vector Machine (SVM) regression technique relies on feature classification and generates an interclass hyperplane and minimizes the vector lengths and variance between the features and the plane. The SVM is compatible with the majority of kernel types, including Euclidean, Gaussian, Exponential, and Dirichlet kernels [23]. The objective function for SVM regression contains a coefficient generated from the cost analysis that aids in determining the flatness of the created hyperplane [24]. This allows the user to change the SVM technique to fit unique datasets.

4.1.2. Adaptive Boosting (AdaBoost). Adaptive Boosting is a boosting machine learning technique in which strong learning algorithms augment weak learning algorithms. AdaBoost must define the number of beginning students (n) as a parameter [25]. During the training phase, AdaBoost develops learners with low accuracy who improve based on their predecessors [26]. Using this method, the AdaBoost dynamically modifies the training weight based on the performance of the fundamental learning algorithms [27].

4.1.3. Random Forest (RF). Random Forests are ensemble models that use many decision trees as base-learners to obtain more precise outcomes. Individual trees are generated from training data using random parameters as their roots and nodes using the bootstrap sampling method [28]. Multiple decision trees are more stable than a single tree because they reduce overfitting and average the outcomes [26]. The number of trees in the forest at each binary node, the number of randomly selected predictors, and the lowest number of observations at the nodes of the trees are the three primary parameters for random forests [29].

4.1.4. K-nearest Neighbor (KNN). The supervised KNN is a machine learning algorithm that can be used to tackle both classification and regression problems. In regression problems, the input data set is comprised of k that is most similar to the training data sets utilized in the highlighted set. The outcome of KNN regression is the object's characteristic value, which is the mean value of k's nearest neighbors. As

Table 1: Statistics of parameters of the training and testing datasets.

Statistical parameter	Dataset	Input variable													Output variable
		D_{10} (mm)	D_{30} (mm)	D_{60} (mm)	D_{90} (mm)	C_C	C_U	GM	FM	R	UCS_{min} (MPa)	UCS_{max} (MPa)	γ (kN/m^3)	σ_n (MPa)	τ (MPa)
Minimum	Total data	0.010	0.560	1.200	2.600	0.100	1.360	0.200	3.000	1.000	1.000	5.000	9.320	0.002	0.005
	Training	0.010	0.560	1.200	2.600	0.100	1.360	0.200	3.000	1.000	1.000	5.000	9.320	0.002	0.005
	Testing	0.010	0.560	1.200	2.600	0.100	1.470	0.200	3.000	1.000	1.000	5.000	9.320	0.021	0.024
Maximum	Total data	33.900	42.400	80.100	100.000	22.270	1040.000	6.000	8.800	6.000	250.000	400.000	38.900	4.205	3.921
	Training	33.900	42.400	80.100	100.000	22.270	1040.000	6.000	8.800	6.000	250.000	400.000	38.900	4.205	3.921
	Testing	33.900	42.400	50.000	99.000	22.270	1040.000	6.000	8.800	5.000	100.000	250.000	38.900	3.223	2.492
Mean	Total data	4.463	7.860	18.280	39.927	2.404	69.561	2.903	6.142	4.327	73.691	168.455	20.799	0.734	0.662
	Training	4.867	8.465	19.287	40.386	2.199	53.324	2.788	6.250	4.364	75.045	170.682	20.766	0.729	0.660
	Testing	2.887	5.442	14.252	38.091	3.226	134.510	3.365	5.709	4.182	68.273	159.545	20.932	0.756	0.668
Standard deviation	Total data	8.875	10.335	14.420	22.432	3.414	193.628	1.278	1.298	0.957	37.975	87.844	4.861	0.785	0.652
	Training	9.179	10.577	15.135	22.018	3.075	156.064	1.243	1.261	0.910	39.230	88.010	4.605	0.780	0.662
	Testing	7.453	9.050	10.349	24.289	4.492	194.958	1.331	1.374	1.131	32.444	87.967	5.854	0.816	0.619

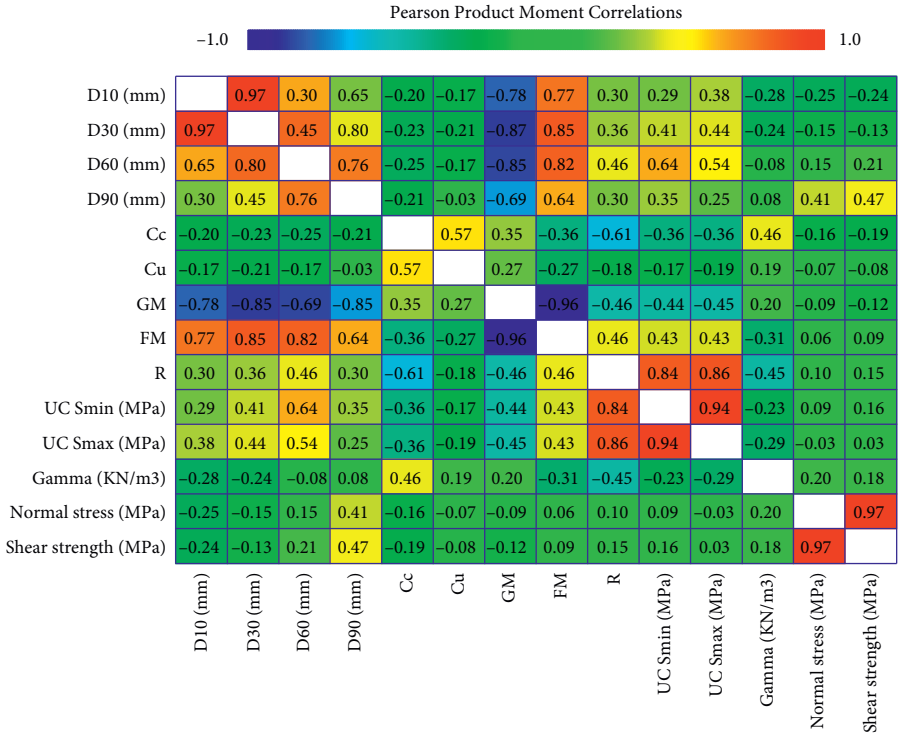


Figure 1: Correlation coefficient between parameters.

the distance metric, a parameter such as Euclidean or Mahalanobis distance can be utilized to locate the k of a data point [30].

4.2. Evaluation Measures. Three quantitative statistical indices, i.e., coefficient of determination (R^2), error in the root mean square ratio to the measured data standard deviation (RSR), and Nash–Sutcliffe coefficient (NSE) were employed to validate and compare the XGBoost model. The following equations characterise the supplied indices:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2}, \tag{7}$$

$$RSR = \frac{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}, \tag{8}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{9}$$

where n is the total number of data; y_i and $\hat{y}_i$ are the actual shear strength and the predicted shear strength, respectively; and $\overline{y}$ is the mean of the actual shear strength.

Values of the coefficient of determination (R^2) that are closer to 1 imply that this model better fits the data. When R^2 is greater than 0.8 and close to 1, the model is deemed robust [31]. The NSE is a normalized statistic that regulates the level of residual variance compared to the variance of the data being measured [32]. The NSE scale ranges from $-\infty$ to 1, with 1 denoting an ideal match. If the NSE value is greater than 0.65, a strong correlation exists [32, 33]. The root mean square error (RMSE)-standard deviation ratio (RSR) is computed by dividing the RMSE by the standard deviation of the observed data. The RSR varies from 0, representing the optimal value, to a significant positive value. The RSR ranges from the optimal value of 0 to a substantial positive number. Classification ranges are expressed as very good, good, acceptable, and unacceptable. The RSR ranges are $0.000 \le \text{RSR} \le 0.500$, $0.500 \le \text{RSR} \le 0.600$, $0.600 \le \text{RSR} \le 0.700$, and $\text{RSR} > 0.700$, respectively [34].

5. Methodology

The present study is carried out based on the proposed framework that involves four main steps as follows: (1) data preparation and correlation analysis, (2) development of the model, (3) validation of the proposed model, and (4) sensitivity analysis (Figure 2):

(1) Data preparation and correlation analysis: In this first step, the data of samples from the laboratory were utilized to build the training and testing datasets. The training dataset was constructed using 80% of the total data, while the testing dataset was built from the remaining 20%.

(2) Development of the model: In this second step, the training dataset was applied for training the model based on the XGBoost algorithm. The optimization of user defined parameters is undertaken by carrying out multiple runs with these parameters on the training data and analyzing the performance of the

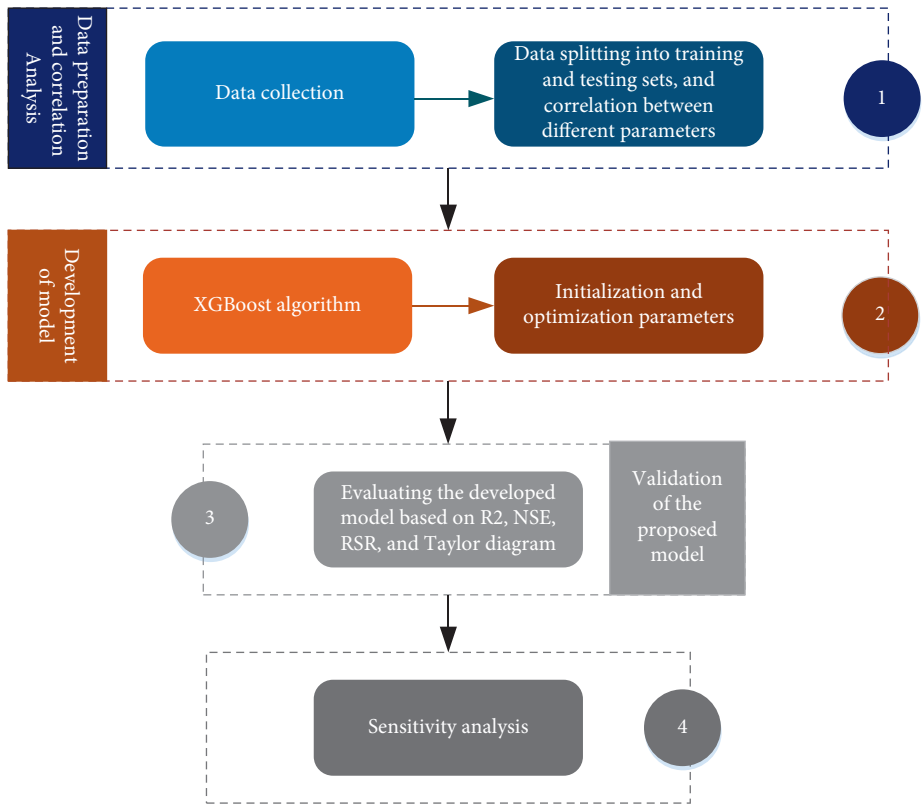


Figure 2: Flowchart illustrates the proposed methodology for present study.

resulting models on testing data. All training and testing operations were conducted out in Orange software.

(3) Validation of the proposed models: In this third step, the testing dataset was adopted for validating the proposed models. Statistical indices including R^2, NSE, and RSR were applied to validate the models. The proposed model is compared to the reference models used in the published literature. Furthermore, Taylor diagram is utilized to illustrate how similar the models (including the proposed XGBoost) are to the reference/observed point position.

(4) Sensitivity analysis: In the last step, sensitivity analysis is used for evaluating the influence of input factors on the shear strength of rockfill material.

6. Results and Discussion

The proposed model that estimates the RFM shear strength is developed using orange software. The predictor variables were provided via an input set (x) defined by $x = [D_{10}, D_{30}, D_{60}, D_{90}, C_c, C_u$, GM, FM, R, $\mathrm{UCS}_{\min}$, $\mathrm{UCS}_{\max}$, γ, and $\sigma_n]$, while the target variable (y) is shear strength (τ) of the rockfill material. Every modelling stage requires the selection of the suitable size of training and testing datasets. Consequently, 80% (132 cases) of the total data were employed to generate models while the remaining 20% (33 cases) of the data were used to test the developed models in this study. The XGBoost model was tuned through trial and error to get an optimal hyperparameters values owing to accurate estimate of the shear strength of rockfill materials. This study optimizes some essential XGBoost parameters and clarifies the definitions of these hyperparameters. The tuning parameters for the model were selected and then changed during the trials until the best metrics from Table 2 were obtained.

The predictive performance of the training and testing datasets is shown in regression form in Figure 3. In terms of training, the XGBoost model produced the best prediction results (i.e., $R^2 = 0.9707$, NSE = 0.9701 and RSR = 0.1729) compared to SVM (i.e., $R^2 = 0.9655$, NSE = 0.9639 and RSR = 0.1899), RF (i.e., $R^2 = 0.9545$, NSE = 0.9542, and RSR = 0.2140), AdaBoost (i.e., $R^2 = 0.9390$, NSE = 0.9388, and RSR = 0.2474), and KNN (i.e., $R^2 = 0.6233$, NSE = 0.6180, and RSR = 0.6181). It is also verified by the findings of R^2, NSE, and RSR in Figure 4 as XGBoost produced lesser RSR, higher R^2, and NSE values compared to SVM, RF, AdaBoost, and KNN models developed in the literature by Ahmad et al. [18] and the parameter optimization is presented in Table 2.

As depicted in Figure 4, the XGBoost model performed the best in terms of R^2, NSE, and RSR (i.e., $R^2 = 0.9676$, NSE = 0.9672, and RSR = 0.1812) compared to SVM (i.e., $R^2 = 0.9656$, NSE = 0.9654, and RSR = 0.1861), RF (i.e., $R^2 = 0.9656$, NSE = 0.9164, and RSR = 0.2891), AdaBoost (i.e., $R^2 = 0.9181$, NSE = 0.8835, and RSR = 0.3414), and KNN (i.e., $R^2 = 0.6304$, NSE = 0.6076, and RSR = 0.6264) in the testing phase. The outcomes of this and a prior study by Ahmad et al. [18] (see Figure 4) demonstrate that the ML method may accurately predict the shear strength of RFMs. The

Table 2: Parameter configuration.

Algorithm	Parameter optimization
XGBoost	n estimators = 40, learning rate = 0.250, maximum depth = 4
SVM	Cost = 8, regression loss epsilon = 0.1, kernel type = radial basis function
RF	Number of trees = 15, limit depth of individual trees = 3
KNN	Number of neighbors = 5, metric = euclidean, weight = uniform
AdaBoost	Number of estimators = 2, learning rate = 0.1, boosting algorithm = SAMME, regression loss function = linear

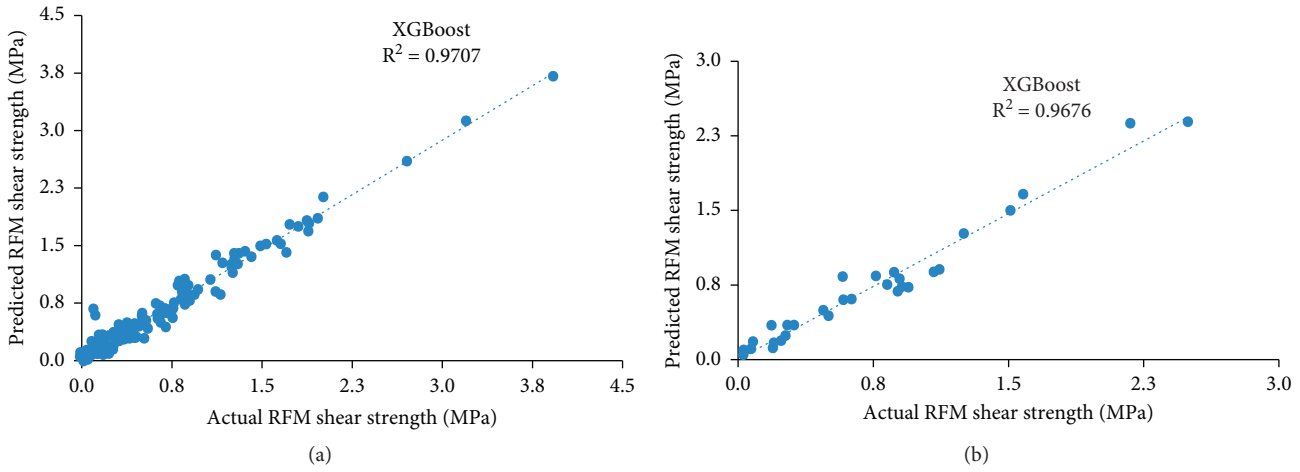


Figure 3: Regression graph of the XGBoost model for (a) training and (b) testing datasets.

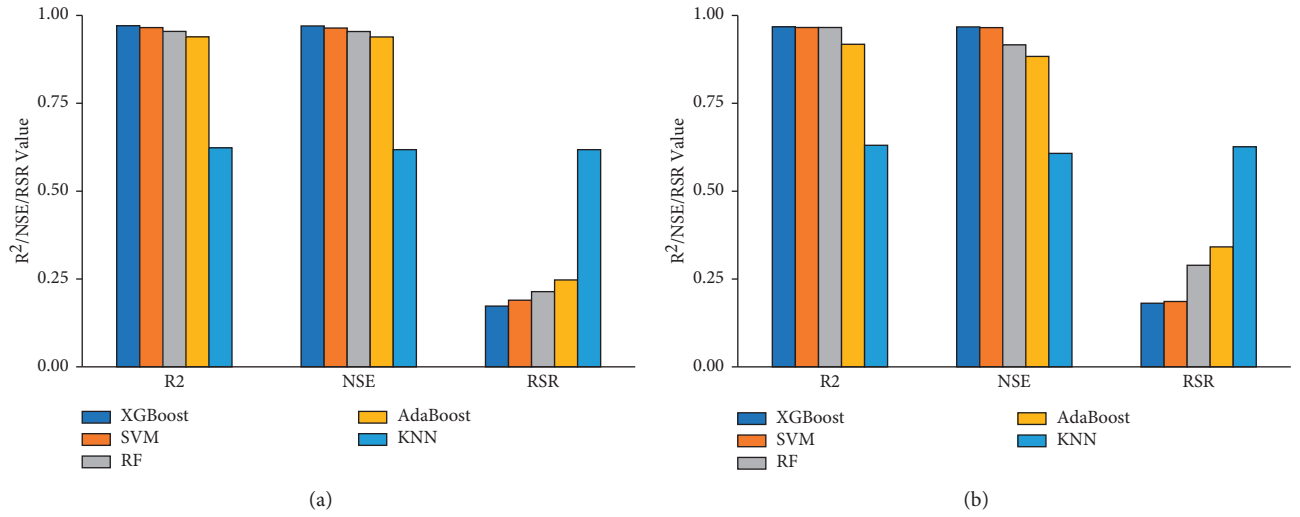


Figure 4: Comparison of R^2, NSE, and RSR values from the XGBoost, SVM, RF, AdaBoost, and KNN models in (a) training; and (b) testing phases.

comparison of study outcomes makes sense because the data sets and inputs are the same. In contrast, the XGBoost model beats the other models in terms of predictive performance and offered a balanced prediction throughout the training and testing data sets. In addition, due to the study's small data set, additional research on other data sets is necessary to establish the most generic model for predicting the shear strength of RFM.

The difference between the actual and predicted shear strength of RFM is represented in Figure 5 by comparing the results of the training and testing sets. The proposed XGBoost model is satisfactory for predicting the RFM shear strength, barring a few noise points.

Taylor diagram (see Figure 6) is utilized to illustrate how similar the models (including the proposed XGBoost) are to the reference/observed point position based on their correlation, root-mean-square error difference, and amplitude of their variations (represented by their standard deviations). The better the performance, the closer each model point is to the position of the reference/observed point. In terms of predictive ability, the proposed XGBoost model beats the SVM, RF, AdaBoost, and KNN models developed in the literature by Ahmad et al. [18].

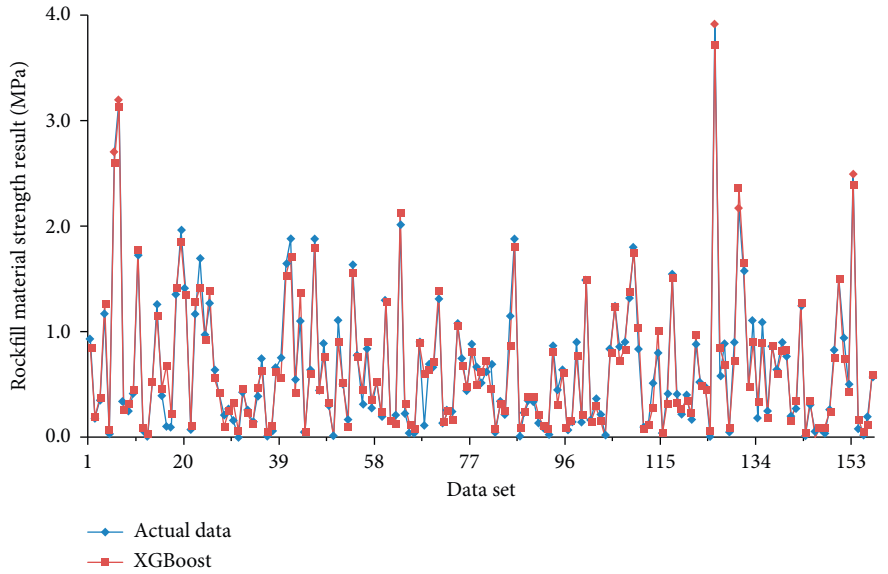


Figure 5: Results of XGBoost model training and testing phases for rockfill material shear strength.

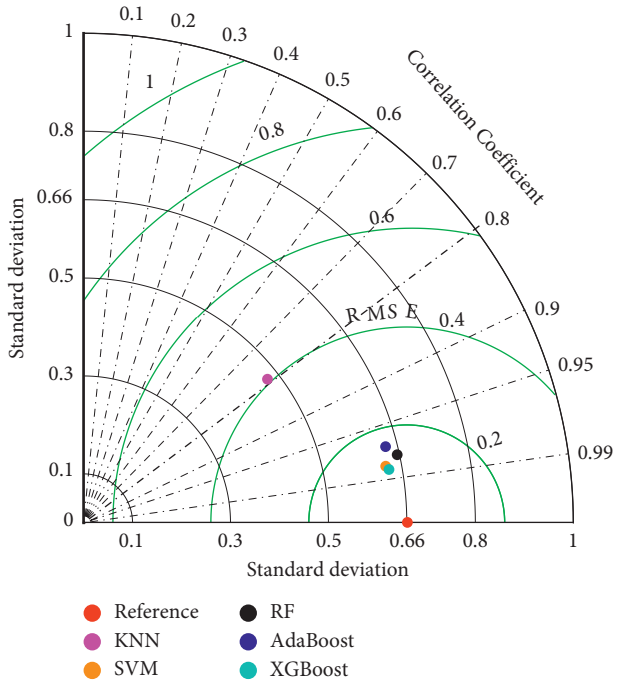


Figure 6: Taylor diagram of the models.

The sensitivity results of the XGBoost model were evaluated utilising Yang and Zang's [35] approach for evaluating the influence of input factors on the shear strength of rockfill material. This approach, which has been the topic of numerous studies [36–41], is as follows:

$$r_{ij} = \frac{\sum_{m=1}^{n} (y_{im} \times y_{om})}{\sqrt{\sum_{m=1}^{n} y_{im}^2 \sum_{m=1}^{n} y_{om}^2}}, \tag{10}$$

where n represents the number of values (i.e., 132); y_{im} and y_{om} denotes input and output variables, respectively. For each input parameter, the r_{ij} value ranges from zero to one, with the greatest r_{ij} values indicating the efficient output variable (i.e., τ). Figure 7 shows the r_{ij} scores for all input variables and demonstrates that σ_n ($r_{ij} = 0.99$) has the greatest effect on the shear strength of rockfill material. Furthermore, Figure 1 shows that the normal stress σ_n has the highest ρ of 0.97 in all other parameters validating the sensitivity analysis results.

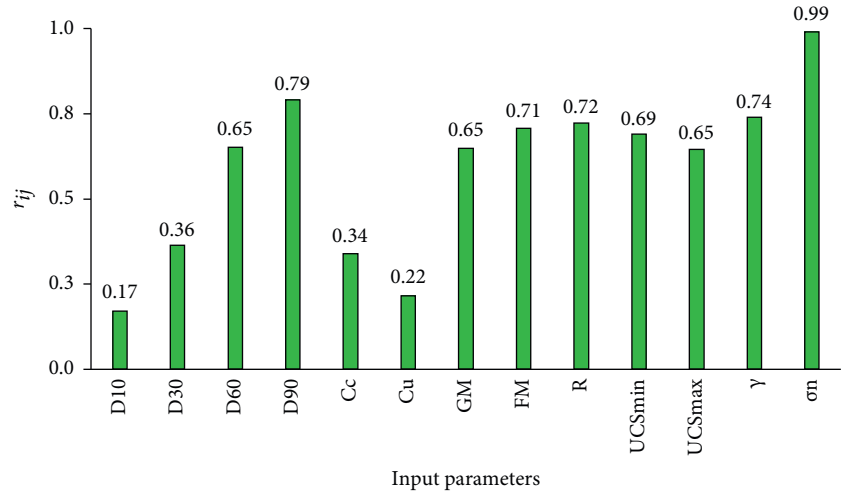


Figure 7: Sensitivity analysis results.

7. Conclusions

Using an XGBoost algorithm, a new prediction model for RFM shear strength is proposed in the current study. Comparisons reveal that the proposed XGBoost model provides the most accurate prediction of the RFM's shear strength when compared to the algorithms developed using the SVM, RF, AdBoost, and KNN model. Important findings found from this study include as follows:

(1) In the test phase, results showed that the XGBoost had the highest power performance ($R^2 = 0.9676$, NSE = 0.9672, and RSR = 0.1812) compared to other machine learning models. Furthermore, based on the scatter plots of actual and predicted values, the XGBoost model exhibited a better fit to the observed data, indicating that it has potential for broader applications in RFM material properties prediction.

(2) Compared to SVM, RF, AdaBoost, and KNN models in the literature, the proposed XGBoost model has a superior predictive capability. In addition, the proposed model is amenable to further modification so that the accumulation of further data will considerably enhance its predictive potential.

(3) The findings of the sensitivity analysis indicate that five parameters, namely, the normal stress, the 90% passing sieve diameters (D_{90}), the dry unit weight, and the ISRM hardness rating, are the most sensitive and important factors for estimating the shear strength of rockfill materials.

(4) The developed XGBoost model gives predictions with the same level of accuracy as existing soft computing methods.

Since the proposed XGBoost model produces predictions based on the input values, interpolation between the input variables is more accurate and reliable than extrapolation. Therefore, the model should not be used for input parameter values beyond the defined range of the study.

Data Availability

The data presented in this study are available in Appendix A, Table A1 (see supplementary file).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The research was partially funded by the Ministry of Science and Higher Education of the Russian Federation under the strategic academic leadership program "Priority 2030" (Agreement 075-15-2021-1333 dated 30.09.2021).

Supplementary Materials

Table A1. Dataset used in the development and validation of the model. (*Supplementary Materials*)

References

[1] S. Abbas, A. Varadarajan, and K. Sharma, "Prediction of shear strength parameter of prototype rockfill material," *IGC-2003, Roorkee*, vol. 1, pp. 5–8, 2003.

[2] A. K. Gupta, "Constitative Modelling of Rockfill Materials," vol. 6, 2000.

[3] K. Venkatachalam, "Prediction of Mechanical Behaviour of Rockfill Materials," vol. 15, 1993.

[4] R. J. Marsal, "Large scale testing of rockfill materials," *Journal of the Soil Mechanics and Foundations Division*, vol. 93, no. 2, pp. 27–43, 1967.

[5] N. D. Marachi, "Strength and deformation, Characteristics of Rockfill Materials," *Report No. TE-69-5 to State of California Department of Water Resources*, U S A, 1969.

[6] N. P. Honkanadavar and K. G. Sharma, "Testing and modeling the behavior of riverbed and blasted quarried rockfill materials," *International Journal of Geomechanics*, vol. 14, no. 6, 2014.

[7] E. Frossard, W. Hu, C. Dano, and P.-Y. Hicher, "Rockfill shear strength evaluation: a rational method based on size effects," *Géotechnique*, vol. 62, no. 5, pp. 415–427, 2012.

[8] N. Honkanadavar and S. Gupta, "Prediction of shear strength parameters for prototype riverbed rockfill material using index properties," *Proceedings of Indian Geotechnical Conf*, vol. 55, pp. 335–338, —2010.

[9] A. Froemelt, D. J. Dürrenmatt, and S. Hellweg, "Using data mining to assess environmental impacts of household consumption behaviors," *Environmental Science & Technology*, vol. 52, no. 15, pp. 8467–8478, 2018.

[10] A. Mahmood, X.-W. Tang, J.-N. Qiu, W.-J. Gu, and A. Feezan, "A hybrid approach for evaluating CPT-based seismic soil liquefaction potential using Bayesian belief networks," *Journal of Central South University*, vol. 27, no. 2, pp. 500–516, 2020.

[11] M. Ahmad, X.-W. Tang, J.-N. Qiu, and F. Ahmad, "Evaluating seismic soil liquefaction potential using bayesian belief network and C4. 5 decision tree approaches," *Applied Sciences*, vol. 9, no. 20, p. 4226, 2019.

[12] M. Ahmad, X. Tang, J. Qiu, F. Ahmad, and W. Gu, "LLDV-A comprehensive framework for assessing the effects of liquefaction land damage potential," in *Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering*, pp. 527–533, ISKE), Dalian, China, November 2019.

[13] M. Ahmad, X.-W. Tang, J.-N. Qiu, F. Ahmad, and W.-J. Gu, "A step forward towards a comprehensive framework for assessing liquefaction land damage vulnerability: exploration from historical data," *Frontiers of Structural and Civil Engineering*, vol. 14, no. 6, pp. 1476–1491, 2020.

[14] M. Ahmad, X. Tang, and F. Ahmad, "Evaluation of liquefaction-induced settlement using random forest and REP tree models: taking pohang earthquake as a case of illustration," *Natural Hazards-Impacts, Adjustments & Resilience, IntechOpen*, vol. 44, 2020.

[15] M. Ahmad, N. A. Al-Shayea, X.-W. Tang, A. Jamal, H. M Al-Ahmadi, and F. Ahmad, "Predicting the pillar stability of underground mines with random trees and C4. 5 decision trees," *Applied Sciences*, vol. 10, no. 18, p. 6486, 2020.

[16] R. Kaunda, "Predicting shear strengths of mine waste rock dumps and rock fill dams using artificial neural networks," *International Journal of Mining and Mineral Engineering*, vol. 6, no. 2, p. 139, 2015.

[17] J. Zhou, E. Li, H. Wei, C. Li, Q. Qiao, and D. J. Armaghani, "Random forests and cubist algorithms for predicting shear strengths of rockfill materials," *Applied Sciences*, vol. 9, no. 8, p. 1621, 2019.

[18] M. Ahmad, P. Kamiński, P. Olczak et al., "Development of prediction models for shear strength of rockfill material using machine learning techniques," *Applied Sciences*, vol. 11, no. 13, p. 6167, 2021.

[19] T. Chen and C. X. Guestrin, "A scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, California CA U S A, August 2016.

[20] H. Nguyen, X.-N. Bui, H.-B. Bui, and D. T. Cuong, "Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study," *Acta Geophysica*, vol. 67, no. 2, pp. 477–490, 2019.

[21] T. van Vuren, *Modeling of Transport Demand–Analyzing, Calculating, and Forecasting Transport Demand: By VA Profillidis and GN Botzoris*, p. 472, Elsevier, Amsterdam, 2018.

[22] Y. Song, J. Gong, S. Gao et al., "Susceptibility assessment of earthquake-induced landslides using Bayesian network: a case study in Beichuan, China," *Computers & Geosciences*, vol. 42, pp. 189–199, 2012.

[23] B. Schölkopf, A. J. Smola, and F. Bach, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*, MIT press, 2002.

[24] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[25] M. Moeini, A. Shojaeizadeh, and M. Geza, "Supervised machine learning for estimation of total suspended solids in urban watersheds," *Water*, vol. 13, no. 2, p. 147, 2021.

[26] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: a review," *Journal of Hydrology*, vol. 598, p. 126266, 2021.

[27] T. Yang, X. Liu, L. Wang, P. Bai, and J. Li, "Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique," *Journal of Water Resources Planning and Management*, vol. 146, no. 2, 2020.

[28] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," *Proceedings of International Conference on Global Research and Education*, pp. 215–227.

[29] L. T. Pham, L. Luo, and A. Finley, "Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds," *Hydrology and Earth System Sciences*, vol. 25, no. 6, pp. 2997–3015, 2021.

[30] V. Prasath, H. A. A. Alfeilat, A. Hassanat et al., "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier--A Review," 2017, https://arxiv.org/abs/1708.04321.

[31] A. H. Gandomi, S. K. Babanajad, A. H. Alavi, and Y. Farnam, "Novel approach to strength modeling of concrete under triaxial compression," *Journal of Materials in Civil Engineering*, vol. 24, no. 9, pp. 1132–1143, 2012.

[32] J. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—a discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, 1970.

[33] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Transactions of the ASABE*, vol. 50, no. 3, pp. 885–900, 2007.

[34] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid, "Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile," *Journal of Hydrology*, vol. 567, pp. 165–179, 2018.

[35] Y. Yang and Q. Zhang, "A hierarchical analysis for rock engineering using artificial neural networks," *Rock Mechanics and Rock Engineering*, vol. 30, no. 4, pp. 207–222, 1997.

[36] R. Shirani Faradonbeh, D. Jahed Armaghani, M. Z. Abd Majid et al., "Prediction of ground vibration due to quarry blasting based on gene expression programming: a new model for peak particle velocity prediction," *International journal of Environmental Science and Technology*, vol. 13, no. 6, pp. 1453–1464, 2016.

[37] W. Chen, M. Hasanipanah, H. Nikafshan Rad, D. Jahed Armaghani, and M. M. Tahir, "A new design of evolutionary hybrid optimization of SVR model in predicting the blast-induced ground vibration," *Engineering with Computers*, vol. 37, no. 2, pp. 1455–1471, 2019.

[38] H. Nikafshan Rad, I. Bakhshayeshi, W. A. Wan Jusoh, M. M. Tahir, and L. K. Foong, "Prediction of flyrock in mine

blasting: a new computational intelligence approach," *Natural Resources Research*, vol. 29, no. 2, pp. 609–623, 2020.

[39] M. H. Ahmad, F. Ahmad, X.-W. Tang et al., "Supervised Learning Methods for Modeling Concrete Compressive Strength Prediction at High Temperature," *Materials*, vol. 14, 2021.

[40] M. Ahmad, M. Amjad, R. A. Al-Mansob et al., "Prediction of liquefaction-induced lateral displacements using Gaussian process regression," *Applied Sciences*, vol. 12, no. 4, p. 1977, 2022.

[41] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, and U. Amjad, "Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation," *Applied Sciences*, vol. 12, no. 4, p. 2126, 2022.

Hindawi
Complexity
Volume 2022, Article ID 8451812, 13 pages
https://doi.org/10.1155/2022/8451812

Research Article

Machine Learning as a Downscaling Approach for Prediction of Wind Characteristics under Future Climate Change Scenarios

Abbas Yeganeh-Bakhtiary,[1] **Hossein EyvazOghli**,[1] **Naser Shabakhty**,[1] **Bahareh Kamranzad**,[2,3,4] **and Soroush Abolfathi**[5]

[1]*School of Civil Engineering, Iran University of Science & Technology (IUST), Narmak, Tehran, Iran*
[2]*Hakubi Center for Advanced Research, Kyoto University, Yoshida Honmachi, Sakyo-ku 6068501, Kyoto, Japan*
[3]*Graduate School of Advanced Integrated Studies in Human Survivability (GSAIS), Kyoto University, Yoshida-Nakaadachi 1, Sakyo-ku, Kyoto 6068306, Japan*
[4]*Department of Physics, Faculty of Natural Sciences, Imperial College London, London, SW7 2AZ, UK*
[5]*School of Engineering, University of Warwick, Coventry, CV4 7AL, UK*

Correspondence should be addressed to Abbas Yeganeh-Bakhtiary; yeganeh@iust.ac.ir

Received 18 May 2022; Accepted 2 August 2022; Published 23 August 2022

Academic Editor: Teddy Craciunescu

Assessment of climate change impacts on wind characteristics is crucial for the design, operation, and maintenance of coastal and offshore infrastructures. In the present study, the Model Output Statistics (MOS) method was used to downscale a Coupled Model Intercomparison Project Phase 5 (CMIP5) with General Circulation Model (GCM) results for a case study in the North Atlantic Ocean, and a supervised machine learning method (M5' Decision Tree model) was developed for the first time to establish a statistical relationship between predicator and predicant. To do so, the GCM simulation results and altimeter remote sensing data were employed to examine the capabilities of the M5'DT model in predicting future wind speed and identifying spatiotemporal trends in wind characteristics. For this purpose, three classes of M5′ models were developed to study the annual, seasonal, and monthly variations of wind characteristics. The developed decision tree (DT) models were employed to statistically downscale the Beijing Normal University Earth System Model (BNU-ESM) global climate model output. The *M5′* models are calibrated and successfully validated against the GCM simulation results and altimeter remote sensing data. All the proposed models showed firm outputs in the training section. Predictions from the monthly model with a 70/30 training to test ratio demonstrated the best model performance. The monthly prediction model highlighted the decreasing trend in wind speed relative to the control period in 2030 to 2040 for the case study location and across all three future climate change scenarios tested within this study. This reduction in wind speed reduces wind energy by 13% to 19%.

1. Introduction

Over the past decades, excessive greenhouse gas emissions have resulted in an accelerated rate of global warming and intensified the effects of climate change. The increase in intensity and frequency of extreme climatic events is exacerbated by climate change, leading to natural hazards such as severe floods and erosion in coastal regions [1–11]. The large-scale effects of climate change are beginning to influence several parts of the world by increasing extreme climatic events. A study by Wei et al. [12] showed that the negative impacts of climate change occur at both global and local scales with detrimental consequences on coastal communities, which are at the forefront of battling against the climate change impacts. Given that climate change will also impact major socioeconomic activities and biodiversity in the coastal region, it is vital to have robust predicting frameworks capable of approximating the key climatic parameters in the future considering different climate projection models. Within the context of climate change, wind

climate in offshore and coastal regions is one of the key parameters that influence wave behaviour and hydroclimate [13]. Turki et al. [7] examined the multiscale components of the monthly extreme surges by considering the climatic parameters, e.g., zonal wind, sea surface temperature, and sea-level pressure along the English Channel coasts. Despite the current availability of vast amount of sensing data, high-resolution downscaled models are still needed to investigate the effects of climate variables based on the future climate change projections [14].

Global climate models (GCMs, aka General Circulation Models) have been developed to generate future projections through large-scale spatiotemporal data of climate variables [15]. At present, GCMs are widely used to predict and simulate the large-scale global climate response to increasing temperature at the surface of oceans. Many studies have been carried out using different GCMs to predict the impact of climate change on the variations of wind characteristics in various regions [16–20]. For instance, Segal et al. [21] employed HadCM2 (Hadley Centre coupled model) with coupling local wind data and showed that the availability of daily average wind power reduces within the range of 0–30% by 2050 over most areas of the United States. Later, Breslow and Sailor [22] also used outputs of global coupled global climate models (CGCMs) with a resolution of 3.75° (both latitude and longitude) and HadCM3 with a resolution of 3.75° at 2.5° (longitude and latitude, respectively). The results of the case study in the United States show that climate change reduces the average wind speed by 10% to 15%. This reduction in wind speed resulted in a 30% to 40% reduction in wind power.

Lionello et al. [23] studied the Adriatic Sea region using ECHAM-4 model data, which were downscaled using statistical methods. A comparison between the present and future climate simulations from Lionello et al. [23] study showed that extreme wave height will decrease in the future. In the UK, the 40 yearly records of data show an increase in winter wind speed by 15% to 20%, which can be linked to climate change consequences [24]. For two scenarios of $A2$ (1961–1990) and $B2$ (2071–2100), Lionello et al. [17] studied the seasonal mean of significant wave height (SWH) for the Mediterranean and predicted a reduction in SWH for the B2 period. For a case study of the Bay of Biscay, France, the results of the ARPEGE-Climate model show a reduction in the wind speed and, consequently, the wave height for the summers during the period of 2061 to 2100 [25]. Kamranzad [19] and Kamranzad et al. [26] investigated the capabilities of CGCM3.1 (Canadian Global Coupled Model Version 3.1) in the prediction of wind characteristics in Persian Gulf, which, considering its semienclosed shape, is vastly different from oceans. Kamranzad et al. [26] predicated that by the year 2100, wind speeds and wind energy will decrease across Persian Gulf according to three different emission scenarios ($A2$, $B1$, and $A1B$). Recently, Goharnejad et al. [27] evaluated both the wave characteristics and wave energy extraction potential in Persian Gulf for greenhouse gas concentration trajectory (representative concentration pathway: RCP)-based GCM simulation outputs of RCP4.5 and RCP8.5 future climate change scenarios and showed that the potential wave energy level will be higher in the southern regions.

Given the importance of the North Atlantic in ocean renewable energy and the heavy investment plans for extending the onshore and offshore renewable energy farms, understanding and quantifying the impacts of climate change on wind and wave characteristics across the North Atlantic are of exceptional importance. However, a limited number of studies have focused on the impacts of climate change on wind characteristics across the North Atlantic Ocean. Wang et al. [16] investigated mean and maximum seasonal variations of SWH based on three emission scenarios and showed that for fall and winter seasons during the twenty-first century, at the middle latitudes of the North Atlantic Ocean, the SWH will decrease, while at the southwest regions of the Atlantic the SWH will increase. Using the ECHAM5 model, Hemer et al. [28] predicted up to a 15% decline in SWH across the midlatitudes of the Atlantic and a 10% SWH reduction in the Southern regions of the Atlantic Ocean. The existing studies and modelling data indicate a decreasing trend in the future wind speed across vast regions of the North Atlantic.

Although GCMs are powerful in predicting the main features of the global atmospheric currents, they are often not capable of vigorously approximating local climate details [29]. Hence, there is a need to develop appropriate tools to downscale GCM climate change forecasts to local and regional scales [30]. Previous studies have adopted three downscaling approaches: empirical, semiempirical, and nesting methods (i.e., dynamical downscaling). In the empirical approach, historical climatic conditions are used to present local analogue scenarios. Such studies are attributable to a qualitative conceptual survey, and results from empirical approaches do not generate a climate forecasting model. Semiempirical (statistical) and nested (dynamical) downscaling approaches use large-scale GCM predictions to develop local climate change scenarios [31]. In the dynamical downscaling approach, a regional climate model (RCM) with the target mesh resolution uses large-scale GCM outputs as the boundary condition for the RCM to produce higher resolution outputs [32]. The major drawbacks of dynamical downscaling methods, which limit their applications in climate change impact assessments, are method complexity, high computational cost, and case-sensitive performance [33].

Statistical downscaling methods are divided into three categories [34, 35]: In the Perfect Prognosis (PP) methodology, a relationship is established between large-scale observational data and locally recorded data [36, 37]. The Model Output Statistics (MOS) method is similar to the PP, except that in this approach, a relationship is created between GCM outputs (predictor) and local climate variables (predictands) [38], and in the Stochastic Weather Generator (SWG) category, this relationship is developed by perturbing probably distribution parameters [39]. Considering the many parameters that are involved in the simulations of GCMs and the scenarios that are intended for the future and simulate trends, the use of the MOS method can be proper for downscaling where only limited observational data are

available in the application of the PP approach. For more information on statistical downscaling methods: [35, 40–44].

The statistical downscaling methods are designed based on two assumptions: (i) the empirical relationships between historical large-scale atmospheric predictors modelled by GCMs, and local climate characteristics can be established, and (ii) the obtained empirical relationships are valid under climate change scenarios [18, 45]. The most popular statistical downscaling approach is transfer functions based on fitting a quantitative relationship between large-scale climate variables and local-scale climate variables. In recent years, machine learning techniques have been adopted to determine the required transfer function in statistical downscaling [46, 47].

Due to the nonlinear time-series nature of climatic processes involved in the predication process, the artificial neural network (ANN), as a self-organizing estimator function, is widely adopted in modelling and forecasting wind characteristics (among others, [31, 48–51]). For example, Sailor et al. [31] adopted ANN technique to downscale and forecast surface wind speeds at three locations across the United States with a high potential for future wind power generation over the next 100 years. Under the future climate change scenarios, they estimated that wind power will decrease between 0.9% and 8%. Nourani et al. [49] employed ANN method to downscale climate variables, including temperature and precipitation, at two study locations (Ardabil and Tabriz, Iran) for single and multi-GCM outputs. Nourani et al. [49] showed that the downscaling method using ANN-based multi-GCM outputs leads to more accurate results.

Despite the advantages of ANN in downscaling and predictions of climatic processes, there are several deficiencies, including the probability of error and mismatch, due to not removing irrelevant data and noneffective parameters and the challenges associated with the training process by increasing the size of input time series. Given the generally nonstationary and large temporal scale (from a few minutes to several decades) of climatic data, the increased computational time required to train ANN models limits their applications [50]. Therefore, in recent years clustering methods were proposed as an alternative to robustly predict climatic variables and overcome the difficulties associated with the ANN downscaling approach. The decision tree (DT) is one of the most popular and efficient data mining techniques for clustering and generating regression models [52]. Decision tree (DT) models are clustering-type models with the advantages of setting out the variable choices logically, simultaneously considering potential options and choices, with tangible and easy-to-understand results [53].

Robustness of the DT models in identifying the effective parameters and understanding interdependencies of complex nonlinear climate variables [29, 50, 53–55] makes them a powerful tool for downscaling of GCM outputs and wind speed prediction. In this study, an M5' DT model too was developed to predict the spatiotemporal variations and trends in wind speed across a case study located in North Atlantic Ocean. The capability of M5' DT was examined for the first time for the prediction of wind speed variations in the coming decades, considering a range of climate change scenarios. The variations in the projected wind speed simulated by GCM (i.e., BNU-ESM model by Ji et al. [56]) were investigated as the case study region of the midlatitudes of the North Atlantic Ocean. The study region was chosen based on its importance and potential for current and future offshore wave and wind renewable energy farms. This study investigates the changes in wind speed over North Atlantic for the years 2030, 2035, and 2040, considering three future climate change scenarios outlined by the Intergovernmental Panel on Climate Change (IPCC: www.ipcc-data.org).

2. Materials and Methods

In this study, two sets of GCM and altimeter data were used to develop and verify a DT model to downscale the GCM outputs and predict the wind speed during 2030–2040. The case study location and details of the methodological approach adopted are described in the following section.

2.1. Study Area. Figure 1 shows the geographical extent and location of the case study area. The case study is an area with high wind speed, and the potential for efficient operations of offshore renewable energy projects was selected as the case study location. The boundaries of the study area were selected based on the GCM grids, covering 340.31°E to 357.19°E in longitude and 57.21°N to 62.79°N in latitude. The downscaling and machine learning model developed was applied to the data for the case study area to predict the wind speed variations under future climate scenarios and evaluate the performance of the proposed method in comparison to the measured data.

2.2. General Circulation Models (GCMs) Dataset. The Earth System Model (Beijing Normal University (BNU)-ESM) developed by Ji et al. [56] is based on climate projection models and is widely used to study climate change impacts, ocean-atmosphere interaction mechanisms, and climate-carbon interactions on temporal scales spanning from a month up to a century. The ESM benefits from several submodels, including atmospheric, ocean, sea ice, and land models. The ESM coupling framework is developed based on the Community Climate System Model version 4 (CCSM4) together with the Community Climate System Model and Community Earth System Model (CCSM/CESM) [56]. In the present study, the GCM historical data of 15 years (between 1991 and 2005) were employed as the control period, and the forecast data under three future climate scenarios, including RCP2.6, RCP4.5, and RCP8.5, were used to predict the climate variables for 2030, 2035, and 2040.

2.3. Altimeter Dataset. Historically, wind characteristics have been studied by analysing time-series records of weather stations that provide reliable measurements of temporal variations of wind characteristics in a fixed position. In recent years, remote sensing techniques such as

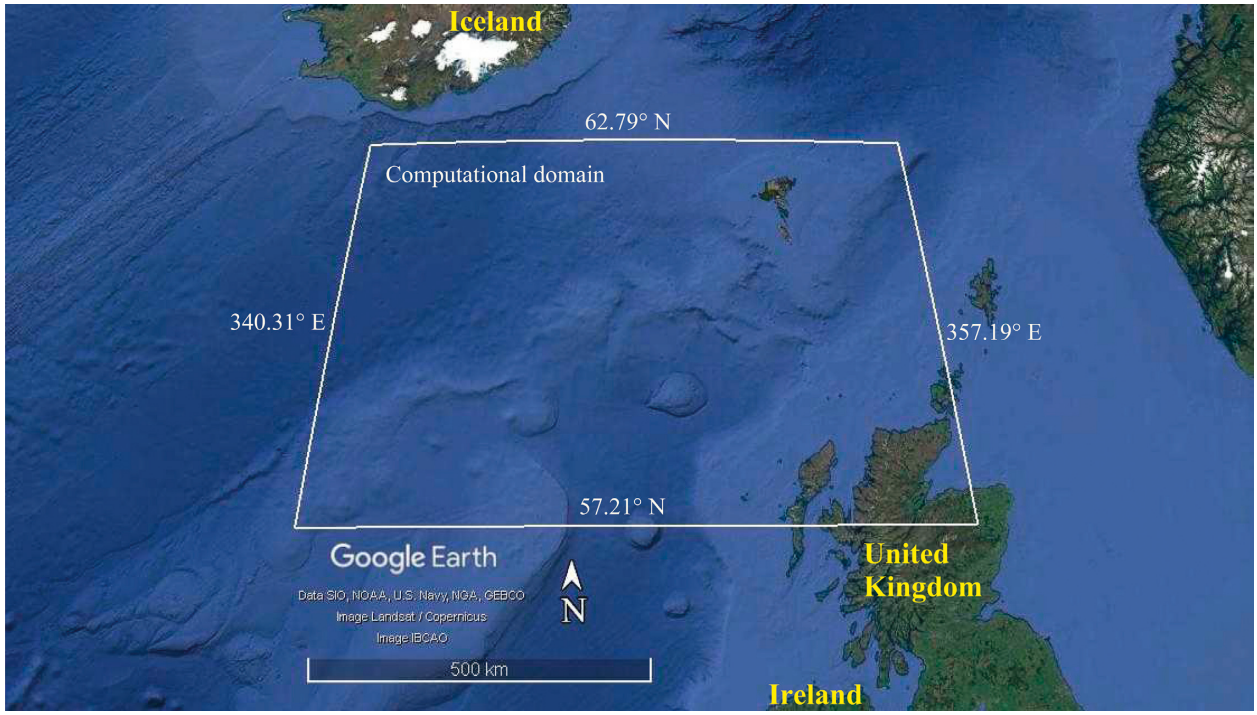


Figure 1: Geographical extent and location of the case study area.

radar measurements and satellite imagery have been widely used as tools to analyse key atmospheric parameters and sea conditions. The spatial distribution of wind characteristics has been analysed and evaluated from remote sensing imageries [57, 58].

Altimeter data are one of the most valuable data sources available that have been employed in previous investigations to verify the performance of a range of climate prediction models (see among others, [59–63]). The altimeter data are recorded along the paths of satellites using remote sensing equipment and provide a firm remote sensing dataset to validate and evaluate the performance of the climate projection models. Altimeter data records are entered in irregular tracks and at different time intervals, which demonstrates the importance of applying appropriate filters to these data to ensure the consistency of the data records. For this purpose, we adopted a 20-minute filter in the time domain and a 1.5-degree filter in the spatial domain to tackle the spatiotemporal nonlinearity of data. The wind speed obtained from altimeter data was used for the downscaling process and validation of the developed DT models. Considering the available period of GCM simulation outputs, the altimeter dataset belonging to the period of 1991–2005 was selected as the control period.

2.4. M5′ Model Tree. Recently, the use of decision tree (DT) algorithms as a robust machine learning technique for prediction of hydroclimatic parameters in coastal and offshore engineering problems has found a growing interest (among others, [51, 54, 55]). In general, the structure of a decision tree is composed of four parts including root, branch, node, and leaves. The root (or first node) is at the top of the tree; also, at the end of the chain of branches and nodes are the leaves (or the last node). Figure 2 shows a schematic of the DT logical structure used as a predictive model. Given that DTs can be classed as a graphical method, interpretations of DT's model outputs are easier compared to other machine learning methods [50].

Nourani et al. [64] investigations on the application of machine learning techniques for predicting climate variables show that as the prediction horizon increases, the accuracy of predictive models is reduced. The reason is due to the nonlinear growth of error propagations in those nonlinear predictive models. However, the issues associated with the error growth are not the case for linear models such as M5′ DTs, as the error in such models will remain constant by increasing the prediction horizon. Thus, the multilinear regression models such as *M5′* model can provide more reliable results in predicting climate variables, compared to nonlinear models. Furthermore, the M5′DT models have superior performance in identifying and selecting the most effective parameters for persistent prediction [50].

2.5. Model Development. Prior to M5′ model development, the control period dataset (historical data) was divided into train and test partitions. The train data were employed for development of the M5′DT models, whereas the test data were utilized for verification and evaluations of the generated M5′ models. Following the data partitioning, the train data were employed to develop three predictive models with M5′ technique. These models were designed with different ratios of test/train data length and three different train/test size ratios of 35/65, 30/70, and 25/75 to evaluate the best predictive model.

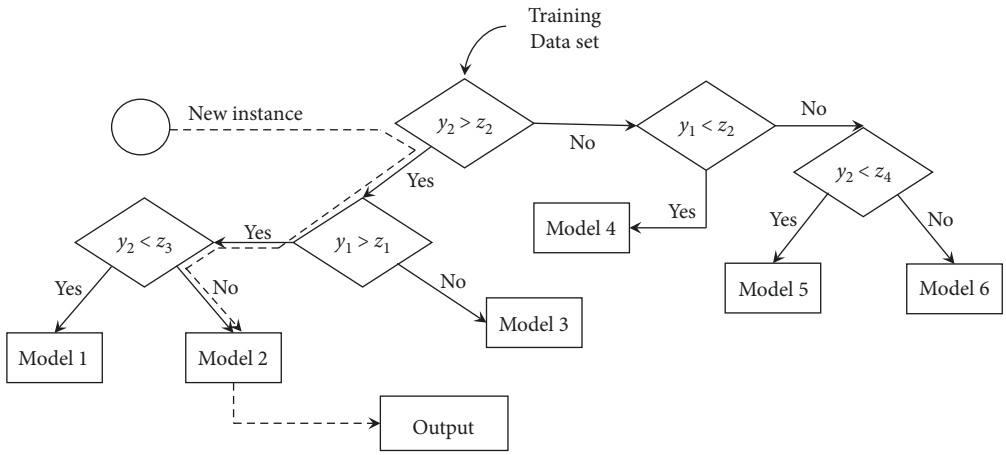


Figure 2: Flow chart of M5′ model prediction procedures [55].

In the first step, the *M5′* models were trained; the process of data selection for training the *M5′* models involved specifying the coordinates of each data instance and determining its corresponding cell from the GCM mesh grid to obtain the wind speed and direction for the specific cell. Based on the instance's time, a time interpolation is performed on the GCM data output. Then, the interpolated wind speeds and direction for a given location and time instance were determined from the GCM cell's corners and introduced to the *M5′* models as an input. In the second step, the trained *M5′* models were verified using the test data from the historical dataset. Then the performance of each model was evaluated using statistical error measures to identify and select the best model for spatial and temporal prediction of wind characteristics across the case study. The final step of the modelling involves prediction of wind speed for the time intervals of 2030 to 2040 based on the three emission scenarios employing the GCM simulation data.

2.6. Performance Evaluation of Predictive Models. Given the nonlinear nature of climatic events and influence of complex marine processes varying at both temporal and spatial scales, predicting the wind characteristics in marine environment is a very difficult and challenging task [65]. Having a strong approach to evaluate and overcome the systematic and random errors of the developed DT models is also crucial. To do so, several statistical assessment criteria, namely correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE), were investigated to evaluate and compare the predictive robustness of the proposed models under future climatic conditions. The correlation coefficient (CC) (1) is adopted to determine the relationship between the predicted wind speed and measured values as:

$$CC = \frac{\sum_{i=1}^{N}\left(\left(O_i - \overline{P}\right)\left(O_i - \overline{O}\right)\right)}{\sqrt{\left(\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2\right)\left(\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2\right)}}, \quad (1)$$

where P_i and O_i, respectively, denote the predicted and measured data (observations), $\overline{P}$ is the average of the predicted data, $\overline{O}$ is the average of the measured data, and N is the number of data points. CC only determines the correlation between measured and predicted values and therefore it is not a sufficient measure to provide a comprehensive understanding of the model's performance. Thus, it is critical issue to employ benchmarks that rigorously determine the model's prediction errors. To this end, the present study employed MAE and RMSE statistical measures in equations (2) and (3).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - O_i\right|, \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - O_i\right)^2}. \quad (3)$$

3. Results and Discussion

To obtain each downscaled wind speed, eight parameters from GCM data were inputted to the *M5′* models, including four wind speeds and four wind directions. The purpose of wind directions participating in the wind speed downscaling was to help the projection model to predict wind speeds more accurately at locations such as the shoreline, where wind directions change rapidly. In the following, the results are presented in two parts: results of the control case and results of the prediction case.

3.1. Control Period. In the control case, three types of prediction models were developed and named as: Annual (*A*), Seasonal (*S*), and Monthly (*M*) models. In the *A*-type model, all control period data were used to develop every single *M5′* model. Where the results showed that in the *A*-type model, although the correlation coefficient between the observational data and the outputs of the DT model in the training period was high, in the test period, the correlation coefficient between the measured and predicted data was low, and the prediction error raised considerably. Since the results of the A-type model were not accurate enough, and

Table 1: Developed models in the present study and their description.

Model type	Description
A (Annual)	One DT model was developed using the Train data, validated by the Test data, and adopted for prediction in the forecast section.
S (Seasonal)	The entire data were categorized into seasons, and four decision tree (DT) models were generated alongside together. Each model is developed and validated with the data of the relevant season and used for forecasting in the season for the projection period. The names of these models based on the seasons are as follows: Swinter: winter's model, Sspring: spring's model, Ssummer: summer's model, and Sfall: fall's model.
M (Monthly)	The entire data were divided into months, and 12 DT models were extended beside together. Each model is developed and verified with the data of the relevant month and used for forecasting in the relevant month for the projection period. The names of these models based on the months are MJan to MDec models for January to December, respectively.

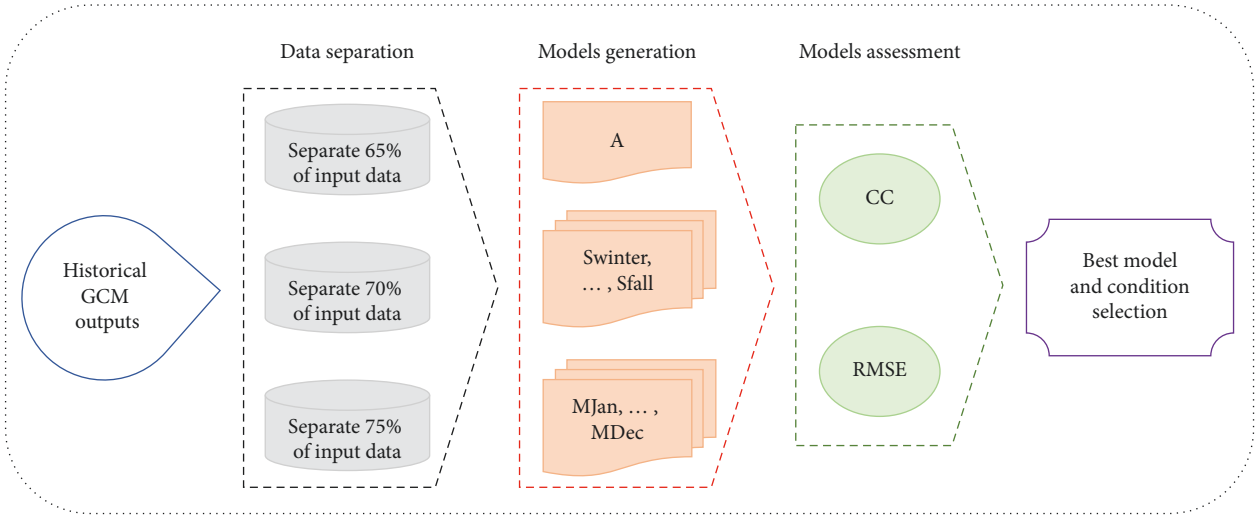


Figure 3: Schematic of the downscaling methodology and predictive modelling processes.

Table 2: Comparison of statistical predictive measures across all the tested scenarios.

Model	Training (70/30)		Verification					
	CC	RMSE	Correlation coefficient			Root mean square error		
			65/35	70/30	75/25	65/35	70/30	75/25
A	0.9028	2.0117	0.3642	0.1237	0.2136	5.7777	17.4996	4.1732
Swinter	0.8944	2.0585	0.0514	0.2003	0.1699	32.723	10.9887	12.8441
Sspring	0.0896	1.8625	0.1296	0.1679	0.6327	15.2014	10.8614	3.0635
Ssummer	0.8934	1.8063	0.0744	0.2707	0.1093	30.627	6.2338	15.2301
Sfall	0.888	2.0465	0.075	0.5276	0.5049	3.9912	3.6316	3.8453
Mean	0.6914	1.9434	0.0826	0.2916	0.3542	20.6356	7.9289	8.7458
MJan	0.8983	2.0922	0.56	0.8054	0.7704	3.7827	2.6749	2.8072
MFeb	0.9046	1.9672	0.5936	0.7791	0.0881	3.6361	2.6831	41.9419
MMar	0.9026	1.9325	0.7804	0.7974	0.6843	2.6535	2.5707	3.0753
MApr	0.8988	1.8653	0.1763	0.7333	0.4716	12.1353	2.6809	4.2891
MMay	0.897	1.7869	0.2977	0.6998	0.0389	5.8888	2.6363	47.1423
MJun	0.8892	1.7279	0.558	0.6979	0.3844	3.0569	2.4745	4.5731
MJul	0.8822	1.6457	0.5116	0.6576	0.6059	2.9252	2.6373	2.6362
MAug	0.8903	1.6815	0.7566	0.7566	0.7392	2.2544	2.2544	2.2925
MSep	0.902	1.9198	0.3506	0.7627	0.0679	5.77	2.6672	40.5325
MOct	0.8815	1.9988	0.0362	0.7565	0.4504	3.86	2.5783	4.0723
MNov	0.889	1.9712	0.0423	0.7174	0.0614	73.324	2.7225	22.6081
MDec	0.8984	2.0211	0.3008	0.7511	0.0912	7.0636	2.7503	24.5093
Mean	0.8945	1.8842	0.4137	0.7429	0.3711	10.5292	2.6109	16.7067

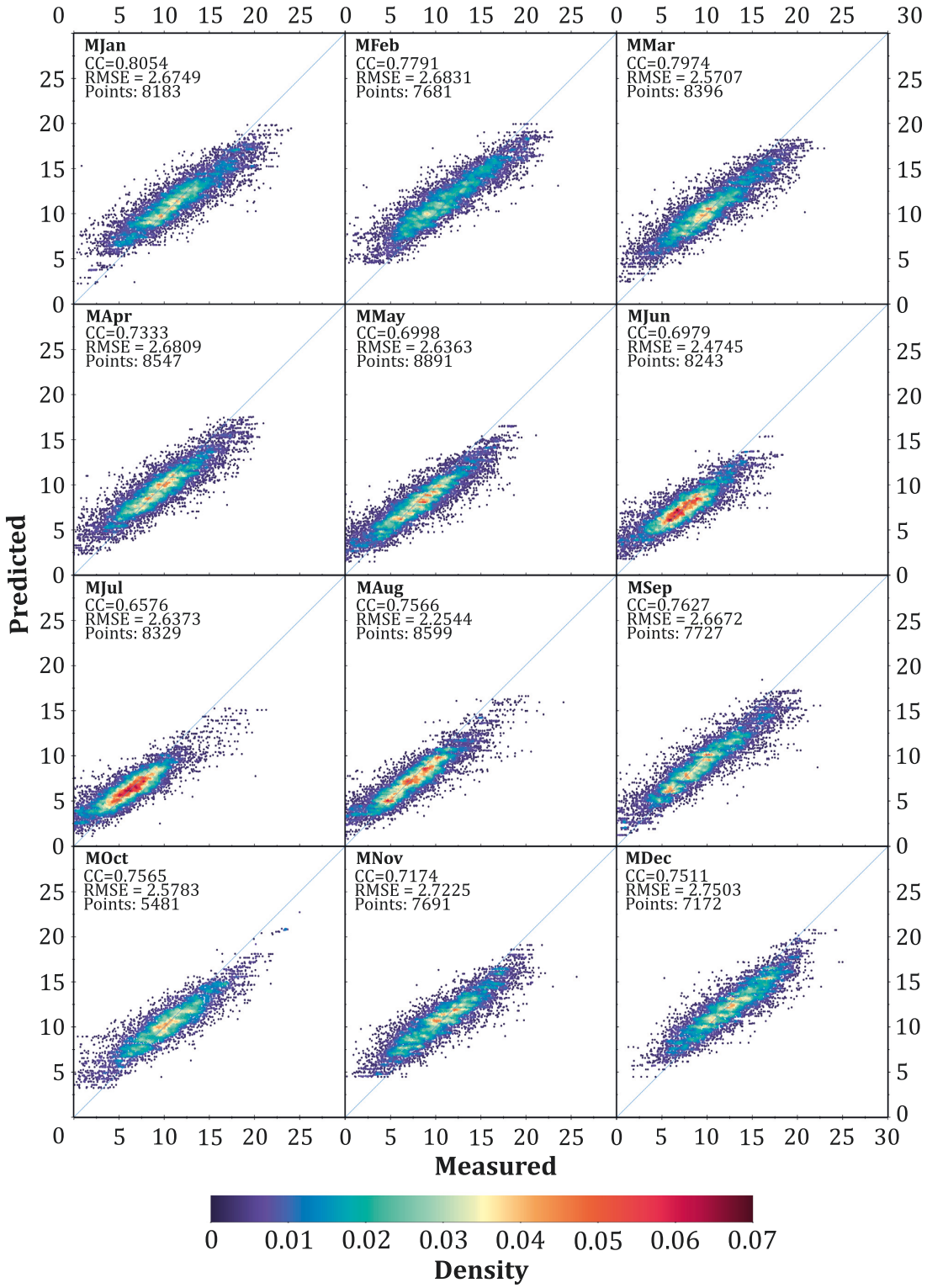


Figure 4: Comparison between the observed data and decision tree (DT) model prediction.

the $M5'$ model acts based on data categories theory, S and M types of the models were developed. Thus, all the data are divided into seasons and months, and the progress of the $M5'$ model development is conducted for each of the categories separately (a summarized description of each model presented in Table 1). Figure 3 presents schematically the downscaling procedure and models' deployment.

Table 1 presents the differences between the models, and as can be seen, the main difference is in the number of models created to downscale GCM's outcomes and predict

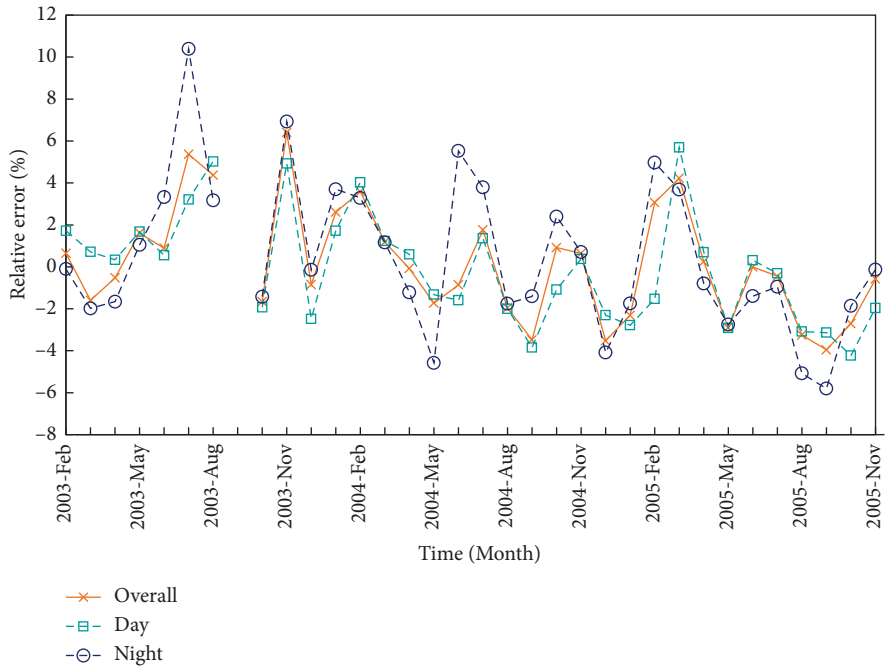


Figure 5: Comparison of the overall, and day and night time prediction errors of M5′ models.

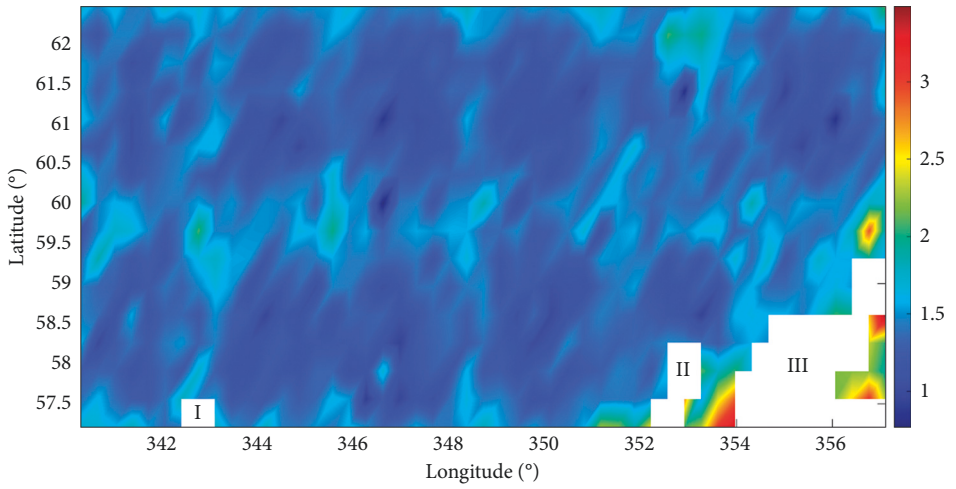


Figure 6: Spatial distribution of mean absolute error (MAE) across the case study region.

the trend of wind changes. The principal idea of the model's development is how the DT model works, which was attempted to prepare more homogeneous input data for each model by dividing the data so that the results reach acceptable accuracy.

Table 2 summarizes the results of the models generated during the training and validation steps. The correlation coefficient (CC) and the root mean square error (RMSE), and the mean of the seasonal and monthly models are reported in this table to compare the generated models' performance. As seen in the training section, the CC for all models is close to 0.9, except for model Sspring, in which CC is less than 0.1. All the models' RMSEs are in the range of 1.6 to 2.0. On the basis of the results, the best model in terms of CC is the *A*-type model with CC = 0.9. Considering the mean of CC, the *A*-type model has also shown to be more efficient. However, considering the RMSE, the annual model performs weakly, and the monthly models often have better performance. In terms of mean values, the *M*-type model with an RMSE of 1.88 has the lowest error, whereas the RMSE of the seasonal and annual models are 1.94 and 2.01, respectively. The annual models have the weakest output among the models. This may be attributed to the high volume of input data leading to the *A*- and *S*-type models, which causes misunderstanding in both classifying the data and establishing a reliable correlation between the input data (wind speed and direction data) and the output data (wind speed data). In contrast, the monthly models were able to detect the trend and establish a proper relationship between the input and output data.

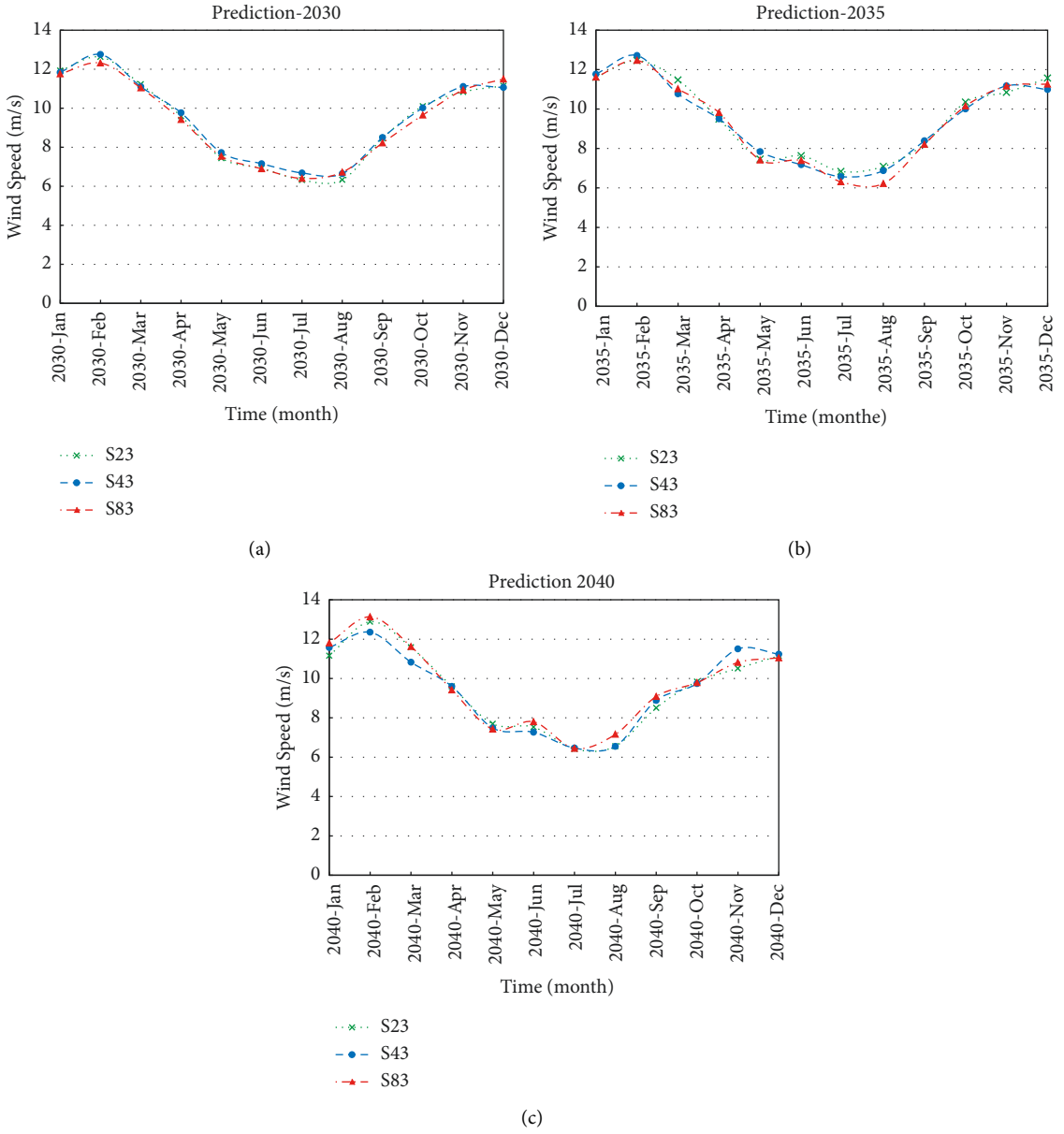


Figure 7: Monthly averaged wind speed predicted based on future climate scenarios (RCP2.6, RCP4.5, RCP8.5) for the year. (a) 2030. (b) 2035. (c) 2040.

In addition, with a close look at Table 2, it can be understood that adopting different historical data for the train and test periods affects the developed model performance and the results. Three modes of data division—65/30, 70/30, and 75/25—and the outcomes of these models' implementation are presented in this paper. As it comes from Table 2, the *A*-type model failed to make accurate predictions in the test section as the best model of *A*-type in 75/25 mode had a CC of 0.21 and RMSE of 4.17. Despite the relative increase of the mean correlation coefficient in the 30/70 and 25/75 modes, *S*-type models have not yet reached an acceptable level, and the errors are still high. The best results belong to the 70/30 mode of *M*-type models with a mean CC of 0.74 and mean RMSE of 2.61 in the test section.

The lack of a persuasive link between the input and output values in the other modes of *M*-type models can be mainly due to the high sensitivity of *M5′* models to the input data and shows the importance of examining various modes of models before adopting them as a prediction model. A detailed appraisal of the selected model (70/30 mode of *M*-type models) is presented in the following. The scatter plots of the measured data versus the models' results are shown to check the projection models' output quality and the correlation between the predicted and observed data (Figure 4). As shown in Figure 4, the *M*-type model outputs had a perfect correlation with the recorded data. It proves the high efficiency of trend recognition and prediction by M-type models.

Table 3: Predicted monthly mean (PMM) of wind speed changes compared to the year 2005.

	Monthly prediction based on Scenarios and Years									
	Historical	RCP2.6			RCP4.5			RCP8.5		
Month	2005	2030	2035	2040	2030	2035	2040	2030	2035	2040
Jan	13.2741	11.93	11.61	11.16	11.84	11.77	11.57	11.76	11.61	11.80
Feb	11.8188	12.63	12.45	12.89	12.75	12.72	12.35	12.32	12.46	13.14
Mar	10.1221	11.22	11.47	11.59	11.10	10.77	10.83	11.04	11.03	11.62
Apr	10.0947	9.590	9.483	9.593	9.763	9.500	9.598	9.416	9.816	9.419
May	9.0151	7.441	7.463	7.694	7.726	7.840	7.480	7.539	7.411	7.422
Jun	7.7712	6.922	7.636	7.530	7.156	7.178	7.268	6.899	7.385	7.797
Jul	7.0615	6.315	6.836	6.414	6.682	6.572	6.464	6.401	6.309	6.447
Aug	8.5289	6.329	7.085	6.555	6.648	6.873	6.550	6.721	6.218	7.160
Sep	9.6920	8.455	8.222	8.503	8.498	8.390	8.884	8.220	8.208	9.083
Oct	11.5291	10.10	10.35	9.839	10.02	10.00	9.741	9.650	10.15	9.788
Nov	11.8781	10.83	10.85	10.51	11.11	11.18	11.50	10.93	11.16	10.81
Dec	11.1440	11.14	11.57	11.15	11.06	10.99	11.24	11.49	11.26	11.05
Mean	10.1608	9.409	9.586	9.452	9.529	9.481	9.457	9.365	9.418	9.628

For a better understanding of the monthly model performance, the average error of wind speed across the entire study area for the predicted values is depicted in Figure 5. As seen from Figures 4 and 5, although the most part of wind speed data vary from around 3 to 15 m/s, the monthly average of predicted values entirely meets the measured values in the test section. In addition to the total relative error (overall error), to get a better picture of the model performance, two relative error curves for day and night periods are also portrayed in Figure 5. As seen, from the Figure, during the daytime, the model prediction is slightly better than during the night time; while, the overall monthly predication of the wind speed was hindcasted precisely. Therefore, it is evident that the monthly model not only shows the proper distribution of wind speed, but it also accurately predicts the wind speed values in the time domain.

In addition to the time domain investigation, it is necessary to evaluate the spatial distribution of prediction errors over the study area. Figure 6 depicts the mean absolute error (MAE) over the entire region as averaged over the validation period. Notably, a filter was applied in portraying the map to be more reliable, and just cells with more than two data were shown in colour. As seen, Figure 6 contains four zones: zone I remained white because there are no observational data in this zone, and therefore no error was calculated. For zones II and III, MAE was also not figured since the relevant cells fully contain land and no observational data are available in these zones. The fourth zone is the coloured area and represents the MAE in the cells with dimensions of 0.4 by 0.4°.

As seen in Figure 6, the MAE in most parts of the study area is less than 2.0 m/s and remains in the acceptable range. It is only near zones II and III that the MAE reaches about 3.0 m/s; the reason for this phenomenon is also the proximity of this part of the region to the land, where the direction and speed of the wind change rapidly. These rapid changes in wind direction disrupt the performance of projection equation and most probably caused significant errors near the coasts.

3.2. Future Projection. In the prediction case, the wind speed is predicted for years of 2030, 2035, and 2040 by implementing the validated monthly model (M-type model in the 70/30 mode) and the GCM prediction data under three future climate change scenarios RCP2.6, RCP4.5, and RCP8.5. Similar to the control stage, the prediction model input was four wind speeds and in four wind directions. In order to simulate as closely as possible to the control period, the time and location of the selected points to anticipate wind speeds in the prediction case are the same as the time and position of the 2004 records, having the highest number of records per year. Thus, the measurement specifications of the 2004 records were used to predict with changing only the components of the "Year" and fixing the other terms (e.g., month, day, hours, minutes, and seconds) and maintaining the longitudes and latitudes.

Figure 7 shows the monthly average wind speed predictions for different future climate change scenarios across the study area. As seen, during 2030 to 2040, the monthly variations of mean wind speed were at a constant range of 6 to 13 m/s, then after reaching its peak in the second month, the decreasing trend began to reach 11 m/s in August. The wind speed trends under the three future climatic scenarios are rather close to each other for those three years. However, the monthly mean wind speed fluctuations increased from 2030 to 2040, and most fluctuations were related to the RCP8.5 scenario.

For further investigation, the predicted monthly mean (PMM) wind speed values are compared with the corresponding wind speed in 2005 (the last year of the GCM historical simulation outputs). Table 3 lists the changes in PMM values. As shown in Table 3, it is only in February and March that the PMM wind speed is higher; in December, on the other hand, there are no significant changes compared to 2005. For the rest of the years, the projected PMM wind speed was declining. Also, the annual mean values of wind speed (the bottom row of Table 3) were reduced comparing to 2005. The reduction ratio of the annual mean wind speed is approximately 7% to 8%, which is in line with the reported results of previous studies (e.g., [16, 28]). Reduction in wind

speed substantially impacts the wind energy due to climate change. Considering the Segal et al. [21] equation for wind power related to the third order of wind speed, the 7% to 8% reduction in annual mean wind speed indicates a 13% to 19% decrease in wind energy.

4. Conclusion

The intensity and frequency of extreme climatic events are exacerbated due to the impacts of climate change in different parts of the world. Predicting the wind characteristics under future climate change scenarios is very vital for evaluating the performance of existing and future marine engineering projects. This study adopts M5′ DT technique for the first time as a MOS multilinear downscaling method for predicting wind speed trends across a case study region in North Atlantic Ocean. The GCM simulation outputs and altimeter remote sensing data were used to train and validate the developed $M5'$ models. Three climate change scenarios of RCP2.6, RCP4.5, and RCP8.5 were adopted for deriving the predictions from $M5'$ models. Three downscaling models, including annual model (A), seasonal models (S), and monthly models (M), were tested for predicting wind speed from DT models. The performance of three $M5'$ models with train-test data ratios of 75/25, 70/30, and 65/35 was investigated to determine the best performing $M5'$ model. Data from the control case were used to choose the best model for predicting wind speed under future climate scenarios. The following conclusion can be drawn:

(i) The detailed analysis of the prediction results from the $M5'$ technique indicates the robustness and appropriateness of the proposed models for assessing wind characteristics under future climatic scenarios. The proposed approach for downscaling has successfully predicted the trend of monthly averaged wind changes in the study area.

(ii) Model performance evaluations were conducted employing appropriate statistical measures. As a result, a detailed analysis of $M5'$ models shows appropriate performance in the training section; whereas, the monthly models provided more reliable predictions in the test section. Moreover, the developed models performed in several modes, and the monthly model in the mode of 70/30 presented the best performance. This diversity in the models' performances in different modes indicated the sensitivity of the $M5'$ DT model to the input data.

(iii) Based on the monthly model predictions, in 2030, 2035, and 2040, the average monthly wind speed values for all the three future climate change scenarios were close. However, from 2030 to 2040, the range of monthly mean wind speed oscillations increased, which in the RCP8.5 scenario was more evident than in the other two scenarios.

(iv) A comparison between the PMM wind speeds and the year 2005 shows that the PMM wind speeds decreased remarkably. Hence, wind energy has also been experimented with a reduction that could be of interest in renewable energy studies and projects in the study area and would challenge the economic exploitation of these resources.

Data Availability

All observed data used during this research are openly available at (ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/data/). Also, the data that support the findings of this study are available upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Dr. Seyed Mostafa Siadatmousavi, Mr. Mostafa Beyramzadeh, and Mr. Amir Molajou for their helpful technical feedback.

References

[1] S. Dong, M. Salauddin, S. Abolfathi, Z. H. Tan, and J. M. Pearson, "The influence of geometrical shape changes on wave overtopping: a laboratory and SPH numerical study," in *Coasts, Marine Structures and Breakwaters*, pp. 1217–1226, ICE Publishing, London, 2017.

[2] M. Armanfar, H. Goharnejad, M. Z. Niri, and W. Perrie, "Assessment of coastal vulnerability in Chabahar Bay due to climate change scenarios," *Oceanologia*, vol. 61, no. 4, pp. 412–426, 2019.

[3] A. Fitri, R. Hashim, S. Abolfathi, and K. N. Abdul Maulud, "Dynamics of sediment transport and erosion-deposition patterns in the locality of a detached low-crested breakwater on a cohesive coast," *Water*, vol. 11, no. 8, p. 1721, 2019.

[4] S. Dong, S. Abolfathi, M. Salauddin, Z. Tan, and J. Pearson, "Enhancing climate resilience of vertical seawall with retrofitting - a physical modelling study," *Applied Ocean Research*, vol. 103, Article ID 102331, 2020.

[5] M. Salauddin, J. O'Sullivan, S. Abolfathi, S. Dong, and J. Pearson, "Distribution of individual wave overtopping volumes on a sloping structure with a permeable foreshore," *Coastal Engineering Proceedings*, vol. 36v, p. 54, 2020.

[6] S. Dong, S. Abolfathi, M. Salauddin, and J. Pearson, "Spatial distribution of wave-by-wave overtopping at vertical seawalls," *Coastal Engineering Proceedings*, vol. 36v, p. 17, 2020.

[7] I. Turki, N. Massei, B. Laignel, and H. Shafiei, "Effects of global climate oscillations on intermonthly to interannual variability of sea levels along the English channel coasts (NW France)," *Oceanologia*, vol. 62, no. 2, pp. 226–242, 2020.

[8] A. Yeganeh-Bakhtiary, H. Houshangi, and S. Abolfathi, "Lagrangian two-phase flow modeling of scour in front of vertical breakwater," *Coastal Engineering Journal*, vol. 62, no. 2, pp. 252–266, 2020.

[9] H. Gao, B. Liang, and Z. Shao, "A global climate analysis of wave parameters with a focus on wave period from 1979 to 2018," *Applied Ocean Research*, vol. 111, Article ID 102652, 2021.

[10] M. Salauddin, J. J. O'Sullivan, S. Abolfathi, and J. M. Pearson, "Eco-engineering of seawalls—an opportunity for enhanced

climate resilience from increased topographic complexity," *Frontiers in Marine Science*, vol. 2021, Article ID 674630, 2021.

[11] M. Salauddin, J. O'Sullivan, S. Abolfathi, and J. M. Pearson, "Extreme wave overtopping at ecologically modified sea defences," *EGU General Assembly*, vol. 2020, 2020.

[12] J. Wei, A. Hansen, Y. Zhang et al., "Perception, attitude and behavior in relation to climate change: a survey among CDC health professionals in Shanxi province, China," *Environmental Research*, vol. 134, pp. 301–308, 2014.

[13] D. E. Reeve, Y. Chen, S. Pan, V. Magar, D. Simmonds, and A. Zacharioudaki, "An investigation of the impacts of climate change on wave energy generation: the Wave Hub, Cornwall, UK," *Renewable Energy*, vol. 36, no. 9, pp. 2404–2413, 2011.

[14] R. Kabir, H. T. A. Khan, E. Ball, and K. Caldwell, "Climate change impact: the experience of the coastal areas of Bangladesh affected by cyclones sidr and aila," *Journal of Environmental and Public Health*, vol. 2016, Article ID 9654753, 2016.

[15] P. Camus, F. J. Mendez, and R. Medina, "A hybrid efficient method to downscale wave climate to coastal areas," *Coastal Engineering*, vol. 58, no. 9, pp. 851–862, 2011.

[16] X. L. Wang, F. W. Zwiers, and V. R. Swail, "North Atlantic Ocean wave climate change scenarios for the twenty-first century," *Journal of Climate*, vol. 17, no. 12, pp. 2368–2383, 2004.

[17] P. Lionello, S. Cogo, M. B. Galati, and A. Sanna, "The Mediterranean surface wave climate inferred from future scenario simulations," *Global and Planetary Change*, vol. 63, no. 2-3, pp. 152–162, 2008.

[18] B. Kamranzad, A. Etemad-Shahidi, V. Chegini, and S. Hadadpour, "Assessment of CGCM 3.1 wind field in the Persian Gulf," *Journal of Coastal Research*, vol. 65, pp. 249–253, 2013.

[19] B. Kamranzad, "Assessment of the changes in average wind speed in Chabahar, Gulf of Oman, due to climate change," *Journal Of Marine Engineering*, vol. 10, no. 19, pp. 13–20, 2014.

[20] C. W. Zheng, C. Y. Li, and X. Li, "Recent decadal trend in the North Atlantic wind energy resources," *Advances in Meteorology*, vol. 2017, Article ID 7257492, 8 pages, 2017.

[21] M. Segal, Z. Pan, R. W. Arritt, and E. S. Takle, "On the potential change in wind power over the US due to increases of atmospheric greenhouse gases," *Renewable Energy*, vol. 24, no. 2, pp. 235–243, 2001.

[22] P. B. Breslow and D. J. Sailor, "Vulnerability of wind power resources to climate change in the continental United States," *Renewable Energy*, vol. 27, no. 4, pp. 585–598, 2002.

[23] P. Lionello, A. Nizzero, and E. Elvini, "A procedure for estimating wind waves and storm-surge climate scenarios in a regional basin: the Adriatic Sea case," *Climate Research*, vol. 23, no. 3, pp. 217–231, 2003.

[24] G. P. Harrison and A. R. Wallace, "Sensitivity of wave energy to climate change," *IEEE Transactions on Energy Conversion*, vol. 20, no. 4, pp. 870–877, 2005.

[25] E. Charles, D. Idier, P. Delecluse, M. Deque, and G. Le Cozannet, "Climate change impact on waves in the Bay of Biscay, France," *Ocean Dynamics*, vol. 62, no. 6, pp. 831–848, 2012.

[26] B. Kamranzad, A. Etemad-Shahidi, V. Chegini, and A. Yeganeh-Bakhtiary, "Climate change impact on wave energy in the Persian Gulf," *Ocean Dynamics*, vol. 65, no. 6, pp. 777–794, 2015.

[27] H. Goharnejad, E. Nikaein, and W. Perrie, "Assessment of wave energy in the Persian Gulf: an evaluation of the impacts of climate change," *Oceanologia*, vol. 63, no. 1, pp. 27–39, 2021.

[28] M. A. Hemer, J. Katzfey, and C. E. Trenham, "Global dynamical projections of surface ocean wave climate for a future high greenhouse gas emission scenario," *Ocean Modelling*, vol. 70, pp. 221–245, 2013.

[29] R. Schnur and D. P. Lettenmaier, "A case study of statistical downscaling in Australia using weather classification by recursive partitioning," *Journal of Hydrology*, vol. 212-213, pp. 362–379, 1998.

[30] F. Giorgi and L. O. Mearns, "Approaches to the simulation of regional climate change: a review," *Reviews of Geophysics*, vol. 29, no. 2, pp. 191–216, 1991.

[31] D. J. Sailor, T. Hu, X. Li, and J. Rosen, "A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change," *Renewable Energy*, vol. 19, no. 3, pp. 359–378, 2000.

[32] H. J. Fowler, S. Blenkinsop, and C. Tebaldi, "Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling," *International Journal of Climatology*, vol. 27, no. 12, pp. 1547–1578, 2007.

[33] S. Ghosh and C. Misra, "Assessing hydrological impacts of climate change: modeling techniques and challenges," *The Open Hydrology Journal*, vol. 4, no. 1, pp. 115–121, 2010.

[34] D. Maraun, F. Wetterhall, A. M. Ireson et al., "Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user," *Reviews of Geophysics*, vol. 48, no. 3, Article ID RG3003, 2010.

[35] K. Srinivasa Raju and D. Nagesh Kumar, *Impact of Climate Change on Water Resources*, Springer, Singapore, 2018.

[36] R. L. Wilby and T. M. L. Wigley, "Precipitation predictors for downscaling: observed and general circulation model relationships," *International Journal of Climatology*, vol. 20, no. 6, pp. 641–661, 2000.

[37] R. M. Trigo and J. P. Palutikof, "Precipitation scenarios over Iberia: a comparison between direct GCM output and different downscaling techniques," *Journal of Climate*, vol. 14, no. 23, pp. 4422–4446, 2001.

[38] X. C. Zhang, "Spatial downscaling of global climate model output for site-specific assessment of crop production and soil erosion," *Agricultural and Forest Meteorology*, vol. 135, no. 1-4, pp. 215–229, 2005.

[39] J. Chen, F. P. Brissette, D. Chaumont, and M. Braun, "Performance and uncertainty evaluation of empirical downscaling methods in quantifying the climate change impacts on hydrology over two North American river basins," *Journal of Hydrology*, vol. 479, pp. 200–214, 2013.

[40] M. S. Khan, P. Coulibaly, and Y. Dibike, "Uncertainty analysis of statistical downscaling methods," *Journal of Hydrology*, vol. 319, no. 1-4, pp. 357–382, 2006.

[41] J. Chen, F. P. Brissette, D. Chaumont, and M. Braun, "Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America," *Water Resources Research*, vol. 49, no. 7, pp. 4187–4205, 2013.

[42] D. Mullan, J. Chen, and X. J. Zhang, "Validation of non-stationary precipitation series for site-specific impact assessment: comparison of two statistical downscaling techniques," *Climate Dynamics*, vol. 46, no. 3-4, pp. 967–986, 2016.

[43] J. M. Gutiérrez, D. Maraun, M. Widmann et al., "An intercomparison of a large ensemble of statistical downscaling methods over Europe: results from the VALUE perfect

predictor cross-validation experiment," *International Journal of Climatology*, vol. 39, no. 9, pp. 3750–3785, 2019.

[44] X. Zhang, M. Shen, J. Chen, J. W. Homan, and P. R. Busteed, "Evaluation of statistical downscaling methods for simulating daily precipitation distribution, frequency, and temporal sequence," *Transactions of the ASABE*, vol. 64, no. 3, pp. 771–784, 2021.

[45] M. A. Sunyer, H. Madsen, and P. H. Ang, "A comparison of different regional climate models and statistical downscaling methods for extreme rainfall estimation under climate change," *Atmospheric Research*, vol. 103, pp. 119–128, 2012.

[46] D. A. Sachindra, K. Ahmed, M. M. Rashid, S. Shahid, and B. Perera, "Statistical downscaling of precipitation using machine learning techniques," *Atmospheric Research*, vol. 212, pp. 240–258, 2018.

[47] A. Davanlou Tajbakhsh, V. Nourani, and A. Molajou, "Hybrid wavelet-M5 modeling in rainfall-runoff process forecast," *Iran Water Resources Research*, vol. 15, no. 2, pp. 1–10, 2019.

[48] V. Nourani, M. T. Alami, and M. H. Aminfar, "A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 3, pp. 466–472, 2009.

[49] V. Nourani, A. Molajou, S. Uzelaltinbulat, and F. Sadikoglu, "Emotional artificial neural networks (EANNs) for multi-step ahead prediction of monthly precipitation; case study: northern Cyprus," *Theoretical and Applied Climatology*, vol. 138, no. 3-4, pp. 1419–1434, 2019.

[50] V. Nourani, Z. Razzaghzadeh, A. H. Baghanam, and A. Molajou, "ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method," *Theoretical and Applied Climatology*, vol. 137, no. 3-4, pp. 1729–1746, 2019.

[51] D. Avila, G. N. Marichal, I. Padrón, R. Quiza, and A. Hernandez, "Forecasting of wave energy in canary islands based on artificial intelligence," *Applied Ocean Research*, vol. 101, Article ID 102189, 2020.

[52] S. H. Vakili, "Forecasting of monthly precipitation using M5 model tree and classic statistical methods (Case study: oroumieh synoptic station) (Technical note)," *IRAN-WATER RESOURCES RESEARCH*, vol. 13, no. 4, pp. 179–183, 2018.

[53] V. Nourani and A. Molajou, "Application of a hybrid association rules/decision tree model for drought monitoring," *Global and Planetary Change*, vol. 159, pp. 37–45, 2017.

[54] J. Mahjoobi and A. Etemad-Shahidi, "An alternative approach for the prediction of significant wave heights based on classification and regression trees," *Applied Ocean Research*, vol. 30, no. 3, pp. 172–177, 2008.

[55] S. Abolfathi, A. Yeganeh-Bakhtiary, S. M. Hamze-Ziabari, and S. Borzooei, "Wave runup prediction using M5′ model tree algorithm," *Ocean Engineering*, vol. 112, pp. 76–81, 2016.

[56] D. Ji, L. Wang, J. Feng et al., "Description and basic evaluation of beijing normal university Earth system model (BNU-ESM) version 1," *Geoscientific Model Development*, vol. 7, no. 5, pp. 2039–2064, 2014.

[57] S. Lehner and H. Günther, "Extreme wave statistics from radar data sets," in *Proceedings of the 2004 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1880–1883, Anchorage, AK, USA, September 2004.

[58] S. Wei, S. Yang, and D. Xu, "On accuracy of SAR wind speed retrieval in coastal area," *Applied Ocean Research*, vol. 95, Article ID 102012, 2020.

[59] P. A. E. M. Janssen, B. Hansen, and J.-R. Bidlot, "Verification of the ECMWF wave forecasting system against buoy and altimeter data," *Weather and Forecasting*, vol. 12, no. 4, pp. 763–784, 1997.

[60] R. P. da Rocha, S. Sugahara, and R. B. da Silveira, "Sea waves generated by extratropical cyclones in the South Atlantic Ocean: hindcast and validation against altimeter data," *Weather and Forecasting*, vol. 19, no. 2, pp. 398–410, 2004.

[61] S. Guinehut, C. Coatanoan, A.-L. Dhomps, P. Y. Le Traon, and G. Larnicol, "On the use of satellite altimeter data in argo quality control," *Journal of Atmospheric and Oceanic Technology*, vol. 26, no. 2, pp. 395–402, 2009.

[62] V. G. Polnikov, F. A. Pogarskii, N. S. Zilitinkevich, and A. A. Kubryakov, "Use of along-track altimeter data to verify numerical wave models," *Izvestiya - Atmospheric and Oceanic Physics*, vol. 55, no. 9, pp. 1089–1097, 2019.

[63] B. Oztunali Ozbahceci, A. R. Turgut, A. Bozoklu, and S. Abdalla, "Calibration and verification of century based wave climate data record along the Turkish coasts using satellite altimeter data," *Advances in Space Research*, vol. 66, no. 10, pp. 2319–2337, 2020.

[64] V. Nourani, A. Davanlou Tajbakhsh, A. Molajou, and H. Gokcekus, "Hybrid wavelet-M5 model tree for rainfall-runoff modeling," *Journal of Hydrologic Engineering*, vol. 24, no. 5, Article ID 04019012, 2019.

[65] S. Emmanouil, S. G. Aguilar, G. F. Nane, and J. J. Schouten, "Statistical models for improving significant wave height predictions in offshore operations," *Ocean Engineering*, vol. 206, Article ID 107249, 2020.

Hindawi
Complexity
Volume 2022, Article ID 5876035, 14 pages
https://doi.org/10.1155/2022/5876035

Research Article

A Comprehensive Method for Improving the Quality of Open Government Data and Increasing Citizens' Willingness to Use Data by Analyzing the Complex System of Citizens and Organizations

Mohammad Moradi,[1] Mojtaba Mazoochi,[1] and Mohammad Ahmadi [2]

[1]*ICT Research Institute, Tehran, Iran*
[2]*Faculty of Engineering and Computer Science, Department of Software engineering, Khatam Al-Nabieen University, Kabul, Afghanistan*

Correspondence should be addressed to Mohammad Ahmadi; mohammad.moradi@ut.ac.ir

Received 5 May 2022; Revised 3 June 2022; Accepted 3 August 2022; Published 21 August 2022

Academic Editor: Andrea Murari

In recent years, the amount of data in the world is growing rapidly. Data growth also occurs in the government sector. All ministries and institutions at every level are data producers. These government-owned data have a high potential if they can be used properly. Open government data can stimulate innovation and economic growth and enhance business models. In order to increase the willingness of citizens to use open government data and enjoy the benefits mentioned, the quality of open government data needs to be improved. The quality of open government data encompasses a variety of dimensions and criteria. Also, the importance of each dimension and criterion in increasing the quality of open government data is different. Therefore, we are faced with a complex system that requires proper decision-making and management. In fact, we are dealing with decision-making in the complex management system. Given the importance of this issue, the purpose of this study is to provide a new and comprehensive method to improve the quality of open government data and increase the willingness of citizens to use the data by considering the complex network of citizens and organizations. For this purpose, library studies have been used to extract comprehensive and effective dimensions and criteria. The statistical population includes all articles related to the criteria of improving the quality of open government data and increasing the willingness of citizens to use the data. The probabilistic sampling method of simple random samples has been used, and 10 articles in this field have been reviewed. After extracting the criteria as well as the data of 112 governmental organizations and institutions related to each criterion from the open data portal, the complex network of citizens and governmental organizations and institutions has been analyzed in order to identify high-degree centrality organizations. Then, the data characteristics of the organizations that were most desired by the citizens were extracted using data mining techniques including the regression model. Also, field method and multicriteria decision-making technique including the DEMATEL technique have been used to express the solutions and identify the cause-and-effect relationships between the solutions. The criteria extracted in improving the quality of open government data and increasing the willingness of citizens to use the data are included: "data originality," "license openness," "up-to-datedness," "data access," "metadata completeness," "number of data sets," "format openness," "nondiscrimination," "understandable," "number of categories of data sets," "free," "lack of missing data," "data request ability," "visualization," "feedback," and "data subject matter." Based on the results obtained from the analysis of the complex network and the regression model, the criterion of "society subject" with a coefficient of 72.564 and a positive sign has the greatest impact on increasing the number of citizens' visits to open government data. After that, the criterion of "format openness" with a coefficient of 52.682 and a positive sign has the second rank in increasing the number of visits. Extracting comprehensive and effective criteria in improving the quality of open government data and increasing citizens' willingness to use data, calculating the weight and importance of each criterion by analyzing the complex network of citizens and organizations, as well as providing solutions, can help managers in decision-making and proper management in the complex system of citizens and government organizations.

1. Introduction

Open data refer to nonconfidential data that are made available without any restrictions on use or distribution [1]. Open government data are tools for empowering citizens and giving them access and permission to use data generated by the government sector, so that they can use, store, redistribute, and integrate data with other data sources [2]. In addition, open government data can be defined as data belonging to a government entity that is published for free use, reuse, and redistribute [3].

Providing information in the form of open data will reduce corruption, gain public trust, and build a democratic society. Open data provide more opportunities to monitor governance activities. For example, it makes the way the budget is spent transparently and the effects clear. It also encourages citizens to be more involved in overseeing governance. In the corporate sphere, open data primarily help the entity itself to be aware of the existence of data in the organization and to avoid parallel and costly activities to collect data that have already been done by the organization. Open government data enable citizens to participate in decision-making processes with informed and structured procedures [4]. Open government data can stimulate innovation and economic growth and enhance business models [5].

In order to increase citizens' desire for open government data and enjoy the benefits mentioned, the quality of open government data needs to be improved. The quality of government data encompasses a variety of dimensions and criteria. Also, the importance of each dimension and criterion in increasing the quality of open government data is different. Therefore, we are faced with a complex system that requires proper decision-making and management. In fact, we are dealing with decision-making in complex management systems. In this study, first, comprehensive and effective dimensions and criteria in increasing the quality of open government data and increasing the willingness of citizens to use the data have been extracted using library studies. In the next phase, an attempt has been made to determine the extent of citizens' willingness to use the data of each organization by presenting a new and creative method and considering the complex network of citizens and government organizations and institutions providing open data. Then, the characteristics of the data that the citizens wanted were examined, and based on the obtained results, the importance and weight of the dimensions and effective criteria in increasing the quality of open government data were calculated. For this purpose, data mining techniques and regression model have been used. Also, after analyzing the results, multicriteria decision-making technique has been used to provide solutions and their importance. The method presented in this study and the results obtained can help managers in the decision-making and proper management of organizations providing open government data to increase data quality and thus increase the desire of citizens to use and enjoy the benefits of open data (as a kind of decision-making in the complex system of citizens and organizations providing open government data).

In the following, in Section 2, the related works and the disadvantages and limitations of previous researches and the reason for dealing with the current research are stated. Section 3 describes the research method, which consists of different phases and includes the analysis of a complex network of citizens and organizations to calculate the weight and importance of criteria for improving the quality of open government data and increasing citizens' willingness to use the data. In Section 4, the results are presented based on the different phases expressed in the research methodology section and the findings are discussed. Findings include the extraction of comprehensive dimensions and criteria in improving data quality and increasing citizens' willingness to use data, extraction of data of government organizations and institutions based on each criterion, results of complex network analysis of citizens and organizations and identification of organizations with high centrality, use of data mining techniques involves a regression model to analyze the characteristics of the data set related to organizations with high degree centrality and calculate the weight and importance of each characteristic and criterion, and to provide solutions and identify the cause-and-effect relationships between the solutions using decision-making trial and evaluation laboratory (DEMATEL) technique. Finally, in Section 5, the conclusion is given.

2. Related Works

In this section, research conducted to improve the quality of open government data and increase the willingness of citizens to use the data is reviewed. Nikiforova and McBride conducted a study entitled "Open government data portal usability: A user-centered usability analysis of 41 open government data portals." Confirming the importance of portal usability for the data reuse process, this study helps to explain some of the initial insights by asking two questions: "How can the usability of open government data portals be evaluated and compared in different contexts?" and "What are usually the practical aspects of open government data portals?". To answer these research questions, a set of 41 open government data portals have been selected for usability analysis based on the feedback of 40 users. According to the results of this study, the lack of interaction between users with open government data portals in cases such as providing feedback or requesting data sets is one of the main problems of open government data portals. Therefore, governments should focus on developing open government data ecosystems and increasing the interoperability of these portals [6].

Zhang and Xiao have examined the framework for evaluating the quality of open government data. The purpose of this study is to create a common framework as a reference for evaluating the quality of open government data. In this research, 10 qualitative studies have been combined in a

common reference framework to evaluate the quality of open government data. Based on a seven-step analysis, a common reference framework for evaluating the quality of open government data is presented, which includes six criteria: accuracy, accessibility, completeness, up-to-datedness, stability, and comprehensibility [7].

De Juana-Espinosa and Luján-Mora evaluated open government data portals in the EU from 2015 to 2017. This study presents data collected from open government data portals in 28 EU countries. Several parameters and criteria observed over a period of 3 years in open national data portals have been identified and recorded to create this data set. The data are obtained manually from existing public information sources and official open government data portals that are freely available on the Web. In this study, the criteria "Existence of a link from an open government data portal to the source site of the data set provider," "Existence of social network plugins" to discuss users' experiences in using an open government data portal, "Support for various data set formats" and "Data search and filtering capability" have been proposed as criteria for evaluating open government data portals [8].

Zheng et al. in a study entitled "Evaluating global open government data: Methods and status" first examines seven methods of evaluating open government data by regularly comparing and analyzing their frameworks, criteria, and methods. Based on this analysis, a framework for evaluating the performance of open government data has been developed for all UN member states. According to the results, most of the current evaluation programs focus on data and foundation and pay less attention to software platforms, use and impact. This study shows that in 2018, 34 countries (18%) score "very high," 40 countries (21%) score "high," 43 countries (22%) score "medium," while 76 countries (39%) have received a "low" score [9].

Dahbi et al. have conducted a study entitled "Toward an evaluation model for open government data portals." In this study, the authors define an evaluation model for open government data portals based on several main dimensions that have a great impact on their application. Specified dimensions are information richness that deals with adapting the portal to the needs of the user in terms of content; detection capability associated with tools and mechanisms that increase data access to the portal; reusability, which deals with the openness of the data published on the portal and the possibility of reusing them; and interaction, which is related to the openness of the portal for user feedback, cooperation, and interaction with published data. The proposed evaluation model has been used to evaluate four national open government data portals [10].

Vetrò et al. offer an approach to measuring the quality of open government data sets. They asked open government data users about the challenges to open government data quality. In general, they suggest that the data set be evaluated for completeness, accuracy, traceability, comprehensibility, compliance, and expiration. In other words, the ideal data set should include complete and accurate data, be machine-readable, have metadata, be updated, be accurate, and be traceable in terms of source. This is a very complete and useful framework for policymakers who want to control the quality of data sets in open government projects or programs [11].

Dawes et al. introduced a framework for evaluating the quality of open data portals at the national level and provided a set of criteria for evaluating data quality problems in open government data portals. These criteria were applied to 12 portals, and several dimensions of data quality were introduced. These dimensions included the existence of standards in data formats, the existence of metadata, machine readability, and the up-to-datedness of data [12].

Misuraca and Viscusi discuss the framework for evaluating the compliance of open government data based on quality. These criteria include three dimensions of quality: completeness, accuracy, and up-to-datedness [13]. Harrison et al. focus on evaluating metadata quality [14].

As seen in previous research, each of the studies focused on a specific dimension of improving the quality of open government data and increasing citizens' willingness to use the data. Also, in the researches done so far, the weight and importance of each of the dimensions and criteria have not been specified. Considering the weight and importance of each dimension and criterion based on the degree of willingness of citizens to use the data so far has not been considered in previous research. In this study, different and comprehensive dimensions and criteria in improving the quality of open government data and increasing the willingness of citizens to use data as well as calculating the importance of each criterion based on the analysis of a complex network of citizens and open data providers are presented. Also, after reviewing the obtained results, the solutions and the causal relationships between them are discussed.

3. Materials and Methods

In this research, the type of research based on the purpose is applied research. In the first phase, library studies are conducted to extract comprehensive dimensions and criteria for improving the quality of open government data and increasing citizens' willingness to use the data. The statistical population includes all articles related to the dimensions and criteria for improving the quality of open government data and increasing the willingness of citizens to use the data. The probabilistic sampling method of simple random samples has been used, and 10 articles in this field have been reviewed.

In the second phase, the organizations, including all government organizations and institutions present in the open data portal (https://data.gov.ir/), are examined and the data related to each criterion extracted in the first phase are calculated for each organization. The number of government organizations and institutions surveyed was 112.

In the third phase, the complex network of citizens and government organizations and institutions providing open data is analyzed. This network is a directional network. The nodes in this network are citizens and government organizations and institutions, and the links represent data visits. For example, in Figure 1, the citizen nodes are shown in

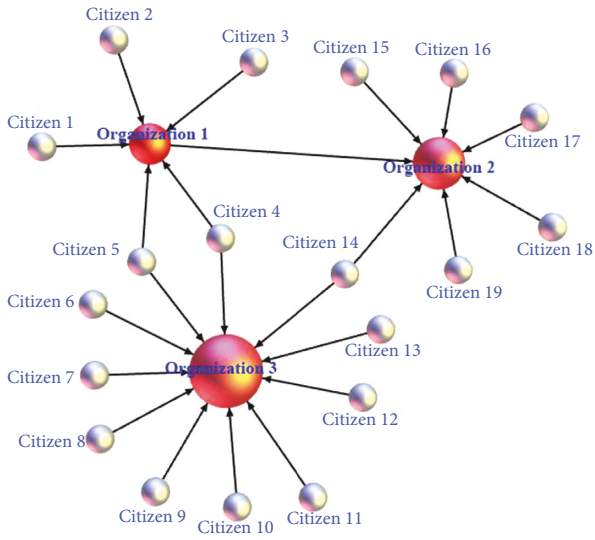


Figure 1: A complex network of citizens and government organizations and institutions.

silver and the organizations nodes are shown in red. The link from citizen i to organization j indicates that citizen i has visited the published open data of organization j. In this network, there can be directional links from citizen to an organization or organization to another organization. But there is no possibility of directional link between citizens or organization to citizen in this network. It is also possible to create a link from one citizen to several organizations, meaning that one citizen can view data from multiple organizations.

Degree centralization has been used to identify organizations whose data are more desirable. The absolute degree centrality of node vi is calculated as follows [15]:

$$c_D(i) = \text{degree of vertex } i. \tag{1}$$

The relative degree centrality of node vi can be calculated as follows [15]:

$$C_D(i) = \frac{c_D(i)}{(n-1)}. \tag{2}$$

In formula (2), $n-1$ is the largest possible degree of a network with n nodes. This information can be extracted through the open data portal based on the number of visits to the data set that each organization has provided openly.

In the fourth phase, data mining techniques were used to extract the data characteristics of the organizations that were more interested (had a higher degree centrality). For this purpose, regression model has been used. The criteria extracted in the first phase were used as attributes, and the data extracted in the second phase, which are the data of organizations based on each criterion, were used as a record. The "visit rate" attribute has been used as a label. The output of this model is the coefficients related to each quality criterion that determine the weight and importance of that criterion.

In the fifth phase, the results are reviewed and solutions are presented. For this purpose, field methods and multicriteria decision-making techniques have been used. The DEMATEL technique was used to identify the cause-and-effect relationships between the solutions. The research process diagram is shown in Figure 2.

Also, the reasons for using each method stated in the research process are summarized in Table 1.

4. Results and Discussion

This section deals with the results of the five phases described in the previous section and discusses them.

4.1. Extracting Comprehensive and Effective Criteria in the Quality of Open Government Data and Increasing the Willingness of Citizens to Use the Data. In this section, comprehensive and effective criteria for the quality of open government data and increasing the willingness of citizens to use data in three dimensions of open data, data transparency, and interaction are stated. References related to each dimension and criterion are also specified.

4.1.1. Open Data

(1) Data Accuracy [*10*, *16*]. This dimension deals with the originality of the data, the absence of lost data, and the up-to-datedness. Missing data prevent the use and reuse of data and have a major impact on the quality of programs that reuse data. The following criteria relate to the data accuracy dimension:

(a) Data originality [11, 16].

(b) Lack of missing data [10].

(c) Up-to-datedness [10, 16, 17]: this criterion evaluates the up-to-datedness of the data in the published data set. For each O_i organization, this score is calculated based on data published in the last five years using formula (3). T_j represents the number of data published in j years ago.

$$O_i = \sum_{j=0}^{4} \left(1 - \frac{j}{10}\right) \times T_j. \tag{3}$$

(2) Discoverability [*10*]. Discoverability dimension deals with tools and mechanisms that increase data access and search. In other words, users should be able to search and access the relevant data set in a simple and efficient way. This will not happen if the metadata is not provided. Metadata provides a better understanding of the importance of data and data structure and helps users access the data they need [18]. Assessing the discoverability dimension requires a thorough evaluation of the descriptive metadata and the availability of data access features. The following criteria relate to the discoverability dimension.

(a) Metadata completeness [9, 10, 17, 19]: this criterion evaluates the completeness of descriptive metadata. For each data set, the completeness of the descriptive metadata fields is evaluated. These fields include title, description, tags, publisher, and more.

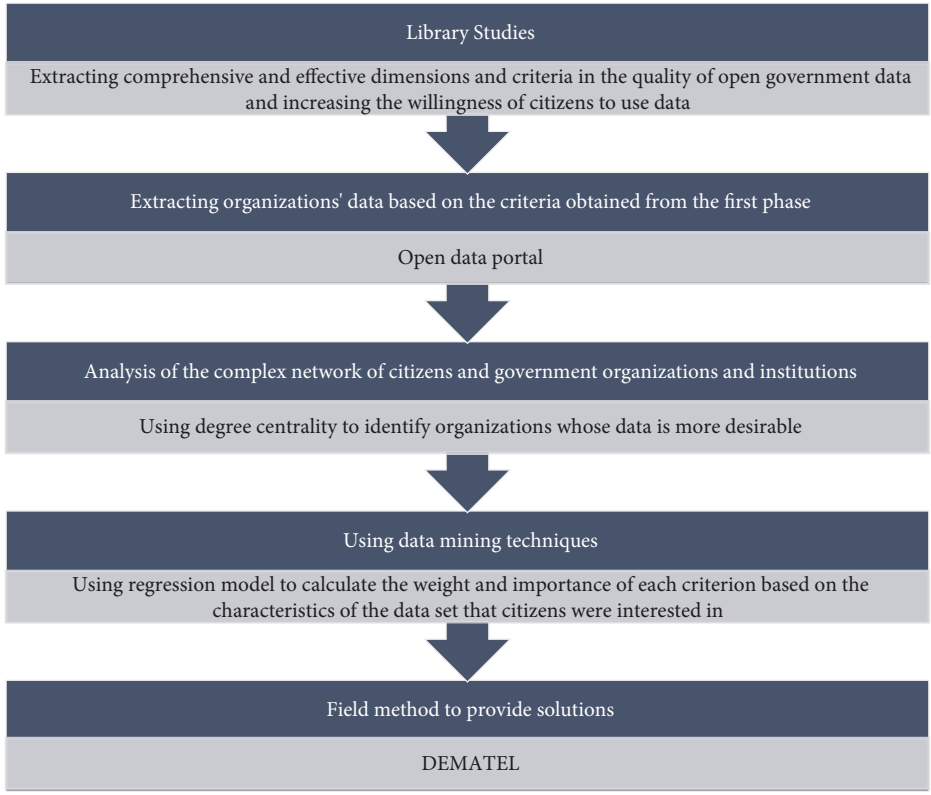


Figure 2: Diagram of the research process.

Table 1: Reasons for using each method in the research process.

Phase	Method	Reason for using the method
1	Library studies	The reason for using library studies has been to extract comprehensive and effective dimensions and criteria in the quality of open government data.
2	Extracting organizations' data based on the criteria obtained from phase 1	The reason for extracting the data of the organizations in the open data portal was to identify the characteristics of each organization. In fact, the data of this phase constitute attributes in phase 4.
3	Analysis of the complex network of citizens and organizations	In addition to the data extracted in phase 2, which identifies the characteristics of each organization, it is necessary to identify the organizations whose data have been most desired and used by citizens. For this purpose, the indegree of each organization in the complex network of citizens and organizations, which indicates the number of visits to the data of the organization, has been used. In fact, the data of this phase form the labels in phase 4.
4	Using data mining techniques	After extracting the data from phase 2 as attributes and the data from phase 3 as labels, it is necessary to analyze the data in order to identify the positive or negative impact of each attribute on the label (number of visits) and also to calculate the weight and importance of each attribute.
5	Field method to provide solutions	The reason for using this phase is to extract solutions and identify cause-and-effect relationships between the solutions in order to improve citizens' interaction with open government data. Identifying the cause-and-effect relationships between solutions will help managers spend more time and cost on the solutions that have the greatest impact on other solutions.

(b) Data access [9–11, 17]: this criterion evaluates the existence of attributes that increase data discovery, in particular, the existence of three attributes: search, sort, and filter, which receive their value in the range [0, 1] according to the existence of these attributes.

(3) Richness of Information [10]. Richness of information measures the extent to which a user needs are met in terms of the amount of data. The following criteria are related to the richness of information dimension.

(a) Number of data sets [10, 11, 17, 19]: this criterion evaluates the number of data sets that an organization has openly provided.

(b) Number of categories of data sets [19].

(c) Data subject matter [19]: citizens may be more interested in some issues. Based on the open data portal survey, the subjects of the entire data set were extracted as follows:

(i) Heights
(ii) Planning
(iii) Water effects
(iv) Animal and plant ecology
(v) Borders
(vi) Images/maps/land cover
(vii) Location
(viii) .Weather
(ix) Society
(x) Health
(xi) Management
(xii) Environment
(xiii) Farming
(xiv) Science and research education
(xv) Energy
(xvi) Structure
(xvii) Economy
(xviii) Transportation
(xix) Earth sciences

(d) Data request ability [7, 10]: this criterion measures the degree of openness to user requests. In other words, it examines the possibility of allowing users to request new data sets. Depending on the availability of the data request, two values of 1 or 0 are assigned.

4.1.2. Data Transparency

(1) Reusability [*10*, *16*]. The value of open government data is realized only after its reuse [3]. Open government data are considered reusable when the data are released under an open license and there is unrestricted access, reuse, and redistribution of data. It must also be published electronically and machine-readable. Reusability also deals with features that provide an easy way to reuse data, such as applications and the API (Application Programming Interface). The following criteria relate to the reusability dimension:

(a) License openness [10, 16]: this criterion evaluates the openness of the data set license for reuse.

(b) Format openness [10, 16, 19]: this criterion evaluates the openness of the data format. For each D_n data set, the FOI_n score based on the data format is assigned as follows:

- If the format is machine-unreadable: $FOI_n = 0$ (e.g., PDF).
- If the format is machine-readable: $FOI_n = 1$ (e.g., JSON, CSV).

(c) Free [16].

(d) Nondiscriminatory [11, 16]: access to and reuse of data are the same for all individuals and legal entities.

(2) Understandable [16]

4.1.3. Interactivity

(1) Feedback [*10*, *16*, *17*]. This criterion examines the existence of features related to collaboration, feedback, and evaluation and assesses the existence of three possibilities: commenting on the data set, ranking the data set, and feedback on the portal.

(2) Visualization [*10*, *17*, *19*]. This criterion evaluates the existence of visualization tools and features such as maps, diagrams, or programs for visualizing and interacting with data.

4.1.4. Chart of Comprehensive and Effective Criteria in the Quality of Open Government Data and Increase the Willingness of Citizens to Use the Data. Figure 3 shows a chart of comprehensive and effective criteria expressed in the quality of open government data and the increasing willingness of citizens to use the data based on the dimension, criteria, reference, and year of publication of the reference.

4.2. Extracting Organizations' Data Based on the Extracted Criteria. In this phase, the data of government organizations and institutions present in the open government data portal are extracted based on the criteria extracted in the previous phase. The number of organizations in this portal that provide open data was 112. Figure 4 shows a data extraction of government organizations and institutions present in the open government data portal based on the stated criteria.

4.3. Analysis of the Complex Network of Citizens and Government Organizations. In this phase, the complex network of citizens and government organizations and institutions present in the open government data portal were examined. In order to identify organizations and government institutions with a high degree centrality, input links from citizens to organizations that represent the indegree of each organization were calculated. This information was available based on the number of visits to each organization's data set in the open government data portal. Figure 5 shows a view of the extracted data, the number of visits to each organization, and its data set.

4.4. Using Data Mining Techniques to Calculate the Weight and Importance of Each Criterion. After extracting comprehensive and effective criteria in the quality of open government data and increasing the willingness of citizens to use the data, as well as extracting data from government organizations based on these criteria and calculating the number of visits to each organization's data set based on the complex network analysis of citizens and organizations, the weight and importance of each criterion need to be determined. As mentioned earlier in this study, we intend to

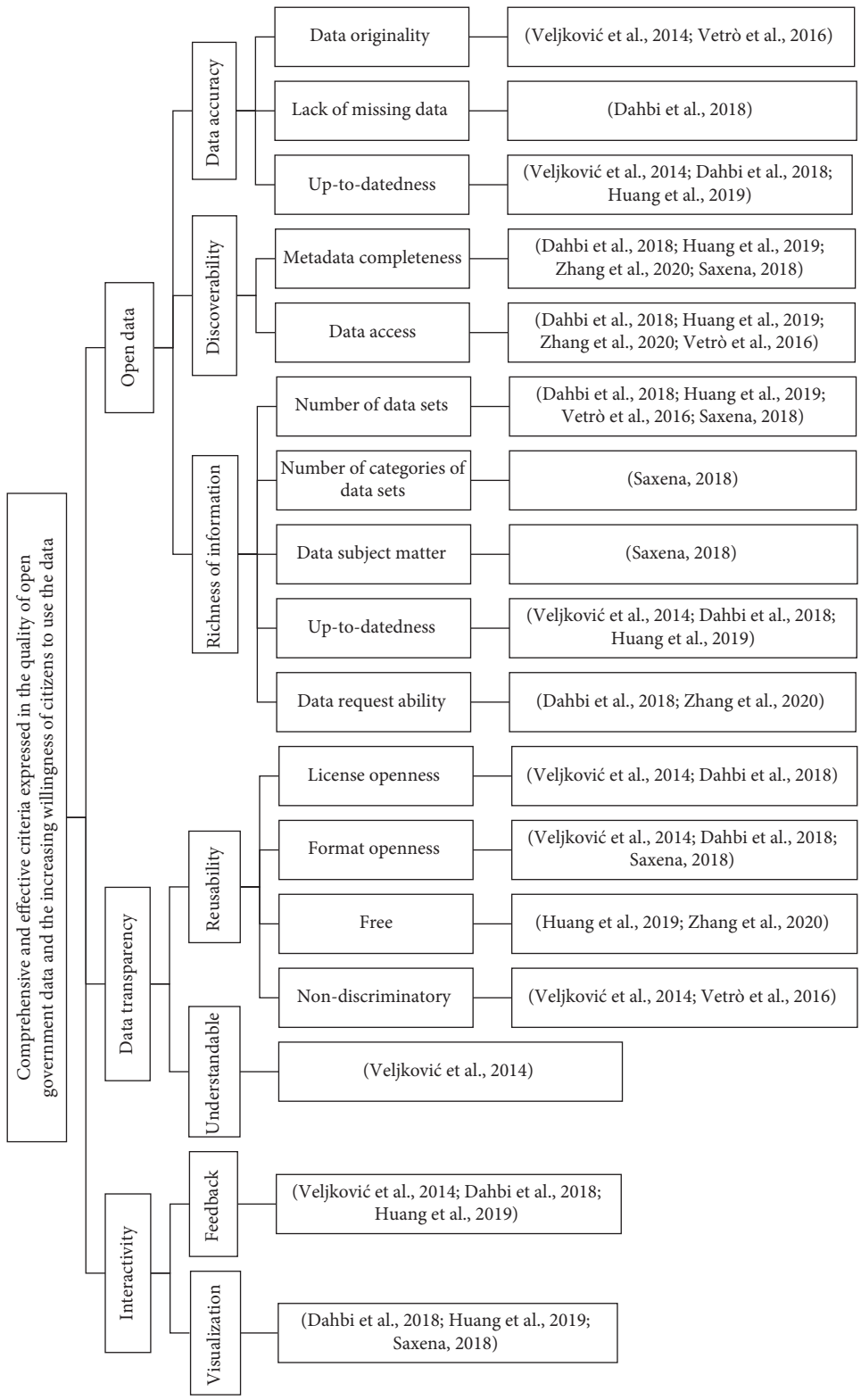


Figure 3: Chart of effective criteria in the quality of open government data and increasing the willingness of citizens to use data based on dimension, criteria, reference, and year of reference.

calculate the weight and importance of each criterion based on the degree of willingness that citizens have so far from the data set of each organization. In this way, first the data sets of the organizations that have been desired by the citizens are identified. Then, the characteristics of this data set are extracted, and based on this, the weight and importance of each criterion are determined. Therefore, the criteria extracted in the first phase are considered as attributes. The

I	H	G	F	E	D	C	B	A
)ata request ability	Up-to-datedness	Visualization	Format openness	Up-to-datedness	Data access	Number of categories of data sets	Number of data sets	Criterion name / Organization name
0	1765.6	1762	0	1765.6	223	3	2098	Geology organization
0	516	0	645	516	645	8	645	Statistical Center
0	235.9	36	226	235.9	226	1	337	Roads and Transportation Organization
0	73.5	0	0	73.5	0	5	81	Iran Health Insurance Organization
0	82.7	43	53	82.7	54	11	97	mmunication Technology Organization of
0	103.6	0	16	103.6	16	1	107	The official newspaper of the country
0	27.3	39	0	27.3	0	2	39	Lake Urmia Rehabilitation Headquarters
0	61	70	0	61	0	2	70	stitute of Communication and Information
0	46.8	52	0	46.8	0	6	52	ousing and Urban Development Research
0	28.3	46	0	28.3	0	10	46	Isfahan Regional Water Company
0	15.1	6	4	15.1	4	3	16	Social Security Organization
0	28.8	31	0	28.8	0	2	32	Railway Research Center
0	30.9	3	0	30.9	0	1	32	/ice President for Science and Technolog
0	18.9	20	0	18.9	0	1	20	National mapping agency
0	20.3	29	0	20.3	0	5	29	Space Agency
0	13.5	0	15	13.5	15	1	15	zation of Municipalities and Villages of the
0	9	0	0	9	0	1	10	Literacy Movement Organization
0	12.6	0	0	12.6	0	1	14	Airports and Air Navigation Company
0	13	0	0	13	0	1	13	Corporate Audit
0	22.8	0	0	22.8	1	3	24	Nomadic Affairs Organization
0	11.5	0	0	11.5	0	1	12	Space Research Institute

Figure 4: View of the extracted data of organizations based on each criterion.

H	G	F	E	D	C	B	A
ng data	Data originality	Data access	Metadata completeness	Data request ability	Number of categories of data sets	Number of data sets	Number of visits	Criterion name / Organization name
	2098	223	2098	0	3	2098	141118	Geology organization
	645	645	645	0	8	645	10355	Statistical Center
	337	226	337	0	1	337	6171	Roads and Transportation Organization
	81	0	81	0	5	81	4809	Iran Health Insurance Organization
	97	54	97	0	11	97	3670	mmunication Technology Organization of
	107	16	107	0	1	107	3258	The official newspaper of the country
	39	0	39	0	2	39	3163	Lake Urmia Rehabilitation Headquarters
	70	0	70	0	2	70	2815	stitute of Communication and Information
	52	0	52	0	6	52	2346	ousing and Urban Development Research
	46	0	46	0	10	46	2038	Isfahan Regional Water Company
	16	4	16	0	3	16	1757	Social Security Organization
	32	0	32	0	2	32	1729	Railway Research Center
	32	0	32	0	1	32	1585	/ice President for Science and Technolog
	20	0	20	0	1	20	1319	National mapping agency
	29	0	29	0	5	29	1307	Space Agency
	15	15	15	0	1	15	1264	zation of Municipalities and Villages of the
	10	0	10	0	1	10	1202	Literacy Movement Organization
	14	0	14	0	1	14	1114	Airports and Air Navigation Company
	13	0	13	0	1	13	1095	Corporate Audit
	24	1	24	0	3	24	1081	Nomadic Affairs Organization
	12	0	12	0	1	12	1066	Space Research Institute

Figure 5: View of the extracted data, the number of visits to each organization, and its data set.

records include the data extracted from each organization based on these criteria, which were extracted in the second phase. The "number of visits" attribute has been used as a label. Also, the names of government organizations and institutions are considered as ID. In this phase, the regression model is used to analyze the data and calculate the weight and importance of each criterion. The output of this model is the coefficients that determine the weight and importance of each criterion. Rapid Miner software has been used for this purpose. Figure 6 shows the operators used in this software.

The Split Data operator is used to divide the data into two sets of training and testing data. 70% of the data were considered as training data, and the remaining 30% as test data. Figure 7 shows the training data, including attributes, records, IDs, and labels. The number of training data was 78, which were randomly selected from a total of 112 available cases.

The linear regression operator is used to analyze the data and calculate the regression coefficients. Figure 8 shows the coefficients. The positive sign of the coefficient indicates the positive effect of the specified criterion on the label, which is the number of visits to the data set. The more positive and larger the coefficient of a criterion, the greater its impact on increasing the number of citizens visiting open government data. A negative sign indicates the negative impact of a specified criterion on the number of visits. The negative and larger the coefficient of a criterion, the greater its impact on reducing the number of visits to open government data.

As can be seen in Figure 8, the "society subject" criterion, with a coefficient of 72,564 and a positive sign, had the greatest impact on the label, which is an increase in the number of visits to open government data. This means that citizens were more interested in data sets that were relevant to the society. After that, the criterion of "format openness" with a coefficient of 52.682 and a positive sign has the second rank in increasing the number of visits. Therefore, citizens were more interested in data sets that can be read by a machine. Criteria for "metadata completeness," "number of data sets," "understandable," "data originality," "free," "lack of missing data," "nondiscriminatory," and "open license" being a positive sign, they gained the next ranks in increasing the number of visits to open government data. The important point is that the criterion of "number of categories of data sets" has a negative sign, which means that citizens are more inclined to data set that is more focused on a particular subject. Also, the "farming subject" with a negative sign and a coefficient of 160.413 had the most negative impact on the number of open government data views. This means that citizens were reluctant to visit the farm data set.

In order to evaluate the regression model and to evaluate the accuracy of the obtained coefficients, test data

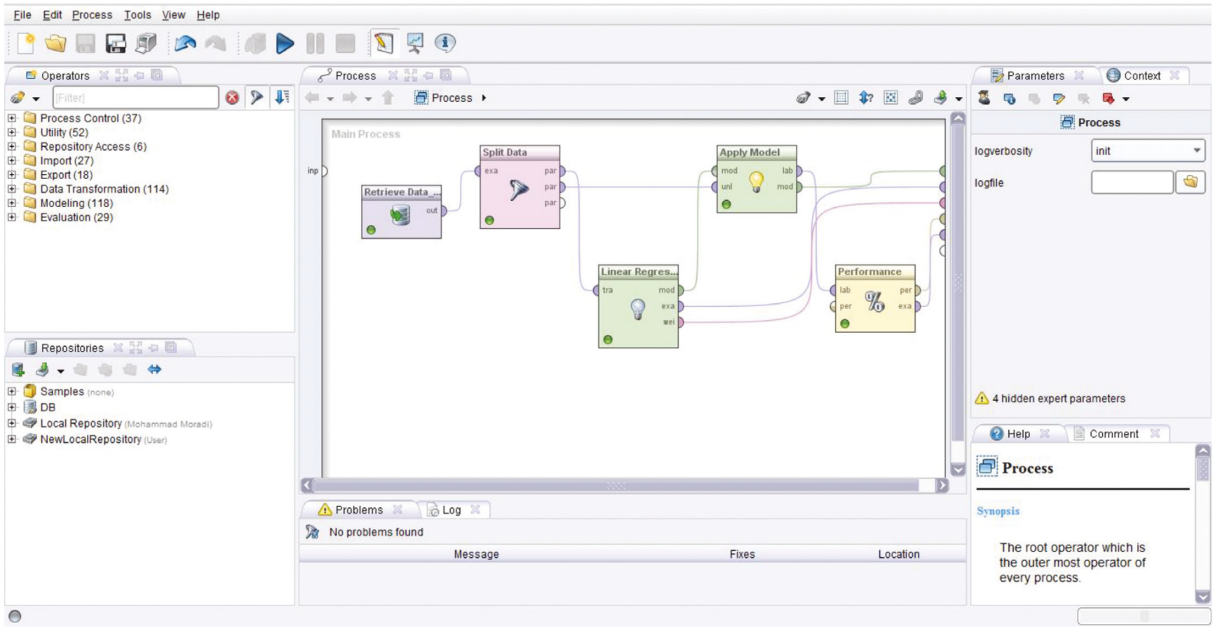


Figure 6: Operators used in Rapid Miner software to calculate regression coefficients.

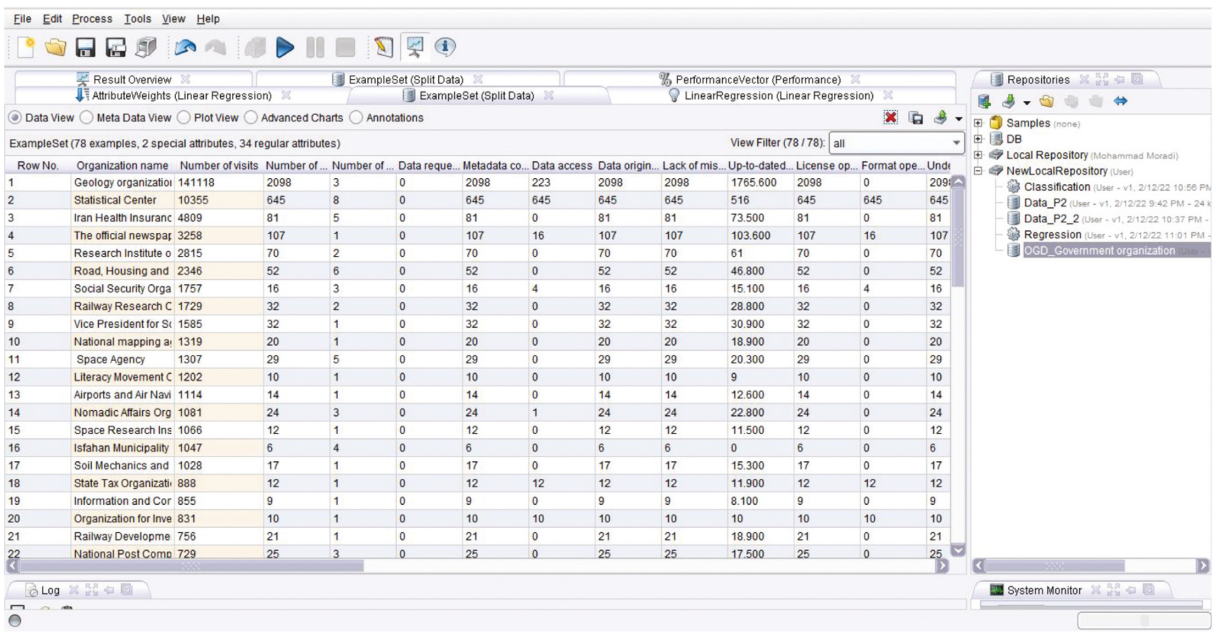


Figure 7: Training data.

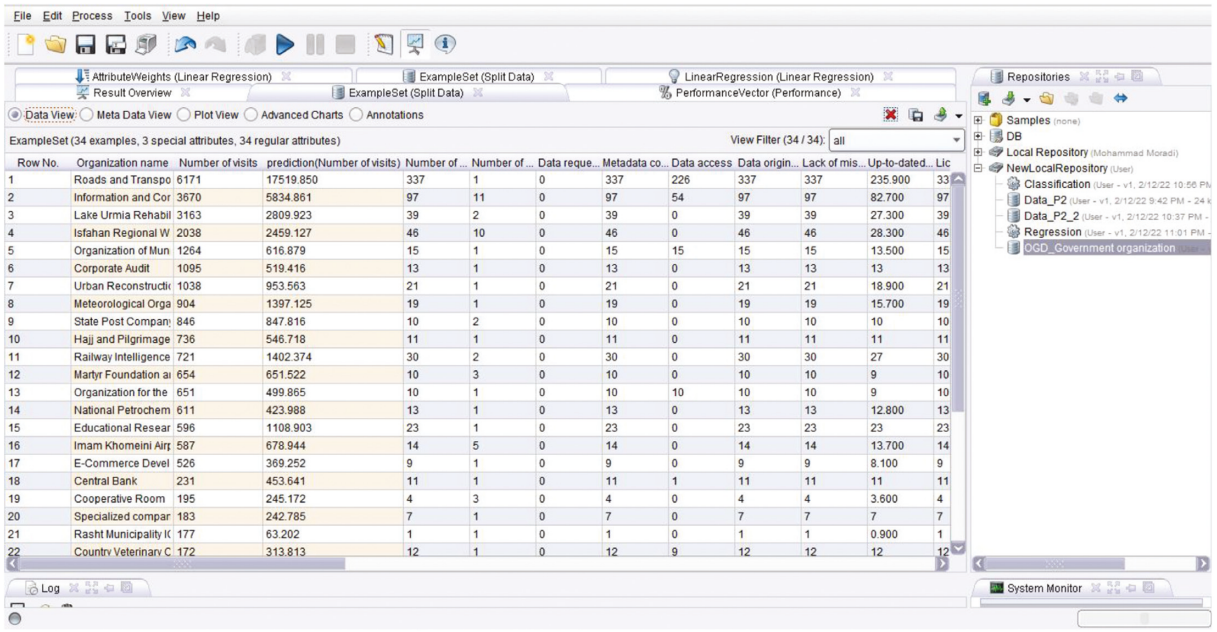
were used. 30% of the total data were used as test data. The number of test data was 78, which were randomly selected from a total of 112 items. Figure 9 shows the test data with the predicted rate for the label, which is the number of visits of open government data using the regression model generated. The actual number of visits is also shown.

Root-mean-square error (RMSE) was also used to evaluate the generated regression model, which was equal to 1998.435. According to the obtained RMSE, the value of normal root-mean-square error (NRMSE) was 0.014 (1.4%), which, according to [20] because it is less than 10%, indicates the desired state of the regression model.

4.5. Solutions. In this section, solutions to increase citizens' visits to organizations' data sets are stated. This section includes extracting solutions through library studies, as well as identifying the cause-and-effect relationships between the solutions based on the DEMATEL technique.

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
Number of data sets	10.349	9.285	53.912	0.051	1.115	0.385	
Number of categories of data sets	-24.826	344.183	-21.768	0.979	-0.072	0.943	
Metadata completeness	11.196	9.285	58.323	0.051	1.206	0.322	
Data access	-43.375	7.264	-263.771	0.856	-5.971	0.000	****
Data originality	8.719	9.285	45.421	0.051	0.939	0.360	
Lack of missing data	8.306	9.285	43.268	0.051	0.895	0.383	
License openness	8.053	9.285	41.949	0.051	0.867	0.397	
Format openness	52.682	7.088	401.612	0.997	7.432	0	****
Understandable	8.773	9.285	45.699	0.051	0.945	0.357	
Free	8.402	9.285	43.767	0.051	0.905	0.377	
Non-discriminatory	8.085	9.285	42.117	0.051	0.871	0.395	
Borders subject	-77.242	245.458	-352.390	1.000	-0.315	0.757	
Images / Maps / Land cover	-12.094	167.598	-52.889	1.000	-0.072	0.943	
Location subject	-27.968	220.930	-145.792	1.000	-0.127	0.901	
Society subject	72.564	124.799	384.651	0.998	0.581	0.569	
Health subject	-9.888	71.144	-72.756	1.000	-0.139	0.891	
Management subject	-42.155	42.181	-270.373	1.000	-0.999	0.330	
Farming subject	-160.413	31.526	-1406.931	0.997	-5.088	0.000	****
Science and research education su	-25.032	80.365	-90.318	1.000	-0.311	0.760	
Energy subject	-41.681	83.672	-158.187	0.999	-0.498	0.625	
Structure subject	-59.139	22.011	-430.373	0.997	-2.687	0.011	**
Economy subject	-34.340	22.894	-203.815	0.997	-1.500	0.177	
Transportation subject	-26.228	93.489	-68.123	1.000	-0.281	0.783	
(Intercept)	56.208	∞	?	?	0	1	

Figure 8: Regression coefficients based on each criterion.



Row No.	Organization name	Number of visits	prediction(Number of visits)	Number of ...	Number of ...	Data reque...	Metadata co...	Data access	Data origin...	Lack of mis...	Up-to-dated...	Lic
1	Roads and Transpo	6171	17519.850	337	1	0	337	226	337	337	235.900	33
2	Information and Cor	3670	5834.861	97	11	0	97	54	97	97	82.700	97
3	Lake Urmia Rehabil	3163	2809.923	39	2	0	39	0	39	39	27.300	39
4	Isfahan Regional W	2038	2459.127	46	10	0	46	0	46	46	28.300	46
5	Organization of Mun	1264	616.879	15	1	0	15	15	15	15	13.500	15
6	Corporate Audit	1095	519.416	13	1	0	13	0	13	13	13	13
7	Urban Reconstructi	1038	953.563	21	1	0	21	0	21	21	18.900	21
8	Meteorological Orga	904	1397.125	19	1	0	19	0	19	19	15.700	19
9	State Post Compan	846	847.816	10	2	0	10	0	10	10	10	10
10	Hajj and Pilgrimage	736	546.718	11	1	0	11	0	11	11	11	11
11	Railway Intelligence	721	1402.374	30	2	0	30	0	30	30	27	30
12	Martyr Foundation a	654	651.522	10	3	0	10	0	10	10	9	10
13	Organization for the	651	499.865	10	1	0	10	10	10	10	9	10
14	National Petrochem	611	423.988	13	1	0	13	0	13	13	12.800	13
15	Educational Resear	596	1108.903	23	1	0	23	0	23	23	23	23
16	Imam Khomeini Airp	587	678.944	14	5	0	14	0	14	14	13.700	14
17	E-Commerce Devel	526	369.252	9	1	0	9	0	9	9	8.100	9
18	Central Bank	231	453.641	11	1	0	11	1	11	11	11	11
19	Cooperative Room	195	245.172	4	3	0	4	0	4	4	3.600	4
20	Specialized compar	183	242.785	7	1	0	7	0	7	7	7	7
21	Rasht Municipality I	177	63.202	1	1	0	1	0	1	1	0.900	1
22	Country Veterinary C	172	313.813	12	1	0	12	9	12	12	12	12

Figure 9: Test data with the predicted value of the label and the actual value of the label.

4.5.1. Extraction of Solutions. Most of the challenges, such as the lack of up-to-date data, the small number of data sets presented, the lack of items such as data visualization, etc., are that government organizations are reluctant to share their data. The following are solutions and incentives to encourage government organizations to share data.

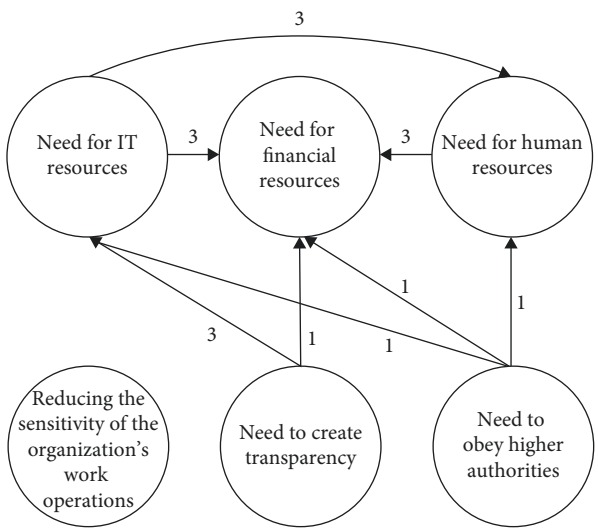


Figure 10: The direct-relation graph of solutions.

(1) Need for Human Resources [*21–23*]. External human resources, such as expertise and ideas, are vital to government organizations pursuing service innovation. Human resources include the manpower, ideas, knowledge, and skills needed to achieve the goals of the organization. Government organizations are looking for human resources and individuals to come up with ideas to help innovate public services. Therefore, it is necessary for government organizations to increasingly trust the expertise of foreign innovators to innovate services. Therefore,

(i) The relative need for human resources has a positive relationship with the dependence of the government organizations on foreign innovators. The greater this dependence, the more organizations will have to provide government data more openings to attract foreign innovators.

(2) Need for Financial Resources [*21*, *24*, *25*]. Financial resources here refer to the monetary capital needed by government organizations to achieve their goals of service innovation. Therefore, government organizations may rely on external sources for financial assistance. Therefore,

(i) The need for financial resources is positively related to the dependence of government organizations on external resources. The greater this dependence, the more organizations will have to provide more open government data to attract external financial resources.

(3) Need for IT Resources [*21*]. The use of IT resources for innovative activities in the services of government organizations is very important. For example, IT resources such as sensors and servers are essential for organizations seeking to innovate in smart city services. Government organizations may need external resources to provide IT resources for service innovation. Therefore,

(i) The relative need for IT resources has a positive relationship with the dependence of the government organizations on external resources. The dependence of government organizations on external sources is also positively related to the dissemination of open data.

(4) Need to Obey Higher Authorities [*21*, *26*]. One of the main factors in the involvement of government organizations in providing open data is higher authorities (for example, local and national governments) who implement formal and informal policies to influence organizations in data sharing. Such regulations or even the informal policies of higher authorities create institutional pressures on government organizations that shape their behavior. Therefore,

(i) The need for government organizations to follow higher authorities is positively related to its open data publishing behavior.

(5) Need to Create Transparency [*21*, *27*]. It is expected that government organizations' need for transparency will affect their open data behaviors. This is the freedom of information (FOI), which is recognized by the United Nations as a fundamental human right. When government organizations share their data with the public through open government data initiatives, citizens can control their activities and thus meet their need for transparency. Therefore,

(i) The need for transparency in government organizations is positively related to its open data publishing behavior.

(6) Reducing the Sensitivity of the Organization's Work Operations [*21*, *28–30*]. The specific work operations of a government organization affect its data-opening behavior. Some government organizations hold and process sensitive information because of their operations in government, which can restrict the provision of open government data. For example, government organizations related to health care need to process private information such as patients' medical records and background information. Similarly, government organizations working in national security, such as defense agencies, deal with limited information. Accordingly, the more sensitive the operations of organizations, the less data sharing will result. Therefore,

(i) The sensitivity of a government organization's operation is negatively related to its open data sharing behavior.

4.5.2. Analysis of Cause-and-Effect Relationships between Solutions Based on DEMATEL Technique. In this section, the solutions are examined based on the cause-and-effect relationships they have with each other. Identifying the cause-and-effect relationships between solutions will help managers invest more in solutions that have a greater impact on other solutions. The DEMATEL technique has been used for this purpose. Figure 10 shows the direct-relation graph of solutions based on the DEMATEL technique.

Table 2: The degree of causality and effectiveness between the solutions.

Solutions	R Degree of causality	J Degree of effectiveness	$(R+J)$ Sum of causality and effectiveness	$(R-J)$ $R-J>0 \longrightarrow$ definite cause $R-J<0 \longrightarrow$ definite effect
Need for human resources	0.5	1	1.5	−0.5
Need for financial resources	0	2.167	2.167	−2.167
Need for IT resources	1.25	0.667	1.917	0.583
Need to obey higher authorities	0.701667	0	0.701667	0.701667
Need to create transparency	1.291667	0	1.291667	1.291667
Reducing the sensitivity of the organization's work operations	0	0	0	0

Table 3: Ranking of solutions based on the degree of causality, the degree of effectiveness, and the degree of interaction with other factors.

Rank	Ranking based on the degree of causality	Ranking based on the degree of effectiveness	Ranking based on the degree of interaction with other factors
1	Need to create transparency	Need for financial resources	Need for financial resources
2	Need for IT resources	Need for human resources	Need for IT resources
3	Need to obey higher authorities	Need for IT resources	Need for human resources
4	Need for human resources	Need to obey higher authorities	Need to create transparency
5	Need for financial resources	Need to create transparency	Need to obey higher authorities
6	Reducing the sensitivity of the organization's work operations	Reducing the sensitivity of the organization's work operations	Reducing the sensitivity of the organization's work operations

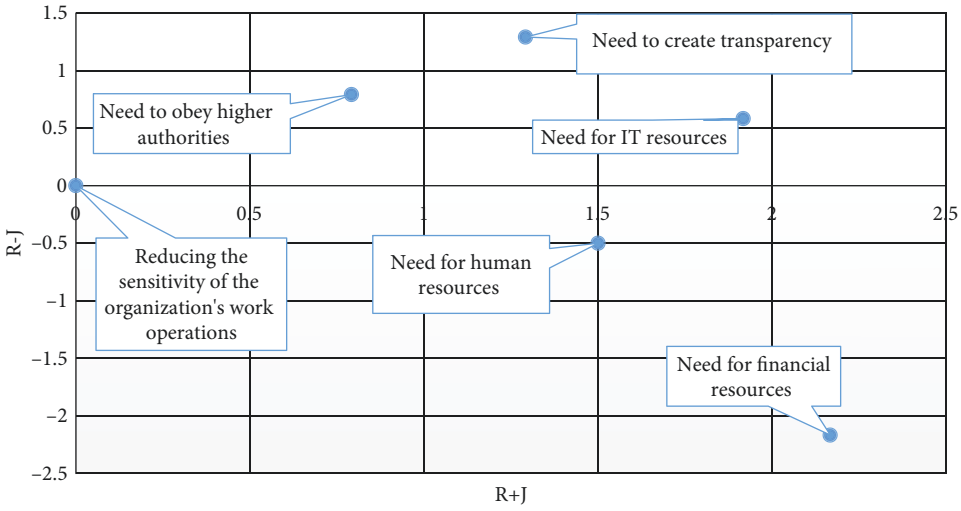


Figure 11: Impact chart between solutions.

The degree of causality and effectiveness between the solutions are shown in Table 2.

The ranking of solutions according to the degree of causality, the degree of effectiveness, and the degree of interaction with other factors is shown in Table 3.

According to Table 3, the "need to create transparency" solution has the greatest impact on other solutions. The "need for financial resources" solution is most affected by other factors. Also, the "need for financial resources" solution has the most interaction with other solutions. The $R-J$ is positive for "need for IT resources," "need to obey higher authorities," and "need to create transparency." Therefore, these solutions are causes in system. $R-J$ is also negative for "need for financial resources" and "need for human resources" solutions. Therefore, these solutions are the effects in the system. In Figure 11, by considering $(R+J)$ on the horizontal axis and $(R-J)$ on the vertical axis, the final position of the solutions in the system is specified. Factors above the axis $(R+J)$ are the causes and factors below the axis $(R+J)$ are the effects.

5. Conclusion

Despite the fact that some governmental organizations and institutions provided their data openly, the interaction of citizens with open government data was not favorable. This

issue can be caused by various factors such as the subject of the data set, the lack of format openness, the lack of up-to-datedness, and the lack of use of visual tools. Therefore, it is necessary to determine the effective criteria for the quality of open government data and increase the willingness of citizens to use the data. Unfortunately, each research has focused only on a specific dimension of open government data, and there is no comprehensive set of criteria. Also, the importance of each dimension and criterion is not considered. More importantly, the weight and importance of each criterion have not been calculated based on citizens' preferences to use different open government data. Also, the solutions and the importance of each of them have not been studied.

In this research, in the first phase, by studying and reviewing the articles, comprehensive and effective criteria were extracted in the quality of open government data and increasing the citizens' willingness from the data. The criteria extracted were included: "data originality," "license openness," "up-to-datedness," "data access," "metadata completeness," "number of data sets," "format openness," "nondiscriminatory," "understandable," "number of categories of data sets," "free," "lack of missing data," "data request ability," "visualization," "feedback," and "data subject matter." In the second phase, the data of 112 governmental organizations and institutions present in the open government data portal were extracted based on each of the stated criteria. In the third phase, in order to identify the organizations and their data sets that were most desired by citizens, the complex network of citizens and government organizations and institutions was analyzed. In order to identify organizations and government institutions with a high degree centrality, input links from citizens to organizations that represent the indegree of each organization were calculated. This information was available based on the number of visits to each organization's data set. In the fourth phase, data mining techniques including regression model were used to identify the data characteristics that were most desired by citizens. The output of the model was a coefficient that determined the positive or negative impact of each criterion as well as the weight and importance of that criterion. According to the results, the criterion of "society subject" with a coefficient of 72.564 and a positive sign had the greatest impact on increasing the number of citizens' visits to open government data. After that, the criterion of "format openness" with a coefficient of 52.682 and a positive sign has the second rank in increasing the number of visits. Criteria for "metadata completeness," "number of data sets," "understandable," "data originality," "free," "lack of missing data," "nondiscriminatory," and "license openness" being a positive sign gained the next ranks in increasing the number of citizens visiting open government data. The "number of categories of data sets" coefficient had a negative sign, meaning that citizens were more inclined to have a data set that focused more on a particular subject. Also, the "farming subject" with a negative sign and a coefficient of 160.413 had the most negative impact on the number of citizens' visits to open government data. Most of the challenges faced by higher ranking criteria such as format openness, small data sets, lack of data updates, etc., were that organizations were reluctant to present their data openly. Therefore, in the fifth phase, it was stated to provide solutions and incentives to increase the willingness of organizations to present their data openly. Based on the study and review of articles, six solutions include "need to create transparency," "need for IT resources," "need to obey higher authorities," "need for human resources," "need for financial resources," and "reducing the sensitivity of the organization's work operations" were extracted. Also, the cause-and-effect relationships between solutions were identified using the DEMATEL technique. Based on the results, the "need to create transparency" solution has the greatest impact on other solutions. The "need for financial resources" solution is most affected by other factors. Also, the "need for financial resources" solution has the most interaction with other solutions.

Extracting comprehensive and effective criteria in improving the quality of open government data and increasing citizens' willingness to use data, calculating the weight and importance of each criterion by analyzing the complex network of citizens and organizations, as well as providing solutions, can help managers make proper decisions and manage complex systems of citizens, government organizations, and institutions providing open government data in order to increase citizens' willingness to use the data and reap the benefits of open government data.

Data Availability

Data are available in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. Janssen, "Open government data and the right to information: opportunities and obstacles," *Journal of Community Informatics*, vol. 8, p. 2, 2012.

[2] M. Solar1, F. Daniels, and R. Lopez, "Automatic generation of roadmaps for open data," in *Electronic Government and Electronic Participation: Joint Proceedings of Ongoing Research, Posters, Workshop and Projects of IFIP EGOV 2014 and EPart*, vol. 21, p. 95, IOS Press, 2014.

[3] I. Susha, A. Grönlund, and M. Janssen, "Driving factors of service innovation using open government data: an exploratory study of entrepreneurs in two countries," *Information Polity*, vol. 20, no. 1, pp. 19–34, 2015.

[4] A. Halonen, "Being open about data," *Analysis of the UK Open Data Policies and Applicability of Open Data*, Finnish Institute in London, London, 2012.

[5] H. Yu and D. G. Robinson, "The new ambiguity of open government," *UCLA L. Rev. Discourse*, vol. 59, p. 178, 2011.

[6] A. Nikiforova and K. McBride, "Open government data portal usability: a user-centred usability analysis of 41 open government data portals," *Telematics and Informatics*, vol. 58, p. 101539, 2021.

[7] H. Zhang and J. Xiao, "Quality assessment framework for open government data: meta-synthesis of qualitative research,

2009-2019," *The Electronic Library*, vol. 38, no. 2, p. 1437, 2020.

[8] S. De Juana-Espinosa and S. ují-Mora, "Open government data portals in the European Union: a dataset from 2015 to 2017," *Data in Brief*, vol. 29, p. 105156, 2020.

[9] L. Zheng, W.-M. Kwok, V. Aquaro, X. Qi, and W. Lyu, "Evaluating global open government data: methods and status," *In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, pp. 381–391, 2020.

[10] K. Y. Dahbi, H. Lamharhar, and D. Chiadmi, "Toward an evaluation model for open government data portals," in *International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*, pp. 502–511, Springer, Cham, 2018.

[11] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: definition and application to Open Government Data," *Government Information Quarterly*, vol. 33, no. 2, pp. 325–337, 2016.

[12] S. S. Dawes, L. Vidiasova, and O. Parkhimovich, "Planning and designing open government data programs: an ecosystem approach," *Government Information Quarterly*, vol. 33, no. 1, pp. 15–27, 2016.

[13] G. Misuraca and G. Viscusi, "Is open data enough? E-governance challenges for open government," *International Journal of Electronic Government Research*, vol. 10, no. 1, pp. 18–34, 2014.

[14] T. M. Harrison, T. A. Pardo, and M. Cook, "Creating open government ecosystems: a research and development agenda," *Future Internet*, vol. 4, no. 4, pp. 900–928, 2012.

[15] E. Estrada, *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, United Kingdom, 2012.

[16] N. Veljković, S. Bogdanović-Dinić, and L. Stoimenov, "Benchmarking open government: an open data perspective," *Government Information Quarterly*, vol. 31, no. 2, pp. 278–290, 2014.

[17] R. Huang, C. Wang, X. Zhang, D. Wu, and Q. Xie, "Design, develop and evaluate an open government data platform: a user-centered approach," *The Electronic Library*, vol. 37, no. 3, p. 287, 2019.

[18] J. Attard, F. Orlandi, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.

[19] S. Saxena, "Open government data (OGD) in six Middle East countries: an evaluation of the national open data portals," *Digital Policy, Regulation and Governance*, vol. 20, no. 4, p. 90014, 2018.

[20] G. Fu, *Modeling Water Availability and its Response to Climatic Change for the Spokane River Watershed*, Washington State University, United Status, 2005.

[21] Y. Zhenbin, A. Kankanhalli, S. Ha, and G. K. , "What drives public agencies to participate in open government data initiatives? An innovation resource perspective," *Information & Management*, vol. 57, no. 3, p. 103179, 2020.

[22] I. Mergel, "Opening government: designing open innovation processes to collaborate with external problem solvers," *Social Science Computer Review*, vol. 33, no. 5, pp. 599–612, 2015.

[23] D. Tapscott, A. D. Williams, and D. Herman, "Government 2.0: transforming government and governance for the twenty-first century," *New Paradigm*, vol. 1, p. l5, 2008.

[24] S. P. Taylor, "Innovation in the public sector: dimensions, processes, barriers and developing a fostering framework," *International Journal of Research Science & Management*, vol. 5, no. 1, pp. 28–37, 2018.

[25] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: the internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol. 5, no. 3, pp. 60–70, 2016.

[26] D. Zheng, J. Chen, L. Huang, and C. Zhang, "E-government adoption in public administration organizations: integrating institutional theory perspective and resource-based view," *European Journal of Information Systems*, vol. 22, no. 2, pp. 221–234, 2013.

[27] H. Yu and D. G. Robinson, "The new ambiguity of open government," *UCLA L. Rev. Discourse*, vol. 59, p. 178, 2011.

[28] A. Appari and M. E. Johnson, "Information security and privacy in healthcare: current state of research," *International Journal of Internet and Enterprise Management*, vol. 6, no. 4, pp. 279–314, 2010.

[29] S. E. Goodman and R. Ramer, "Global sourcing of IT services and information security: prudence before playing," *Communications of the Association for Information Systems*, vol. 20, p. 50, 2007.

[30] C. Coglianese, "The transparency president? The Obama administration and open government," *Governance*, vol. 22, pp. 529–544, 2009.

Hindawi
Complexity
Volume 2022, Article ID 8062969, 19 pages
https://doi.org/10.1155/2022/8062969

WILEY | Hindawi

Research Article

Ontology of Mathematical Modeling Based on Interval Data

Mykola Dyvak,[1] **Andriy Melnyk**,[1] **Artur Rot**,[2] **Marcin Hernes**,[2] **and Andriy Pukas**[1]

[1]*Department of Computer Science, West Ukrainian National University, 11 Lvivs'ka Str., Ternopil 46000, Ukraine*
[2]*Faculty of Management, Wroclaw University of Economics and Business, Komandorska 118/120, Wroclaw 53-345, Poland*

Correspondence should be addressed to Andriy Melnyk; melnyk.andriy@gmail.com

Received 8 February 2022; Revised 9 April 2022; Accepted 13 June 2022; Published 19 July 2022

Academic Editor: Andrea Murari

An ontological approach as a tool for managing the processes of constructing mathematical models based on interval data and further use of these models for solving applied problems is proposed in this article. Mathematical models built using interval data analysis are quite effective in many applications, as they have "guaranteed" predictive properties, which are determined by the accuracy of experimental data. However, the application of mathematical modeling methods is complicated by the lack of software tools for the implementation of procedures for constructing this type of mathematical models, creating an ontological model that operates by the categories of the subject area of mathematical modeling, regardless of the modeling object proposed in this article. This approach has made it possible to generate tools for mathematical modeling of various objects based on the interval data analysis for any software development environment selected by the user. The technology of creating the software on the basis of the developed ontological superstructure for mathematical modeling using the interval data for different objects, as well as various forms of user interface implementation, is presented in this article. A number of schemes, which illustrate the technology of using the ontological approach of mathematical modeling based on interval data, are presented, and the features of its interpretation when solving environmental monitoring problems are described.

1. Introduction

Mathematical modeling is one of the main tools that allows describing the object in a simple form, exploring it, and predicting behavior. Mathematical modeling is understood as the process of building a model and its application to certain applied problems [1–4].

Mathematical modeling processes consist of a large number of procedures, which are mainly implemented in the relevant tools, that is, in the form of certain software systems [3, 4].

Examples of these software environments are Matlab, GNU Octave, Scilab, and SageMath. These tools are multipurpose and well developed. However, practitioners often need to use more specialized tools for building mathematical models, as well as to adapt existing tools to nonstandard conditions that are absent in the noted environments. In this case, there are difficulties in using and interpreting such tools because the simulation procedures are hidden from the researcher, and this makes it difficult to use them by making appropriate software changes [4–8].

In this case, the most appropriate solution is to create an ontological description of certain methods of mathematical modeling. It describes in detail the components of a model building process and its application. Then this ontological description is used to generate appropriate software. This approach, on the one hand, will allow the integration of the created software in various applied systems and, on the other hand, will make changes to existing software [4, 9–12].

The availability of ontological descriptions of modeling processes based on certain methods makes it possible to unify the software used for a wide range of tasks. It enables, based on experience, a repository of mathematical model creation that can be used to model a wide range of mathematically similar properties [13–23].

The positive effect of this approach will be a significant simplification of the process of creating tools for both the modeling processes organization and their application to applied problems.

One of the directions of mathematical modeling is the inductive approach, which is based on a self-organized process

of the evolutionary transition from primary data to explicit mathematical models that reflect the patterns of functioning of simulated objects and systems, which are implicit in existing experimental research and statistical data [24–27].

An important feature of the inductive approach implementation is the nature of the uncertainty in information data sets (probabilistic, interval, fuzzy), as this approach is based on methods of data analysis. In a number of works [28–30], the ontological approach for the construction of the mathematical models within the framework of the inductive approach is based on a group of methods of data handling (GMDH). Within the framework of the proposed approach, the key parameters for the main components of the modeling process are identified, which determine the possibility of generalization and expediency of constructing multifunctional software modules in the development of computer inductive modeling tools based on GMDH [26, 31, 32]. Since the mentioned approach has a complex structure, which is interpreted using Protege [33–36] and does not contain applied software-interpreted solutions, its practical use in other approaches to mathematical modeling is not advisable. The use of such an approach is time-consuming to formalize the subject area and, due to the complexity of its presentation within the Protege system, will not contribute to support among the developers of the appropriate applied software solutions [19, 37, 38].

Another direction in mathematical modeling according to the inductive approach is presented by the methods of mathematical modeling based on interval data [39–43]. The multiple estimates of the parameters of the "input-output" model, built on the results of an experiment in which the output variables are obtained in interval form, are the peculiarities of these methods [44, 45].

As a result of the application of the methods of interval analysis, instead of one "input-output" model, there is a corridor (set) of equivalent interval models of the system. The properties of the obtained models depend on the chosen method of sets of parameter estimation. Preferably, sets of parameter estimates can be presented in the forms of a polyhedron, a multidimensional ellipsoid, or a rectangular parallelepiped that specifies the intervals of parameter values [46, 47].

Given that the methods of systems modeling, based on the analysis of interval data, require minimal information about the research system, their applications significantly expand the class of research systems [48].

However, these methods are limited for use by both researchers and users-practitioners due to the lack of developed ontological description for this area of mathematical modeling, which would make it possible to expand the scope of application of the existing interval models for a particular subject area and to develop new models. An example, in this case, is the field of building mathematical models for medicine [41] or environmental monitoring, in particular, the description of mathematical models based on interval data for the processes of air pollution by harmful emissions from vehicles [46–48]. The long-term experience of the authors of this work in creating and applying this type of model has shown that in the case of changes in the state of the environment, or conditions for obtaining interval data, most built interval models lose accuracy or become inadequate. The application of the ontological superstructure to the process of development and use of models significantly expands the possibilities of modeling the characteristics of these systems and increases the accuracy of the model in specific cases. Simply put, an ontological model as an "add-on" can use the "switch" functions to select the best model from the repository, depending on changes in the simulation environment.

The need for automated, systematic, and reusable mathematical models as an environment for knowledge obtaining, accumulating, and reusing is fully justified in the context of a large amount of information about knowledge, which is generated and stored.

Therefore, the aim of this article is to create an ontology of mathematical modeling based on interval data, which would expand the possibilities for researchers dealing with the objects of different nature, data on which were obtained in interval form, as well as for practitioners who can use it for modeling processes in medicine, environmental monitoring, etc.

2. Statement of the Problem of Mathematical Modeling Based on Interval Data

The problem of object modeling based on interval data is considered in [42, 47]. The authors of the interval approach declare that it has a number of advantages over the stochastic (probabilistic) approach. Among them is the absence of a requirement to research the statistical characteristics of the simulated object. As it is known, this reduces the number of experiments (data sampling). Therefore, the interval approach is more useful for researching the object properties in conditions of limited data sampling. A declarative approach to presenting knowledge about object modeling methods based on interval data analysis makes it possible to develop tools for using this approach by both researchers and practitioners. To develop a declarative ontology, the basic concepts of this approach should be considered.

First, the basic concept refers to a method of presenting data in the form of intervals of possible values of the simulated characteristic:

$$\left[z^{-}_{i,j,h,k}; z^{+}_{i,j,h,k}\right], i = 0, \ldots, I,\ j = 0, \ldots, J, h = 0, \ldots, H,\ k = 0, \ldots, K, \tag{1}$$

where $z^{-}_{i,j,h,k}, z^{+}_{i,j,h,k}$ are, accordingly, the lower and upper bounds of intervals of possible values of the output characteristic at a point with discretely given spatial coordinates $i = 0, \ldots, I$, $j = 0, \ldots, J$, $h = 0, \ldots, H$ (for objects with distributed parameters) and time discrete $k = 0, \ldots, K$ (for dynamic objects, for example, a dynamic of air pollution from vehicles in discrete time).

Note that in the measuring experiment, the lower and upper bounds can be set by the relative error of the measuring device: $z^{-}_{i,j,h,k} = z_{i,j,h,k} - z_{i,j,h,k} \cdot \varepsilon$ and $z^{+}_{i,j,h,k} = z_{i,j,h,k} + z_{i,j,h,k} \cdot \varepsilon$, where $z_{i,j,h,k}$ is the measured value of characteristic; ε is a relative error of measuring.

Representation of experimental data in interval form (1) is reasonable in cases: when the measurement error significantly exceeds the methodological errors and modeling errors, intervals

(1) set the tolerance bounds of deviations of the simulated characteristic of the object from the nominal, under conditions of known maximum values of errors in the experiment.

Next, it is necessary to determine the mathematical object to represent the object model. In this case, it is limited to a discrete linear model in general

$$v_{i,j,h,k} = \overrightarrow{f}^{T}\left(v_{i-d,j-d,h-d,k-d}, v_{i-d+1,j-d,h-d,k-d}, \ldots, v_{i-1,j-1,h-1,k-1}, \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}\right) \cdot \overrightarrow{g}, \quad i = d, \ldots, I, j = d, \ldots, J, h = d, \ldots, H, k = d, \ldots, K, \tag{2}$$

where $\overrightarrow{f}^{T}(\bullet)$ is a vector of basic functions, in general nonlinear, with the help of which the values of the simulated characteristic of the object are transformed, as well as the input variables at discrete space points and for a certain time are discrete.

As a result of performing the procedure of structural identification, a discrete model is determined, in particular: the vector of basic functions $\overrightarrow{f}^{T}(\bullet)$; sets and dimension of vectors of input variables (controls) $\overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}$; d is an order of a discrete model, which as is known is equivalent to the order of a differential equation analogous to a discrete model. To implement a discrete model, it is also necessary to specify the initial conditions, i.e., the value of each element in the set $v_{0,0,0,0}, \ldots, v_{d-1,d-1,d-1,d-1}, \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}$ for certain discrete, as a rule, initial one, and set the value of the components in the parameters vector $\overrightarrow{g}$.

If the general form of the discrete model is known, for example, due to physical considerations, it remains to identify the parameters $\overrightarrow{g}$ in a way to ensure maximum agreement of the simulated characteristic of the object with the experimentally obtained values of this characteristic. This task is called the parametric identification task [42].

Let's assume that the vector of estimates $\overrightarrow{\hat{g}}$ of parameters $\overrightarrow{g}$ in the difference operator (2) is obtained on the basis of interval data analysis. Substituting a vector of parameter estimates $\overrightarrow{\hat{g}}$ from difference operator instead of the vector of their true values $\overrightarrow{g}$ in expression (2) together with the specified initial interval values of each element of the set $[\hat{v}_{0,0,0,0}], \ldots, [\hat{v}_{d-1,d-1,d-1,d-1}]$ and given vectors of input variables $\overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}$ an interval estimate of the simulated characteristic $[v_{i,j,h,k}]$ at points with discrete spatial coordinates $i = d, \ldots, I$, $j = d, \ldots, J$, $h = d, \ldots, H$ and on time discrete $k = d, \ldots, K$ can be obtained:

$$\left[\hat{v}_{i,j,h,k}\right] = \left[\hat{v}^{-}_{i,j,h,k}; \hat{v}^{+}_{i,j,h,k}\right] = \overrightarrow{f}^{T}\left(\left[\hat{v}_{i-d,j-d,h-d,k-d}\right], \ldots, \left[\hat{v}_{i-1,j-1,h-1,k-1}\right], \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}\right) \cdot \hat{\overrightarrow{g}}, \quad i = d, \ldots, I, j = d, \ldots, J, h = d, \ldots, H, k = d, \ldots, K. \tag{3}$$

Now, the problem of parametric identification of the interval discrete model (IDM) based on the interval data analysis can be mathematically formulated.

The conditions of matching the experimental data presented in the interval form (1) with the data obtained on the basis of the macromodel in the form of IDM (3) are formulated as follows:

$$\left[\hat{v}^{-}_{i,j,h,k}\hat{v}^{+}_{i,j,h,k}\right] \subset \left[z^{-}_{i,j,h,k}; z^{+}_{i,j,h,k}\right], \forall i = 0, \ldots, I, \quad \forall j = 0, \ldots, J, \forall h = 0, \ldots, H, \forall k = 0, \ldots, K. \tag{4}$$

Conditions (4) provide obtaining the interval estimates of the simulated characteristic of the object within the intervals of possible values of the characteristic obtained experimentally.

Substitute in equation (4) instead of interval estimates $[\hat{v}^{-}_{i,j,h,k}; \hat{v}^{+}_{i,j,h,k}]$ of the simulated characteristic; its interval values are calculated on the basis of IDM (3) together with taking into account the given initial interval values of each element from a set:

$$\left[\hat{v}_{0,0,0,0}\right] \subseteq \left[z_{0,0,0,0}\right], \ldots, \left[\hat{v}_{i-1,j-1,h-1,k-1}\right] \subseteq \left[z_{i-1,j-1,h-1,k-1}\right], \tag{5}$$

and given vectors of input variables $\overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k}$, and receive the following:

$$\begin{cases} \left[\hat{v}^{-}_{0,0,0,0}; \hat{v}^{+}_{0,0,0,0}\right] \subseteq \left[z^{-}_{0,0,0,0}; z^{+}_{0,0,0,0}\right]; \\ \ldots \\ \left[\hat{v}^{-}_{d-1,d-1,d-1,d-1}; \hat{v}^{+}_{d-1,d-1,d-1,d-1}\right] \subseteq \left[z^{-}_{d-1,d-1,d-1,d-1}; z^{+}_{d-1,d-1,d-1,d-1}\right]; \\ z^{-}_{i,j,h,k} \le \overrightarrow{f}^{T}\left(\left[\hat{v}_{i-d,j-d,h-d,k-d}\right], \ldots, \left[\hat{v}_{i-1,j-1,h-1,k-1}\right], \overrightarrow{u}_{0}, \ldots, \overrightarrow{u}_{k}\right) \cdot \hat{\overrightarrow{g}} \le z^{+}_{i,j,h,k}; \\ i = d, \ldots, I; \; j = d, \ldots, J; \; h = d, \ldots, H; \; k = d, \ldots, K. \end{cases} \tag{6}$$

Therefore, an equation (6) is obtained by substituting interval estimates of initial characteristics $[\widehat{v}^{-}_{i,j,h,k}; \widehat{v}^{+}_{i,j,h,k}]$ (given as initial conditions and predicted on the basis of expression (3) in the remaining nodes of the grid) in conditions (4).

As it is known, the obtained system is an interval system of nonlinear algebraic equations (ISNAE). Therefore, the task of identifying the parameters of IDM (3) under conditions (4) is the task of solving ISNAE in the form (6).

It should be noted that ISNAE (6) is formed recurrently. The total number of interval equations is a product of $I \times J \times H \times K$.

Obviously, the greater the number of equations in the interval system, the more difficult it is to find the ISNAE solution.

Given that this problem cannot be solved for a predetermined number of iterations, this type of problem belongs to NP-complete. The only way to solve it is to do a full search or random search. Given the complexity of the task of IDM parametric identification, to find at least one ISNAE solution, random search methods can be used [42].

These computational schemes for the implementation of the method of IDM parametric identification are based on four-step procedures [44].

Step 1. Set the initial conditions in the form (5).

Step 2. Set the initial $\widehat{\vec{g}}$ or randomly generate the current $\widehat{\vec{g}}$ estimate of the vector of the IDM parameters.

Step 3. Calculate the interval estimates of the simulated characteristic $[\widehat{v}_{i,j,h,k}]$ at points with discrete-given spatial coordinates $i = d, \ldots, I$, $j = d, \ldots, J$, $h = d, \ldots, H$ and on time discrete $k = d, \ldots, K$ using a recurrent scheme (4).

Step 4. Check the "quality" $\delta(\widehat{\vec{g}}_{l})$ of the current approximation of the estimate $\widehat{\vec{g}}$ of the vector of IDM parameters [39, 40].

In this step, assume that the "quality" of the approximation will be higher if the predicted corridor is closer, built on the basis of this parameter vector approximation, to the experimental one.

If the calculated value of "quality" $\delta(\widehat{\vec{g}}_{l})$ of the current approximation of the estimate $\widehat{\vec{g}}$ of the vector of IDM parameters at the current iteration is zero ($\delta(\widehat{\vec{g}}_{l}) = 0$), then the procedure is over; otherwise, go to Step 2.

The quality of the approximation will be quantified as the difference between the centers of the most distant predictive and experimental intervals in the case when they do not intersect, and the width of the intersection of the predictive and experimental intervals is the smallest, for the case of their intersection [40].

Formally, these conditions are written as follows:

$$\delta\left(\widehat{\vec{g}}_{l}\right) = \max_{i=d,\ldots,I, j=d,\ldots,J, h=d,\ldots,H, k=d,\ldots,K} \left\{\left|\mathrm{mid}\left(\left[\widehat{v}_{i,j,h,k}\right]\right) - \mathrm{mid}\left(\left[z_{i,j,h,k}\right]\right)\right|\right\},$$
$$\text{if } \left[\widehat{v}_{i,j,h,k}\right] \cap \left[z_{i,j,h,k}\right] = \varnothing \; \exists i = d, \ldots, I \; \exists j = d, \ldots, J \; \exists h = d, \ldots, H, \; \exists k = d, \ldots, K. \tag{7}$$

$$\delta\left(\widehat{\vec{g}}_{l}\right) = \max_{i=d,\ldots,I, j=d,\ldots,J, h=d,\ldots,H, k=d,\ldots,K} \left\{\mathrm{wid}\left(\left[\widehat{v}_{i,j,h,k}\right]\right) - \mathrm{wid}\left(\left[\widehat{v}_{i,j,h,k}\right], \cap \left[z_{i,j,h,k}\right]\right)\right\}$$
$$\text{if } \left[\widehat{v}_{i,j,h,k}\right] \cap \left[z_{i,j,h,k}\right] \neq \varnothing \; \forall i = d, \ldots, I, \; \forall j = d, \ldots, J, \; \forall h = d, \ldots, H, \; \forall k = d, \ldots, K, \tag{8}$$

where mid (•) and wid (•) are operations for determining the center and width of the interval correspondingly.

Therefore, the problem of parametric identification of interval models of the object is formulated in the form of an optimization task:

$$\delta\left(\widehat{\vec{g}}_{l}\right) \longrightarrow \widehat{\vec{g}}_{l}$$

$$\min, \; \widehat{g}_{jl} \in \left[g^{\mathrm{low}}_{jl}; g^{\mathrm{up}}_{jl}\right], \; j = 1, \ldots, m, \; l = 1, \ldots, S, \tag{9}$$

where the value of the objective function $\delta(\widehat{\vec{g}}_{l})$ is calculated by formula (7) or (8).

Let's consider the problem of IDM structural identification in general (3). The complexity of the task of configuring IDM (3) is that not only the parameters are unknown, but the same is with the structure. In this case, to find the IDM parameters, it is necessary to solve the problem of parametric identification and identify the structure–structural identification. Note that both these tasks are very closely related because parametric identification is a structural stage, and to find one solution to the latter, it is necessary to make many attempts to find the vector of IDM parameters. Note that the "success" of the task of finding the vector of IDM parameters directly depends on the success of the process of selecting its structure. After all, if the defined IDM structure is "unsuccessful," then it is impossible to find a solution of the parametric identification task.

Therefore, parametric identification is a stage of structural identification. When the data is given in interval form, this step is to find estimates of the IDM parameters by solving the ISNAE (6) for some known vector of basic functions (structural elements of the IDM).

To solve ISNAE (6), the method of parametric identification based on random search procedures is used. The application of this method involves, instead of ISNAE (6) solving,

the search for some approximation to its solution, which determines the quality of the current IDM structure [47].

Let's use some notations that are necessary to reveal the essence of the task formulation. Denote by λ_s the current IDM structure

$$\lambda_s = \left\{ f_1^s(\bullet) \cdot g_1^s; f_2^s(\bullet) \cdot g_2^s; \ldots; f_{m_s}^s(\bullet) \cdot g_{m_s}^s \right\}, \tag{10}$$

where $\overrightarrow{f}^s = \left\{ f_1{}^s(\bullet); f_2{}^s(\bullet); \ldots; f_{m_s}^s(\bullet) \right\} \subset F$ is a set of structural elements that specify the current s IDM structure.

Next, denote the following symbols: $m_s \in [I_{\min}; I_{\max}]$ is a number of elements in the current structure λ_s; F is the set of all structural elements, $F = \{f_1(\bullet); \ldots; f_l(\bullet); \ldots; f_L(\bullet)\}$, where $|F| = L$ (power of the set F); $\overrightarrow{g}^s = \left\{ g_1^s; g_2^s; \ldots, g_{m_s}^s \right\}$ is a vector of unknown parameter values. Structural identification aims at finding the IDM structure λ_0 in the form of (10) so that the interval discrete model is formed on its basis [48].

$$\begin{aligned} \left[v_{i,j,h,k}(\lambda_0) \right] = {} & \left[f_1{}^0(\bullet) \right] \cdot \widehat{g}_1^0 + \left[f_2{}^0(\bullet) \right] \cdot \widehat{g}_2^0 \\ & + \ldots + \left[f_{m_0}^0(\bullet) \right] \cdot \widehat{g}_{m_0}^0, \end{aligned} \tag{11}$$

The conditions (4) are true, i.e., the interval estimates of the predicted value of the simulated characteristic are included in the intervals of tolerance values of the simulated characteristic on the set of all discrete.

The quality of the current IDM structure is estimated on the basis of the value of the indicator $\delta(\lambda_s)$, which quantifies the proximity of the current structure to a satisfactory level in terms of providing conditions (4). Afterward, $\delta(\lambda_s)$ will be called the objective function of the optimization task of the structural identification of a mathematical model with guaranteed prognostic properties.

The value of the quality indicator $\delta(\lambda_s)$ for the current IDM structure λ_s is calculated using modified expressions (7) and (8):

$$\delta(\lambda_s) = \max_{i=d,\ldots,I, j=d,\ldots,J, h=d,\ldots,H, k=d,\ldots,K} \left| \left\{ \operatorname{mid}\left(\overrightarrow{f}_s^T \begin{pmatrix} \left[\widehat{v}_{i-d,j-d,h-d,k-d} \right], \\ \ldots, \left[\widehat{v}_{i-1,j-1,h-1,k-1} \right], \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k} \end{pmatrix} \cdot \widehat{\overrightarrow{g}}^s \right) - \operatorname{mid}\left(\left[z_{i,j,h,k} \right] \right) \right| \right\}, \tag{12}$$

$$\delta(\lambda_s) = \max_{i=d,\ldots,I, j=d,\ldots,J, h=d,\ldots,H, k=d,\ldots,K} \left\{ \begin{array}{c} \operatorname{wid}\left(\overrightarrow{f}_s^T \left(\left[\widehat{v}_{i-d,j-d,h-d,k-d} \right], \ldots, \left[\widehat{v}_{i-1,j-1,h-1,k-1} \right], \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k} \right) \cdot \widehat{\overrightarrow{g}} \right) - \\ - \operatorname{wid}\left(\left(\overrightarrow{f}_s^T \left(\left[\widehat{v}_{i-d,j-d,h-d,k-d} \right], \ldots, \left[\widehat{v}_{i-1,j-1,h-1,k-1} \right], \overrightarrow{u}_{i,j,h,0}, \ldots, \overrightarrow{u}_{i,j,h,k} \right) \cdot \widehat{\overrightarrow{g}} \right) \cap \left[z_{i,j,h,k} \right] \right) \end{array} \right\}, \tag{13}$$

$$\text{if } \left[\widehat{v}_{i,j,h,k} \right] \cap \left[z_{i,j,h,k} \right] = \varnothing, \exists i = d \ldots, I, \; j = d, \ldots, J, \exists h = d, \ldots, H, \exists k = d, \ldots, K$$

$$\text{if } \left[\widehat{v}_{i,j,h,k} \right] \cap \left[z_{i,j,h,k} \right] \neq \varnothing, \forall i = d, \ldots, I, \forall j = d, \ldots, J, \forall h = d, \ldots, H, \forall k = d, \ldots, K,$$

where $\operatorname{mid}(\bullet), \operatorname{wid}(\bullet)$ are operations from interval analysis determining the center and width of the intervals, accordingly.

Expression (12) describes the "proximity" of the current structure to a satisfactory level in the initial iterations, and expression (13) in the case of $\delta(\lambda_s) = 0$ ensures the fulfillment of conditions (4).

The task of IDM structural identification is written formally in the form of the task of finding the minimum of the objective function $\delta(\lambda_s)$:

$$\delta(\lambda_s) \xrightarrow{\lambda_s} = \left(f_1^s \overrightarrow{V} \right) \cdot g_{l1}^s, f_2^s \left(\overrightarrow{V} \right) \cdot g_{l2}^s, \ldots, f_{m_s}^s \left(\overrightarrow{V} \right) \cdot g_{lm_s}^s \min, \tag{14}$$

$$\begin{aligned} & m_s \in [I_{\min}; I_{\max}], \; f_1^s(\overrightarrow{V}), f_2^s(\overrightarrow{V}), \ldots, f_{m_s}^s(\overrightarrow{V}) \in F \\ & \widehat{\overrightarrow{g}}^s{}_{jl} \in \left[g_{jl}^{\text{low}}; g_{jl}^{\text{up}} \right], j = 1, \ldots, m, \; l = 1, \ldots, S, \end{aligned} \tag{15}$$

where $m_s \in [I_{\min}; I_{\max}]$ is a number of elements in s interval model structure; $F = \left\{ f_1(\overrightarrow{V}), f_2(\overrightarrow{V}), \ldots, f_m(\overrightarrow{V}) \right\}$ is a set of potential structure elements in a model.

From expressions (12) and (13), it is seen that for the calculated value of the objective function $\delta(\lambda_s)$ for the IDM structure λ_s, the inequality $\delta(\lambda_s) \geq 0$ will be satisfied under

any conditions. Therefore, the objective function $\delta(\lambda_s)$ has a global minimum only at those points for which the equality $\delta(\lambda_s) = 0$ holds. Based on the theory of multiplicity of models [40], it can be stated that in the search space for solutions to the IDM structural identification task, the function $\delta(\lambda_s)$ has many global minima.

The smaller the value of $\delta(\lambda_s)$, the "better" the current IDM structure. If $\delta(\lambda_s) = 0$, then the current IDM structure makes it possible to build an adequate model for which the interval estimates of the predicted characteristic belong to the intervals of possible values of the modeled characteristic.

As it can be seen, the IDM structural identification is reduced to the multiple repeating of the parametric identification problem-solving. Therefore, it is important to develop methods of structural identification, which would reduce the number of iterations of the method for finding an adequate structure of the mathematical model and, accordingly, would reduce the required number of repeating the parametric identification problem-solving.

3. Methods of Mathematical Modeling Based on Interval Data

The previous section presents a four-step procedure for solving the problem of parametric identification. However, to date, the most effective methods for solving this optimization problem are methods based on behavioral models of artificial bee colonies (ABC) [49]. The substantiation of this fact is given in [40, 44].

To build a method of parametric identification, the principles of behavioral models of the bee colony are used.

Initialization phase. Vectors that determine the possible minimum points of the objective function (9) are the vectors of parameter estimates and are denoted by $\widehat{\vec{g}}_l$. In the context of the behavioral model of the bee colony, this means that each vector of the nectar source coordinates corresponds to one l bee that investigates it. Let's set the number of the entire population to be equal to the value S and set the bounds of the parameter estimates

$$\widehat{g}_{jl} \in \left[g_{jl}^{\text{low}}; g_{jl}^{\text{up}}\right], j = 1, \ldots, m, l = 1, \ldots, S. \tag{16}$$

In this phase the following formula is used:

$$\widehat{g}_{jl} = g_{jl}^{\text{low}} + \text{rand}(0,1) * \left(g_{jl}^{\text{up}} - g_{jl}^{\text{low}}\right), \quad j = 1, \ldots, m, l = 1, \ldots, S, \tag{17}$$

where $g_{jl}^{\text{low}}, g_{jl}^{\text{up}}$ are lower and upper bounds of parameter values at the initialization phase.

Notice that in this phase, all the parameters of the algorithm are also configured [42].

The phase of worker bees. In the context of the optimization task, the phase of worker bees means the search for new estimates of solutions (16) with smaller values of the objective function. To calculate the possible points of the local minimum of the objective function, the following formulas are used:

$$\widehat{g}_{jl}^{mcn} = \widehat{g}_{jl} + \Phi_{jl} * \left(\widehat{g}_{jl} - \widehat{g}_{jp}\right), j = 1, \ldots, m, p \neq l = 1, \ldots, S. \tag{18}$$

After calculating the coordinates of the possible points of the minimum $\widehat{\vec{g}}_l^{mcn}$ a pairwise comparison of the existing and current values of the parameter estimates (16) is performed using the objective function:

$$\widehat{\vec{g}}_l = \left\{\widehat{\vec{g}}_l, \text{ if } \delta\left(\widehat{\vec{g}}_l\right) \leq \delta\left(\widehat{\vec{g}}_l^{mcn}\right)\right\} \text{ or } \widehat{\vec{g}}_l = \left\{\widehat{\vec{g}}_l^{mcn}, \text{ if } \delta\left(\widehat{\vec{g}}_l\right) > \delta\left(\widehat{\vec{g}}_l^{mcn}\right)\right\}. \tag{19}$$

The phase of researchers bees. In the context of the optimization task, at this stage, the most probable points (vectors of parameter values) were determined, around which it is necessary to conduct a detailed study of the objective function. It is these points that claim to provide local minima of the objective function. For these purposes, the probabilistic approach is used, namely, the probabilities of the expediency of research are calculated, and each specific point is given by the vector of parameter values from the previously found ones. The expression for calculating the specified probability is as follows:

$$P_l = \frac{1 - \delta\left(\widehat{\vec{g}}_l\right)}{\sum_{l=1}^{S}\left(1 - \delta\left(\widehat{\vec{g}}_l\right)\right)}. \tag{20}$$

It should be noted that in the case of a significant deviation between the values of the objective function $\delta(\widehat{\vec{g}}_l)$, calculated for different points (vectors of parameter values), it is necessary to rewrite formula (20), taking into account the normalization of the values of this function. In this case, the formula takes the following form:

$$P_l = \frac{1}{\delta\left(\widehat{\vec{g}}_l\right)\sum_{l=1}^{S} 1/\delta\left(\widehat{\vec{g}}_l\right)}. \tag{21}$$

Based on the calculated probabilities, the number of points for researching the possible local minima of the objective function from task (9) is determined. However, given that the value of m_l in this formula must be an integer because it determines the number of points in the neighborhood of the studied point to find the minimum of the objective function, the formula will be rewritten as follows:

$$m_l = ToInt\left(P_l \cdot S\right), l = 1, \ldots, S, m_{l=1} = 0, \tag{22}$$

where $ToInt(\cdot)$ is the operator of selection of the integer part from number.

Then the procedure is repeated to determine the points where the lowest value of the objective function is achieved.

To avoid focusing on the local minima of the objective function, the phase of scout bees is used.

The phase of scout bees. This is the phase where new solutions to the optimization problem are randomly calculated again. To do this, formula (18) is used. As mentioned above, in the context of the behavioral model of the bee colony, this means the exhausting of current nectar sources.

Each iteration of calculations involves obtaining a new number of points in addition to the current ones. At the end of each iteration, it has $2S$ points - applicants for research. Therefore, at the end of the iteration, a group selection of points is performed with the smallest value of the objective function $\delta(\vec{g}_l)$, so that their number is equal to the value of S. This procedure is called group selection. The procedure ends under the condition $\delta(\vec{g}_l) = 0$.

Given the analogy between the mathematical formulation of problems of parametric and structural identification of object models, the main phases of the method for structural identification of models of dynamic objects based on the behavioral models of the bee colony are considered.

Initialization phase. In this phase, the main parameters of the method are set: LIMIT; S; $[I_{\min}; I_{\max}]$; $mcn = 0$ is a current iteration number; MCN is the total number of iterations and the set of structural elements is F, and also the initial set Λ_0 (with power S) of the structures λ_s from the set of structural elements F is randomly formed.

In this case, the structural elements will look different than in Table 1. The results of coding the structural elements for the case of developing a model of the characteristics of a dynamic object are shown in Table 1.

Next, to form structures, consider a set of operators. Note that their names and purposes are stored by analogy with the existing method of structural identification built on the ABC.

The phase of worker bees. In the phase of worker bees, the operator $P(\Lambda_{mcn}, F)$, which transforms the structure of the interval model in the form (10), is used. On the current iteration of implementation of the method of structural identification, this operator $P(\Lambda_{mcn}, F)$ forms, on the basis of each of the current structures λ_s of the mathematical model, one "new" structure λ_s', which is close to the current one. Therefore, the operator $P(\Lambda_{mcn}, F)$ converts the set Λ_{mcn} of the current structures λ_s generated on the mcn iteration into the set Λ_{mcn}' structures λ_s' by randomly selecting and replacing part of the elements of the current structure λ_s and also replaces on selected elements from the set $F = \{f_1(\vec{V}), f_2(\vec{V}), \ldots, f_m(\vec{V})\}$. In this case, the set of n_s elements of the current structure that need to be replaced is inversely proportional to the value of the objective function $\delta(\lambda_s)$, which is calculated by formulas (12) or (13).

Table 1: Coding of structural elements for the model of dynamic objects.

No	Structural elements
1	$f_1(\vec{V})$
2	$f_2(\vec{V})$
...	...
m	$f_m(\vec{V})$

Next, in this phase, using the operator $D_1(\lambda_s, \lambda_s')$, pairwise selection is performed to choose the best structure from the two ones: the current and the generated one. To do this, the following formula is used:

$$D_1(\lambda_s, t\lambda_s'): \lambda_s^1 = \begin{cases} \lambda_s, if\ \delta(\lambda_s) \le \delta(\lambda_s'); \\ \lambda', if\ \delta(\lambda_s) > \delta(\lambda_s'). \end{cases} \quad (23)$$

The operator $D_1(\lambda_s, \lambda_s')$ implements the process of synthesis of the set of "best" structures Λ_{mcn}^1 from the current sets Λ_{mcn}, Λ_{mcn}^1. Thus, a set of structures of the first series of formation $\lambda_s^1 \in \Lambda_{mcn}^1, s = 1 \cdots S$ is obtained.

The phase of researchers bees. As already mentioned, in this phase, the number of R_s structures is determined. It will be generated on the basis of each λ_s^1 structure from the set Λ_{mcn}^1. This indicator R_s is calculated by formulas:

$$P_s\left(\lambda_s^1\right) = \frac{1 - \delta\left(\lambda_s^1\right)}{\sum_{s=1}^{S}\left(1 - \delta\left(\lambda_s^1\right)\right)}, s = 1 \ldots S,$$
$$R_s = ToInt\left(P_s\left(\lambda_s^1\right) \cdot S\right), s = 1 \ldots S. \quad (24)$$

Next, in this phase also, the operator $P_\delta(\Lambda_{mcn}, F)$ is used, which converts the current structure into a certain number of R_s structures. In this case, the total number of structures distributed between the current structures is equal to S. Thus, $P_\delta(\Lambda_{mcn}, F)$ means the transformation of each structure λ_s^1 from the set of structures $\lambda_s^1 \in \Lambda_{mcn}^1$ of the first series of formations, generated by iterating the algorithm $mcn = 0$, to the set of structures λ_s', $s = 1 \cdots S$. Replacement of elements in each current structure (or some structures) is carried out randomly on the basis of the calculated value of the number n_s elements in the current structure and is inversely proportional to the value of the objective function $\delta(\lambda_s)$. This substitution is also performed on randomly selected elements from the set $F = \{f_1(\vec{V}), f_2(\vec{V}), \ldots, f_m(\vec{V})\}$.

Also, in this phase, group selection $D_2(\lambda_s^1, \lambda_s')$ of the "best" structure from the current λ_s^1 is performed and the set $\lambda_s' = \{\lambda_1 \ldots \lambda_r \ldots \lambda_{Rs}\}$ is formed in its neighborhood by the values of the objective function. This selection operator, as distinct from the pair selection operator $D_1(\lambda_s, \lambda_s')$, has the following form:

$$D_1\left(\lambda_s^1, \Lambda_s'\right): \lambda_s^2 = \begin{cases} \lambda_s^1, \text{if } (R_s = 0); \\ \lambda_s^1, \text{if}\left(\left(\delta\left(\lambda_s^1\right) \le \delta\left(\lambda_r\right)\right) \wedge \left(R_s \ne 0\right)\right), \forall \lambda_r \in \Lambda_s', r = 1 \ldots R_s; \\ \lambda_r^s, \text{if}\left(\left(\delta\left(\lambda_s^1\right) > \delta\left(\lambda_r\right)\right) \wedge \left(R_s \ne 0\right)\right), \exists \lambda_r \in \Lambda_s', r = 1 \ldots R_s. \end{cases} . \tag{25}$$

Operator (25) implements the process of synthesis of the set of "best" IDM structures Λ_{mcn}^2 from the current sets Λ_{mcn}^1 and $\Lambda_{mcn}^{"} = \{\Lambda_1' \cup \Lambda_2' \ldots \Lambda_s' \ldots \cup \Lambda_S'\}, s = 1 \ldots S$ in the method of ranking all structures by the values of the objective function (12) or (13) with subsequent selection of $s = 1 \ldots S$ structures λ_s^2 by the highest value of the objective function of the optimization task (12), (13). Thus, the set of structures of interval models of the second series of formation Λ_{mcn}^2 is obtained.

Exit from the local minima of the objective function in task (12), (13) is carried out in the phase of scout bees.

The phase of scout bees. To do this, for each current structure λ_s' enter the Limit_S counter, which is incremented by "1" each time. If during pairwise or group selection, the current structure is not "updated," and reset, otherwise. Comparing the value of this counter with some LIMIT constant given in the initialization phase makes it possible to decide whether the current structure has exhausted itself. If the counter $Limit_S$ reaches the value LIMIT, it is no longer appropriate to modify this current structure. This means that the function (14) is in the local minimum. Then, use the operator $P_N(\text{F}, I_{\min}, I_{\max})$, which randomly generates a "new" structure λ_s^2 from the set F of all structural elements randomly, as in the initialization phase, only for one structure. Therefore, such structures will be only a few percent of the S value (of all worker bees).

The procedure is completed under the condition that for some structure in the task of parametric identification, the condition is true: $\delta(\vec{g}_l) = 0$.

The main problem with using these methods is the lack of declarative ontological description, which does not allow developing software environments as a tool. On the other hand, as it is seen from the description of the structural identification task, the main problem for its solving is the formation of a set of potential structural elements of the model $F = \left\{f_1(\vec{V}), f_2(\vec{V}), \ldots, f_m(\vec{V})\right.$ of difference (discrete) equation, which represents a mathematical model of the object. This problem can be solved by the ontological description of the subject area of modeling, i.e., operational ontology. Therefore, solutions to these problems will reduce the complexity of the modeling procedure and adequate models with guaranteed prognostic properties will be obtained.

4. Features of the Ontological Approach Implementation

The need for automated, systematic, reusable mathematical models as an environment for obtaining, accumulating, and reusing knowledge is fully justified in the context of a large amount of information about the process and production of previously generated and stored knowledge. To achieve these goals, as well as in order to expand the possibilities of the researchers of objects of different nature in cases when the data is presented in interval form, it is necessary to build an ontology of mathematical modeling based on interval data.

In the proposed ontological approach to represent the concepts, methods, and tools of mathematical modeling based on interval data, namely the declarative and procedural parts, mathematical knowledge is separated. The declarative part consists of the information needed to build the model, the information obtained from the model, and the corresponding mathematical expressions that represent the model. The procedural part consists of detailed parts of the model, appropriate methods and algorithms for their implementation, and procedures for initializing variables and their interpretations. Among the tools used to build and apply the ontology, Protege and OntoStudio are the most commonly used [33, 34, 50]. Due to their reliability, widespread use, scalability, and extensibility, these tools can also be used in the process of building appropriate ontological models to represent and manage the knowledge they accumulate in the process of mathematical modeling [35, 51, 52]. However, these tools are difficult to integrate into software and hardware systems, which, in particular, are often used in medicine, where the speed and quality of managing decisions are a priority. Therefore, for building an ontology in this paper, the following tools are used:

(i) tools of modern relational databases for information storage [53–55];

(ii) algebra of tuples for the formalized presentation of knowledge and its subsequent program interpretation regardless of the selected software platforms for its implementation, as well as for implementation of effective methods of managing accumulated knowledge [56–59];

(iii) Python and Java as programming languages for the appropriate interpretation of the proposed methods and tools [60–63].

In Figure 1 a general scheme of the relationship between the declarative and procedural parts of the knowledge that is accumulated in the process of mathematical modeling based on interval data within the proposed ontological approach is shown.

The declarative part of the ontological approach consists of an ontology of formalized mathematical models (declarative ontology), which contains model definitions and an information repository. The ontology of using mathematical models (operational ontology) contains design data, operating conditions, and equipment parameters for the use of models. Model ontology consists of a model class that has both attributes and instances.

A class of equations denotes model equations (integral equations, algebraic equations or functions), model

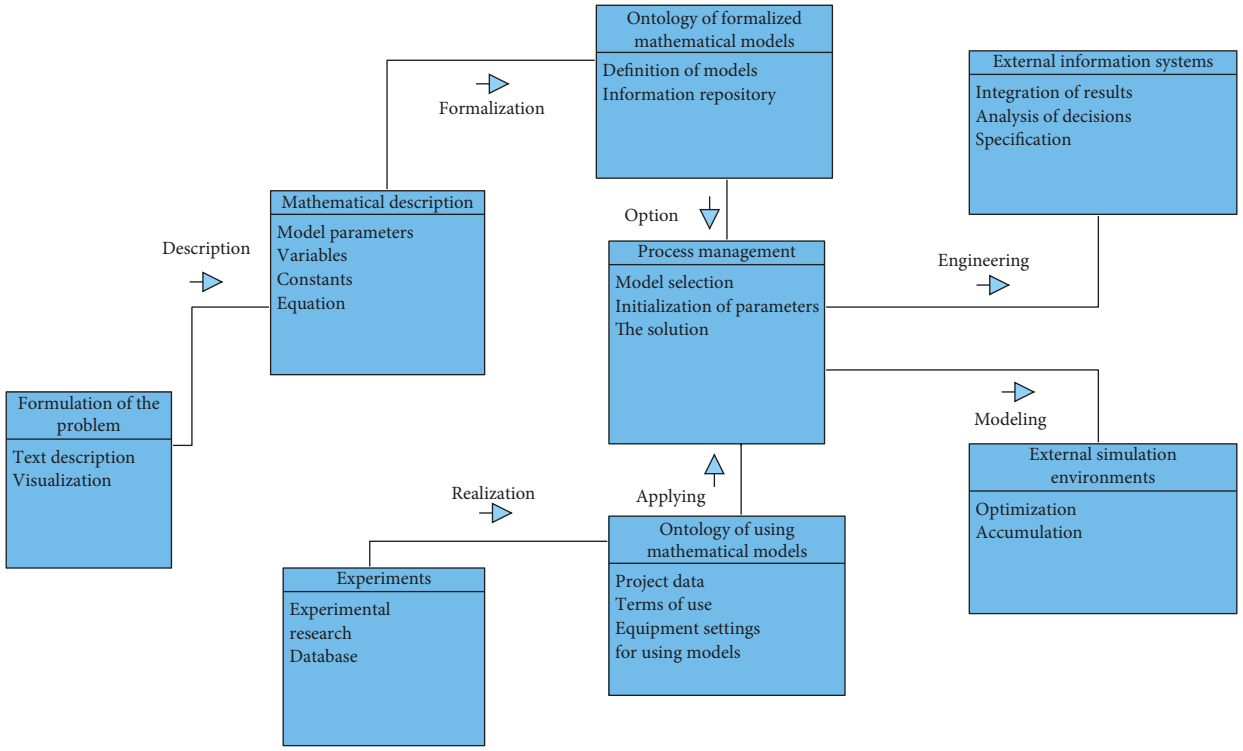


Figure 1: General scheme of implementing an ontological approach to mathematical modeling.

parameters, dependent and independent variables, and universal classes of constants. All of the above attributes of the class describe some knowledge about the mathematical model in a very explicit way, which makes representation more computer-interpreted, systematic, and more generalized in nature.

The feature of the proposed approach is that the components of the model created in this way can be reused. That is, equations, variables, and assumptions from one model can be reused when creating another model or the formed repository of mathematical models can be reused in the process of interpretation in other information systems. Thus, the process of creating mathematical models and their practical use becomes more intuitive and user-oriented, which is not very oriented in the modeling process. Each model in this approach is a specific instance of the ontology model class.

The ontology of formalized mathematical models also contains a functional representation of the model in the form of a graphical interpretation for the diagnosis of inaccuracies based on the improved model.

A subset of concepts and relationships that are fixed in the general ontological model is shown in Figure 2.

The procedural part of the ontological approach consists of a mechanism for construction based on methods of data relationship analysis, which analyzes equations in the ontological interpretation of mathematical models and translates them into expressions that can be interpreted in other external software environments. The general scheme of this approach is shown in Figure 3.

The ontology of a mathematical model consists of an operating class, the subclasses of which are various operations that occur during the implementation of the model and also contain the conditions for the implementation of each operation. This ontology also consists of a class of results, which stores the results of the model solving, as well as the results of experiments.

The model selection process control subsystem creates operators to initialize model parameters with corresponding values, creates associations between index variables and values for which it is denoted, initializes universal constants, collects actual model solution commands, and finds the appropriate solution to a set of equations.

This software-interpreted ontological approach provides the user with a number of additional features in the form of implemented functions. Among these features is symbolic processing, which directly analyzes the equations in different formats and provides their interpretation in different programming languages.

The graphical user interface is designed to display the results of solving (graphs or expressions) along with saving returns to the ontology of mathematical models and is also used to select the best instance of the model that is best suited for use in a particular application area.

Based on the analysis of the structure of interval models, the modeling process, and the features of experiments, the mathematical model from the point of view of the ontological approach is formalized by the following structures:

$$Mm = \{Ma, Mi, Mo, Par, Mr, Mc, SuMth, Mmt\}, \quad (26)$$

where Ma is the subject area within which the mathematical model is constructed or used; Mi are the descriptions of the mathematical model; Mo is a set of objects where the model

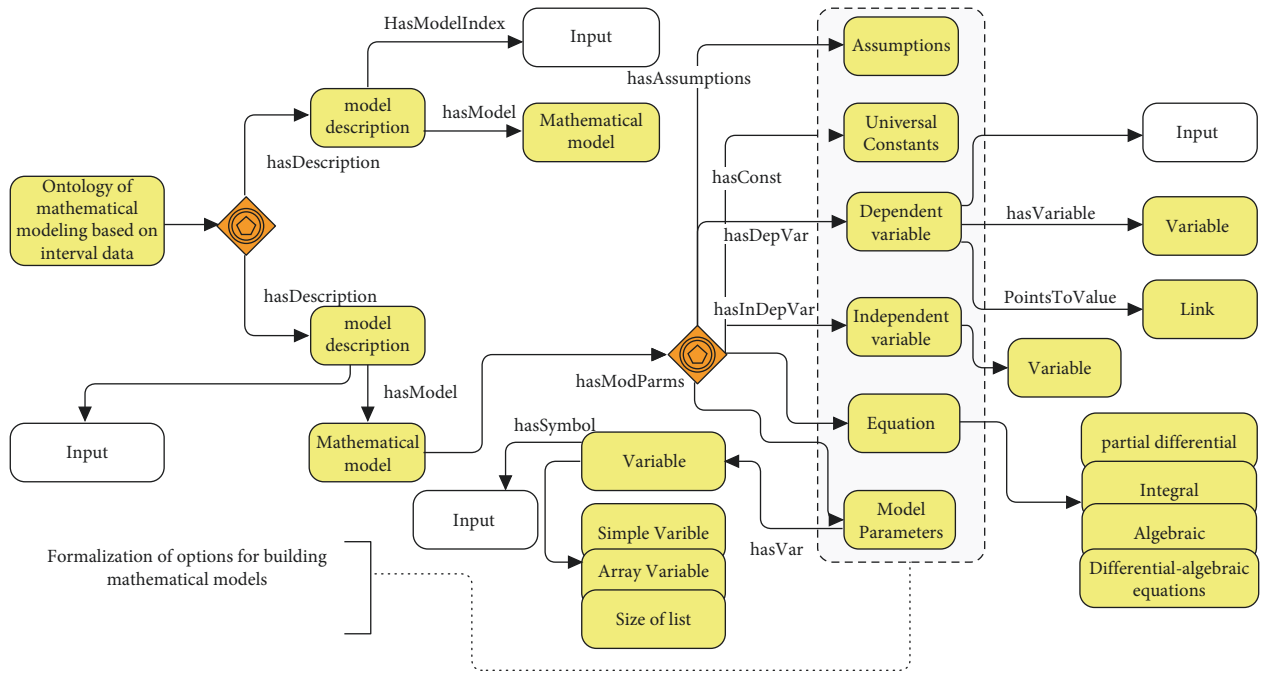


Figure 2: Scheme of description of the ontology of mathematical models on the basis of interval data.

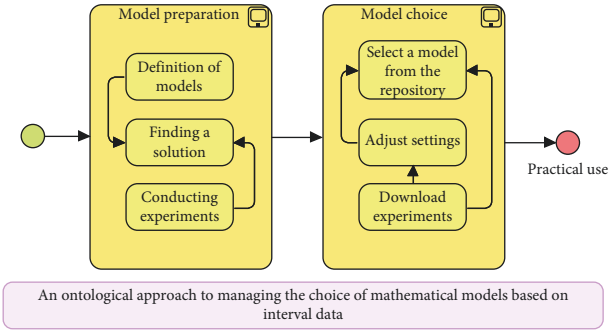


Figure 3: Scheme of implementation of the ontological approach for mathematical modeling based on interval data for practical use.

can be used; Par is a set of parameters; Mr is a set that describes the result of building object models; Mc is a set of characteristics of the experiments; $SuMth$ is a set of methods for structural identification of models; Mmt is a set of methods for identifying model parameters.

In turn, the subject area is described by a tuple

$$Ma = \langle IdMa, NmMa \rangle, \tag{27}$$

where $IdMa$ is the subject area identifier; $NmMa$ is a subject area.

Descriptions of the mathematical model have the following structure:

$$Mi = \langle IdMi, NmMi, IdMa \rangle, \tag{28}$$

where $IdMi$ is the identifier of equation; $NmMi$ is a formalized description of the equations of a mathematical model.

The structure of the description of the set of objects where the model can be used has the following representation:

$$Mo = \langle IdMo, NmMo, IdMa, IdMi \rangle, \tag{29}$$

where $IdMo$ is an object identifier; $NmMo$ is the information that describes the structure of the object of the model usage.

Tuple description of the set of parameters:

$$Par = \langle IdPar, PT, PV, IdMa, IdMi, IdMo \rangle, \tag{30}$$

where $IdPar$ is a parameter identifier; PT is a parameter type; PV are the values of model parameters.

The presentation of the results of building object models is as follows:

$$Mr = \langle IdMr, RNm, IdMa, IdMi, IdMo \rangle, \tag{31}$$

where $IdMr$ is a result identifier; RNm are the statements that describe the result.

The characteristics of the experiments are presented as follows:

$$Mc = \langle IdMc, MA, Dsc, IdMa, NA, IdMo, IdMi, IdPar \rangle, \tag{32}$$

where $IdMc$ is the identifier of the features that affect the experimental conditions; MA are the main characteristics; NA are the alternative characteristics; Dsc is a statement that describes the conditions of mathematical model usage.

Tuple for many methods of model structural identification:

$$SuMth = \langle IdMmt, NmMth, Ac, IdMth \rangle, \tag{33}$$

where $IdMmt$ is a method identifier; $NmMth$ is a method of model structure identification; Ac is the set of statements

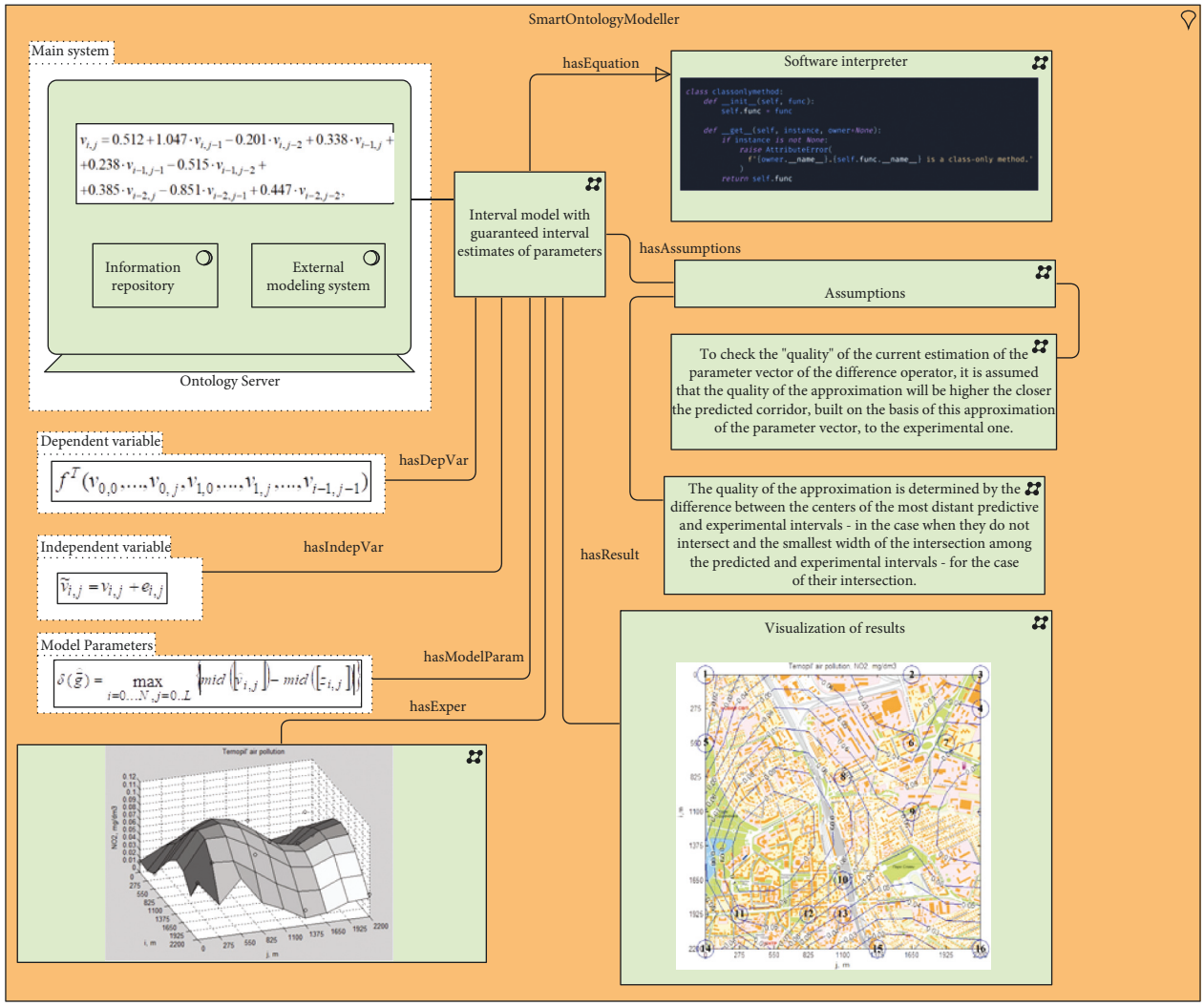


FIGURE 4: An example of implementing the ontology of the interval model for simulating the fields of harmful emission concentrations in the squat layer of the atmosphere in the conditions of large errors of observations in the SmartOntologyModeller environment.

that describes the method; *IdMth* is the identifier of the parametric identification method.

The set of methods for identifying the parameters of the models will be presented as follows:

$$Mth = \langle IdMth, NmMth, Ac \rangle, \tag{34}$$

where *IdMth* is an identifier of the model parameter identification method; *NmMth* is a method of model structure identification; *Ac* is the set of statements that describes the method.

An example of implementation of the ontological approach for constructing models of fields of harmful emission concentrations in the squat layer of the atmosphere in the conditions of large errors of observations is shown in Figure 4.

The scheme of formalization of the mathematical model using the developed tool SmartOntologyModeller reflects the main structural components within the proposed ontological approach. As seen, the information repository with a formalized model description and external modeling environment, which describes the use of software-implemented models (in this case, an interval model with guaranteed interval parameter estimates), is translated to the index representation and stored in the HasEquation attribute. The diagram shows the dependent and independent variables and parameters combined to represent the structure of the interval model with guaranteed interval estimates of the parameters. On the right side of the diagram, the process of using assumptions for the implementation of methods, the conditions of experiments, recommendations for the use of methods, and visualization of simulation results are formalized.

As an option for using the above ontological description, consider the method of constructing a mathematical model for modeling based on interval data.

Let's present this method as a sequence of steps.

(1) The user selects the subject area: *IdMa_C*. The notation "_" means the prefix of choice, C is the selection procedure.

The result is a proposed set of mathematical models for a set of Mi_C objects.

(2) Selection of the object of modeling.

The formal description of this procedure is as follows:

$$Mo\_C = \pi_{IdMo,NmMo} \cdot \left(\sigma_{Mo(IdMa)=IdMa\_C \wedge MoIdMi=IdMi\_C}(\tau(Mo))\right), \tag{35}$$

where π is the projection operation of the tuple algebra, σ is the sampling operation from the set by the given attributes, τ is the ordering operation by the values of the corresponding attributes.

The result of the operation is a selected object with a set of possible models if any of them are in the repository.

(3) Choosing the conditions of application of the model:

$$Mc\_C = \pi_{IdMc,Dsc,IdPar}\left(\sigma_{\substack{Mc(IdMa)=IdMa\_C \wedge Mc(IdMi)= \\ =IdMi\_C \wedge Mc(IdMo)=IdMo\_C}}(Mc)\right). \tag{36}$$

(4) Model selection.

For this case use the following procedure:

$$Mi\_C = \pi_{IdMi,IdMc,NmMi} \cdot \left(\sigma_{Mi(IdMa)=IdMa\_C \wedge McIdMc}(\tau(Mi))\right). \tag{37}$$

(5) To select Mi_C and Mo_C, a set is formed that represents the results of building object models using the following description:

$$Mr\_C = \pi_{IdMr,RNm}\left(\sigma_{\substack{Mr(IdMa)=IdMa\_C \wedge Mr(IdMi)= \\ =IdMi\_C \wedge Mr(IdMo)=IdMo\_C}}(\tau(Mr))\right). \tag{38}$$

If the repository does not have adequate models to describe the object, continue to build models.

(6) Choosing the conditions of model application (characteristics of the experiment):

$$Mc\_C = \pi_{IdMc,Dsc,IdPar}\left(\sigma_{\substack{Mc(IdMa)=IdMa\_C \wedge Mc(IdMi)= \\ =IdMi\_C \wedge Mc(IdMo)=IdMo\_C}}(Mc)\right). \tag{39}$$

(7) The user chooses the method of identifying the model structure

$$Mmt\_C = \pi_{IdMmt,IdMth}\left(\sigma_{\substack{Mmt(IdMa)=IdMa\_C \wedge Mmt(IdMi)= \\ =IdPi\_C \wedge Mmt(IdMo)= \\ =IdMo\_C \wedge Mmt(IdPar)=Par\_C}}(Mmt)\right). \tag{40}$$

(8) Determining the structure of the model and its parameters

$$SuMth\_C = \pi_{IdMmt,Ac,IdPar}\left(\begin{array}{c}\sigma_{SuMth(IdMa)\,=\,IdMa\_C\wedge SuMth(IdMi)\,=\,(\tau(SuMth))} \\ =\, IdMi\_C\wedge SuMth(IdMo) = IdMo\_C\wedge \\ \wedge SuMth(IdMth) = IdPar\_C\end{array}\right),$$
$$Par\_C = \pi_{IdPar,Ac,}\left(\begin{array}{c}\sigma_{Par(IdMa)\,=\,IdMa\_C\wedge Par(IdMi)\,=\,(\tau(Par))} \\ =\, IdMi\_C\wedge Par(IdMo) = IdMo\_C\end{array}\right). \tag{41}$$

The result of this operation is a set of object models.

(9) For certain Mi and Mo, a set is formed that describes the results of model construction:

$$Mr\_C = \pi_{IdMr,RNm}\left(\begin{array}{c}\sigma_{Mr(IdMa)\,=\,IdMa\_C\wedge Mr(IdMi)\,=\,(\tau(Mr))} \\ =\, IdMi\_C\wedge Mr(IdMo) = IdMo\_C\end{array}\right). \tag{42}$$

Performing steps 1-5 makes it possible to choose an adequate model for describing the object in the repository. Steps 1, 2, 6–9 are used in case of the absence of models in the repository.

The proposed ontological description makes it possible to develop the environment for modeling on the basis of interval data.

5. Results and Discussion

The practical implementation of the ontology of mathematical modeling based on interval data leads to the formation of common structural elements based on the specifics of their use for a particular subject area. The practical implementation of software as one of the options for using the developed repository of model experiments in various subject areas within the proposed ontological approach is described in this paper.

As an example of the application of the ontological approach, the problem of building models of fields of harmful emissions concentrations in a squat layer of atmosphere on the basis of macromodels in the form of difference operators is considered, which structure needs to be selected under conditions of coordination with experimental data and when big errors in observations occur. Differential equations in partial derivatives, or their difference analogs, serve as a theoretical basis for modeling the processes of pollutants spreading in the atmosphere. In addition, due to big observation errors, the boundaries of which are usually known, the difference operators are built on the basis of methods of interval data analysis.

Consider the case of describing the field of concentrations of harmful emissions of a substance in the squat layer of the atmosphere by a macromodel in the form of a difference operator (2):

$$v_{i,j} = f^T\left(v_{0,0},\ldots,v_{0,j},v_{1,0},\ldots,v_{1,j},\ldots,v_{i-1,j-1}\right)\cdot\vec{g}, \quad i = 1,\ldots,N,\ j = 1,\ldots,L, \tag{43}$$

where in our case $v_{i,j}$ is the predicted (true) value of the concentration of harmful substances in the squat layer of the atmosphere at a point in the city with discrete coordinates i, j; $\vec{g}$ is unknown vector (dimension $m \times 1$) of parameters of the difference operator.

To estimate the vector of parameters $\vec{g}$ of the difference operator, use the results of observations of the concentration of harmful substances for given discrete coordinates i, j:

$$\tilde{v}_{i,j} = v_{i,j} + e_{i,j},\ i = 1,\ldots,N,\ j = 1,\ldots,L, \tag{44}$$

where $\tilde{v}_{i,j}$ is measured value of the concentration of harmful substances in the squat layer of the atmosphere at a point in the city with discrete coordinates i, j; $e_{i,j}$ are the random limited by the amplitude errors

$$\left|e_{1,j}\right| = \left|e_{2,j}\right| = \cdots = \left|e_{i,j}\right| \le \Delta_{i,j},\ \Delta_{i,j} > 0 \forall i \ldots\ldots, N, \quad j = 1,\ldots,L, \tag{45}$$

which in the general case depend on the discrete values of the space coordinates.

Using the model of observations (44) and taking into account the limitation on the amplitude of the error (45), estimates of the concentration of harmful substances

Table 2: Example of formalized representation of mathematical models based on interval data for air pollution processes by harmful emissions from vehicles.

Attribute	Description	Value
Ma	Subject area	Harmful emissions Atmospheric pollution Emissions from vehicles
Mi	Descriptions of the mathematical model	$[\widehat{v}_j^-; \widehat{v}_j^+] = \widehat{v}_1 \cdot [\widehat{v}_{k-1}^-; \widehat{v}_{k-1}^+] + \widehat{g}_2 \cdot ([\widehat{v}_{k-2}^-; \widehat{v}_{k-2}^+] - [\widehat{v}_{k-1}^-; \widehat{v}_{k-1}^+])$ $[\widehat{v}_0^-; \widehat{v}_0^+] \subset [52{,}25; 57{,}75]$, $[\widehat{v}_1^-; \widehat{v}_1^+] \subset [44{,}65; 49{,}35]$, $\widehat{g}_1 = 0{,}8897; \widehat{g}_2 = -0{,}0261$.
Mo	Set of object characteristics	Distribution of carbon monoxide concentrations Straight section of the street Uniform traffic flow Constant emission capacity
Attr	Set of parameters	$\widehat{v}_k^-$ is a concentration CO at the k moment of time x_k is a distance u_k is an intensity of traffic flows z_k is a measured concentration $[v_k] = [v_k^-; v_k^+]$ are the interval values of carbon monoxide concentration
Mr	Many possible results	Predicted dynamics of daily cycle of changes in carbon monoxide concentrations The concentration of carbon monoxide within the observation errors A set of interval models of atmospheric pollution processes by harmful emissions from vehicles
Mc	Many characteristics of the experiments	Carbon monoxide concentration measurement error 25% Daily cycle of concentration of harmful emissions of motor transport Change in the intensity of traffic flows
Mmt	Many identification methods	Identification with a random search procedure with linear tactics Identification with the procedure of random search on the best attempt Identification with a random search procedure using a directed cone Identification with a random search procedure with adaptation of the random step distribution Identification based on the behavioral model of the bee colony
Mi	Descriptions of the mathematical model	$\widehat{v}_k = 0{,}0149 - 0{,}5788\widehat{v}_{k-2} + 0{,}7425\widehat{v}_{k-3} + 0{,}046\widehat{v}_{k-1}/\widehat{v}_{k-4}$, $k = 4, \ldots, 18$ $\widehat{v}_k = 0{,}124 - 0{,}5764\widehat{v}_{k-2} + 0{,}7078\widehat{v}_{k-3} + 0{,}0473\widehat{v}_{k-1}/\widehat{v}_{k-4} + 0{,}0159\widehat{v}_{k-1}\widehat{v}_{k-2}/\widehat{v}_{k-1}$, $k = 4, \ldots, 18$ $\widehat{v}_k = 0{,}0226 - 0{,}6114\widehat{v}_{k-2} + 0{,}7781\widehat{v}_{k-3} + 0{,}037\widehat{v}_{k-1}/\widehat{v}_{k-4} + 0{,}0282\widehat{v}_{k-1}\widehat{v}_{k-4}/\widehat{v}_{k-2}$, $k = 4, .., 18$
Mo	Set of object characteristics	Dynamics of nitrogen dioxide concentrations Uniform intensity of traffic flows Straight section of the street
Attr	Set of parameters	$\widehat{v}_k^-$ is a concentration NO_2 at the k moment of time u_k is an intensity of traffic flows x_k is a distance
Mr	Many possible results	Intervals of predicted values of nitrogen dioxide concentrations Intervals of measured values of concentrations of nitrogen dioxide Interval model with a simpler structure
Mc	Many characteristics of the experiments	Error of measurement of concentrations of nitrogen dioxide 15% Control intensity of traffic flows Uniform period of measurements
Mi	Descriptions of the mathematical model	$v_{i,j} = 0.512 + 1.047 \cdot v_{i,j-1} - 0.201 \cdot v_{i,j-2} + 0.338 \cdot v_{i-1,j} + 0.238 \cdot v_{i-1,j-1} - 0.515 \cdot v_{i-1,j-2} +$ $+0.385 \cdot v_{i-2,j} - 0.851 \cdot v_{i-2,j-1} + 0.447 \cdot v_{i-2,j-2}$
Mo	Set of object characteristics	Distribution of nitrogen dioxide concentrations Uniform intensity of traffic flows Center part of the city
Attr	Set of parameters	$v_{i,j}$ is a concentration NO_2 in the point with discrete coordinates i, j
Mr	Many possible results	Intervals of predicted values of nitrogen dioxide concentrations Intervals of measured values of concentrations of nitrogen dioxide Interval model with a simpler structure
Mc	Many characteristics of the experiments	Error of measurement of concentrations of nitrogen dioxide 15% Uniform period of measurements

obtained on the basis of experimental data acquire interval representation

$$\left[z_{i,j}^{-}; z_{i,j}^{+}\right] = \left[\left(\widetilde{v}_{i,j} - \Delta_{i,j}\right); \left(\widetilde{v}_{i,j} + \Delta_{i,j}\right)\right], \quad i = 1, \ldots, N,\ j = 1, \ldots, L, \tag{46}$$

where $[z_{i,j}^{-}; z_{i,j}^{+}]$ is a guaranteed interval which includes the true unknown concentration of the substance, i.e.,

$$v_{i,j} \in \left[z_{i,j}^{-}; z_{i,j}^{+}\right] \forall i = 1, \ldots, N, j = 1, \ldots, L. \tag{47}$$

Then, substituting in expression (5) the value of $v_{i,j}$, which is given by the difference operator (43), the conditions for matching the experimental values of concentrations with the simulated ones are obtained.

$$z_{i,j}^{-} \le f^{T}\left(v_{0,0}, \ldots, v_{0,k}, v_{1,0}, \ldots, v_{1,j}, \ldots, v_{i,j}\right) \cdot \overrightarrow{g} \le z_{i,j}^{+} \quad i = 1, \ldots, N, j = 1, \ldots, L. \tag{48}$$

Further, according to the description in paragraph 2, it is necessary to solve the problem of structural and parametric identification of the model using ABC algorithms.

One of the initial structures generated on the basis of the ontological description has the following form:

$$v_{i,j} = g_1 + g_2 \cdot v_{i-1,j} + g_3 \cdot v_{i,j-1} + g_4 \cdot v_{i-1,j-1}. \tag{49}$$

As a result of solving the problem of structural and parametric identification, a difference operator that adequately describes the spatial distribution of concentrations of nitrogen dioxide is obtained:

$$\begin{aligned} v_{i,j} = {} & 0.512 + 1.047 \cdot v_{i,j-1} - 0.201 \cdot v_{i,j-2} + 0.338 \cdot v_{i-1,j} \\ & + 0.238 \cdot v_{i-1,j-1} - \\ & - 0.515 \cdot v_{i-1,j-2} + 0.385 \cdot v_{i-2,j} - 0.851 \cdot v_{i-2,j-1} \\ & + 0.447 \cdot v_{i-2,j-2}. \end{aligned} \tag{50}$$

The mathematical models obtained in this way are stored in the repository.

If the object is changed, then in general the identification scheme remains unchanged.

The authors of this article have developed a number of models not only for predicting the spatial distribution of nitrogen dioxide concentrations for different conditions but also for predicting the dynamics of this harmful substance or the dynamics of carbon monoxide for different conditions. However, for their effective use, it is necessary to obtain a correct ontological description.

Based on the developed method of ontological description of the mathematical modeling of objects on the basis of interval data, some results of such description are shown in Table 2.

Based on the method of choosing a mathematical model within the ontological approach for modeling based on interval data, it is possible to switch models from the information repository depending on the conditions and specifics of the relevant experimental studies. The ability to control the switching process was practically implemented in the web-based information system SmartOntologyModeller.

Table 3: The results of predicting the nitrogen dioxide concentrations at control points.

Point No	i	x_i (m)	j	y_j (m)	$v^{-}_{i,j}$ (mg/dm^3)	$v^{+}_{i,j}$ (mg/dm^3)
1	0	0	0	0	0.011	0.019
2	0	0	6	1650	0.015	0.025
3	0	0	8	2200	0.015	0.025
4	1	275	8	2200	0.030	0.050
5	2	550	0	0	0.015	0.025
6	2	550	6	1650	0.015	0.025
7	2	550	7	1925	0.057	0.095
8	3	825	4	1100	0.065	0.109
9	4	1100	6	1650	0.045	0.075
10	6	1650	4	1100	0.069	0.115
11	7	1925	1	275	0.065	0.109
12	7	1925	3	825	0.068	0.113
13	7	1925	4	1100	0.036	0.060
14	8	2200	0	0	0.056	0.094
15	8	2200	5	1375	0.015	0.025
16	8	2200	8	2200	0.023	0.038

Table 2 contains three columns that correspond to the description of the ontological model, namely: Attribute Description Value. These structural elements represent the subject area, object, modeling conditions (two groups of conditions), variables, etc. Also, for the specified conditions of application, there is a repository of models (4 such models are given in the table).

Thus, having a repository for the specified object (concentrations of harmful emissions in the squat layer of the atmosphere), the first five steps of the above method of choosing a mathematical model for modeling based on interval data can be applied:

Step 1. Selection of the subject area: $IdMa\_C$ is "pollution of the squat layer of the atmosphere by harmful emissions from vehicles."

Step 2. Selection of the Mo_C modeling object is "concentration of nitrogen dioxide emissions from vehicles."

Step 3. Selection of the conditions for the application of the model Mc_C is "error in measuring the concentration of nitrogen dioxide at the level of 15%; control of traffic intensity; uniform period of measurements."

Step 4. Selection of a model from the repository for approximation of the fields of concentrations of nitrogen dioxide emissions from vehicles in Ternopil city, taking into account the results obtained in the previous steps:

$$\begin{aligned} v_{i,j} = {} & 0.512 + 1.047 \cdot v_{i,j-1} - 0.201 \cdot v_{i,j-2} + 0.338 \cdot v_{i-1,j} \\ & + 0.238 \cdot v_{i-1,j-1} - 0.515 \cdot v_{i-1,j-2} + 0.385 \cdot v_{i-2,j} \\ & - 0.851 \cdot v_{i-2,j-1} + 0.447 \cdot v_{i-2,j-2}. \end{aligned} \tag{51}$$

Step 5. For the obtained model, tabular and visual results of its use from the repository can also be

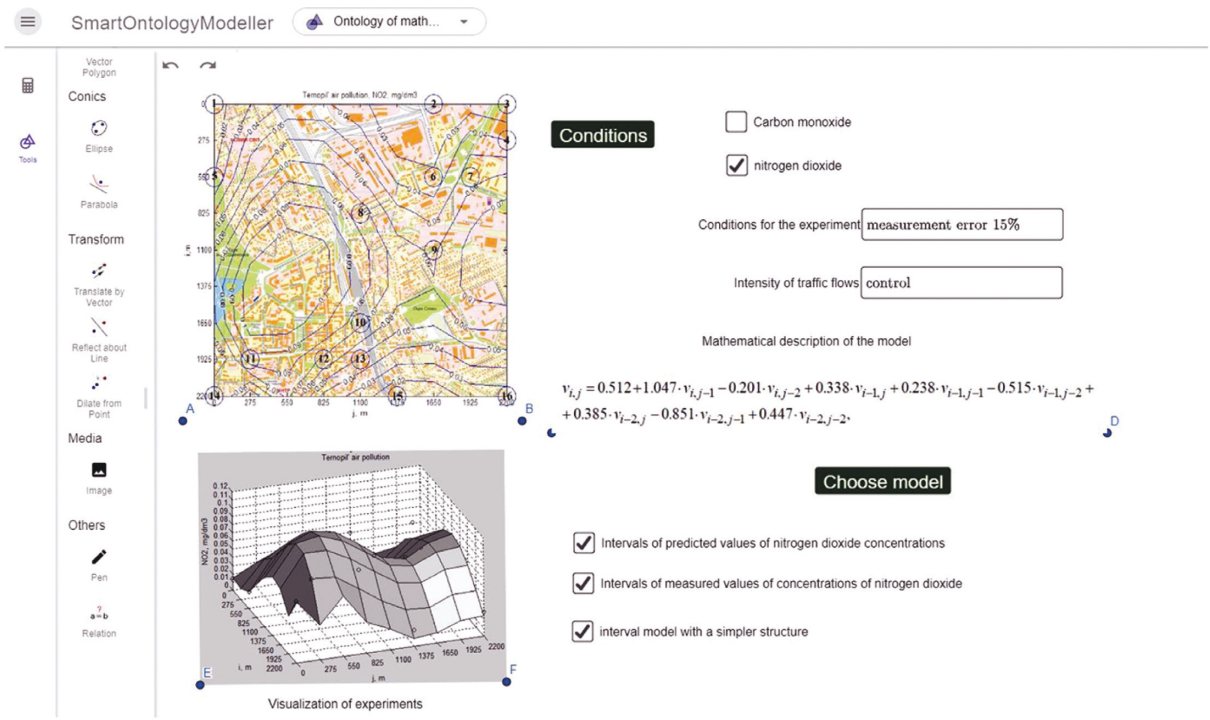


FIGURE 5: Example of switching the mathematical model depending on the change in control characteristics and conditions of experiments in SmartOntologyModeller environment.

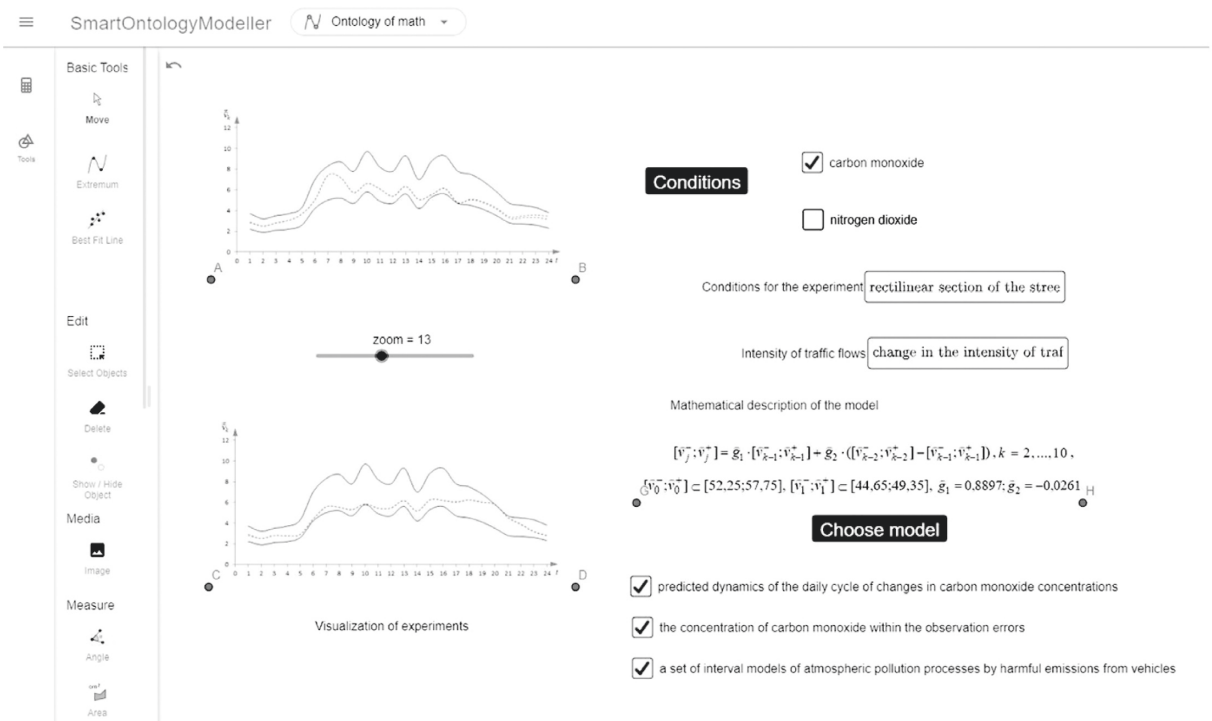


FIGURE 6: Example of switching the mathematical model due to the changes in the conditions of the simulation environment or the conditions of the corresponding experiment.

received. For example, Table 3 compares the results of predicting nitrogen dioxide concentrations and those measured at control points.

Figure 5 shows an example of switching by choosing a mathematical model based on interval data depending on changes in the subject characteristics of the model. Switching

occurs by changing the conditions of the simulation environment.

It should be noted that in the case of another task, such as modeling the dynamics of concentrations of harmful carbon monoxide emissions during the day in a certain area of the city and the existing repository of these models, the scheme of applying the method of choosing a mathematical model for modeling based on interval data will be the same. However, in the fifth step, the results will be presented adequately to the selected object. For this case, the results are presented in Figure 6.

The accuracy of the model of the dynamics of atmospheric pollution by vehicles is characterized by the equivalent accuracy of the measurement experiment. If the conditions of the experiment are changed, the accuracy of the model may also change. The advantage of the proposed approach is the saving of resources, which is achieved through the reuse of the developed model repository for the relevant objects from the repository.

Figure 6 shows the results of the corresponding switching, related to changes in the conditions of tracking traffic flows and according to the characteristics of the section of the street under research.

The connected Python toolkit allows the user to select a sample of the model and the corresponding operational example, after which the operators can build using the appropriate libraries that interpret equations from formatted, indexed parts, initialize model parameters based on the corresponding sample of operation, and finally allow the model to build the necessary solution. When calculating, the results are interpreted in the appropriate graphical interface using graphs, and tables, resulting in files, as well as other results that are stored in the operational part of the mathematical model with the appropriate refinements. This refinement will allow in the future choosing the right models depending on the specifics of the conditions of the experiments and the relevant subject area.

6. Conclusions

The inductive approach to mathematical modeling of complex systems based on interval data is limited to strictly formalized and algorithmic procedures. The proposed ontological superstructure for mathematical modeling of objects based on interval data makes it possible to generate tools in the form of software for building interval models. On the other hand, in the presence of previously constructed interval discrete models, the ontological superstructure makes it possible to create a repository of these models, as well as to manage this repository. In this case, it serves as a "switch" that choose the most accurate and adequate model from the repository of previously created models. The advantage of the proposed approach is illustrated by the example of modeling the processes of air pollution by harmful emissions from vehicles. In particular, the example illustrates the "switching" of the choice of a mathematical model based on interval data depending on changes in the subject characteristics of the model. Switching occurs by changing the conditions of the simulation environment.

In further research, the implementation of tools for integration of the offered ontology in external information systems for the purpose of their expansion and qualitative improvement is planned.

Data Availability

The data cited in this study are available from the published papers or the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by the Ministry of Science and Higher Education in Poland under the program "Regional Initiative of Excellence," 2019–2022, project no. 015/RID/2018/19, total funding amount 10,721,040,00 PLN and partially supported by the Ministry of Education and Science of Ukraine under the grant "Mathematical and computer modeling of objects with distributed parameters based on a combination of ontological and interval analysis" January 2022–December 2024, state registration number 0122U001497.

References

[1] B. Rittle-Johnson, "Developing mathematics knowledge," *Child Development Perspectives*, vol. 11, no. 3, pp. 184–190, 2017.

[2] N. Gorgorió, L. Albarracín, J. Ärlebäck, A. Laine, R. Newton, and A. Villarreal, *Fundamental mathematical knowledge: progressing its specification*, Linköping University Electronic Press, Linköping, Sweden, 2019, http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-157096.

[3] H. C. Hill, M. L. Blunk, C. Y. Charalambous et al., "Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study," *Cognition and Instruction*, vol. 26, no. 4, pp. 430–511, 2008.

[4] P. Suresh, G. Joglekar, S. Hsu et al., "Onto MODEL: ontological mathematical modeling knowledge management," *Computer Aided Chemical Engineering*, vol. 25, pp. 985–990, 2008.

[5] P. Cimiano and J. Völker, "Text2Onto, natural language processing and information systems," in *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, pp. 227–238, Alicante, Spain, June 2005.

[6] P. Grenon and B. Smith, "Foundations of an ontology of philosophy," *Synthese*, vol. 182, no. 2, pp. 185–204, 2011.

[7] M. Husáková and V. Bureš, "Formal ontologies in information systems development: a systematic review," *Information*, vol. 11, no. 2, p. 66, 2020.

[8] Y. Aït-Ameur, I. Ait-Sadoune, K. Hacid, and L. Mohand Oussaid, "Formal modelling of ontologies within Event-B," in *Proceedings of the First International Workshop on Handling IMPlicit and EXplicit knowledge in formal system development*, Xi'an, China, 2017.

[9] M. Slavíčková, "Implementation of digital technologies into pre-service mathematics teacher preparation," *Mathematics*, vol. 9, no. 12, 2021.

[10] C. Lange and M. Kohlhase, "SWiM, emerging technologies for semantic work environments," in *Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications*, J. . Rech, B. Decker, and E. Ras, Eds., IGI Global, Hershey, PA, USA, pp. 47–68, 2008.

[11] C. Lange, "Ontologies and languages for representing mathematical knowledge on the Semantic Web," *Semantic Web*, vol. 4, no. 2, pp. 119–158, 2013.

[12] P. Suresh, S.-H. Hsu, G. V. Reklaitis, V. Venkatasubramanian, and V. Venkatasubramanian, "OntoMODEL: ontological mathematical modeling knowledge management in pharmaceutical product development, 2: applications," *Industrial & Engineering Chemistry Research*, vol. 49, no. 17, pp. 7768–7781, 2010.

[13] A. Asperti, B. Buchberger, Davenport, and H. James, "Mathematical knowledge management," in *Proceedings of the 2nd International Conference, MKM 2003*, vol. 2594Springer, Berlin, Germany, February 2003.

[14] J. Cao, Y. L. He, and Q. Zhu, "An ontology-based procedure knowledge framework for the process industry," *Canadian Journal of Chemical Engineering*, vol. 99, no. 2, pp. 530–542, 2021.

[15] T. Schneider and M. Šimkus, "Ontologies and data management: a brief survey," *KI - Künstliche Intelligenz*, vol. 34, no. 3, pp. 329–353, 2020.

[16] R. Falbo, A. Natali, P. Mian, G. Bertollo, and F. Ruy, "ODE: ontology-based software development environment," *IX Argentine Congress of Computer Science*, pp. 1124–1135, 2003.

[17] A. Belfadel, E. Amdouni, J. Laval, C. Cherifi, and N. Moalla, "Ontology-based software capability container for RESTful APIs," in *Proceedings of the 9th IEEE International Conference on Intelligent Systems (IS 2018)*, pp. 466–473, Madeira, Portugal, September 2018.

[18] B. Moreno Torres, C. Völker, S. M. Nagel, T. Hanke, and S. Kruschwitz, "An ontology-based approach to enable data-driven research in the field of NDT in civil engineering," *Remote Sensing*, vol. 13, no. 12, p. 2426, 2021.

[19] G. R. Roldan-Molina, J. R. Mendez, I. Yevseyeva, and V. Basto-Fernandes, "Ontology fixing by using software engineering technology," *Applied Sciences*, vol. 10, no. 18, p. 6328, 2020.

[20] S. Chimalakonda and K. V. Nori, "An ontology based modeling framework for design of educational technologies," *Smart Learning Environments*, vol. 7, no. 1, 2020.

[21] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K.-S. Kwak, "Traffic accident detection and condition analysis based on social networking data," *Accident Analysis & Prevention*, vol. 151, Article ID 105973, 2021.

[22] F. Ali, S. El-Sappagh, and D. Kwak, "Fuzzy ontology and LSTM-based text mining: a transportation network monitoring system for assisting travel," *Sensors*, vol. 19, no. 2, p. 234, 2019.

[23] F. Ali, P. Khan, K. Riaz, T. Abuhmed, D. Park, and K. S. Kwak, "A fuzzy ontology and SVM-based web content classification system," *IEEE Access*, vol. 5, Article ID 25781, 2017.

[24] H. Madala and A. Ivakhnenko, *Inductive learning algorithms for complex systems modelling*, CRC Press, Boca Raton, FL, USA, 1994.

[25] A. G. Ivakhnenko, "Polynomial theory of complex systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-1, no. 4, pp. 364–378, 1971.

[26] A. Ivakhnenko and G. Ivakhnenko, "The review of problems solvable by algorithms of the group method of data handling (GMDH)," *Pattern Recognition and Image Analysis*, vol. 5, no. 4, pp. 527–535, 1995.

[27] A. Ivakhnenko and V. Lapa, "Cybernetics and forecasting techniques," *Modern Analytic and Computational Methods in Science and Mathematics*, American Elsevier, 8 edition, 1967.

[28] H. Pidnebesna and V. Stepashko, "Ontology-based design of inductive modeling tools," in *Proceedings of the 2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 731–734, Deggendorf, Germany, September 2021.

[29] A. Pavlov, H. Pidnebesna, and V. Stepashko, "Ontology-based approach to construction of intelligent interface for inductive modeling tools," in *Proceedings of the 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 26–29, Zbarazh, Ukraine, September 2020.

[30] H. Pidnebesna, "On constructing ontology of the GMDH-based inductive modeling domain," in *Proceedings of the 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 511–513, Lviv, Ukraine, September 2017.

[31] S. J. Farlow, "The GMDH algorithm of Ivakhnenko," *The American Statistician*, vol. 35, no. 4, pp. 210–215, 1981.

[32] A. Ivakhnenko, "The group method of data handling in long-range forecasting," *Technological Forecasting and Social Change*, vol. 12, no. 2-3, pp. 213–227, 1978.

[33] M. Debellis, *A practical guide to building OWL ontologies using protégé 5.5 and plugins*, 2021.

[34] I. Urbieta, M. Nieto, M. García, and O. Otaegui, "Design and implementation of an ontology for semantic labeling and testing: automotive global ontology (AGO)," *Applied Sciences*, vol. 11, no. 17, p. 7782, 2021.

[35] L. Wendelberg, "An ontological framework to facilitate early detection of 'radicalization' (OFEDR)-A three world perspective," *Journal of Imaging*, vol. 7, no. 3, p. 60, 2021.

[36] M. A. Musen, "The protégé project," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015.

[37] S. W. Tu, H. Eriksson, J. H. Gennari, Y. Shahar, and M. A. Musen, "Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTÉGÉ-II to protocol-based decision support," *Artificial Intelligence in Medicine*, vol. 7, no. 3, pp. 257–289, 1995.

[38] A. Sattar, E. Salwana, M. Nazir, M. Ahmad, A. Kamil, and A. Mahmood, "Comparative analysis of methodologies for domain ontology development: a systematic review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, p. 2020, 2020.

[39] M. Dyvak, O. Papa, A. Melnyk, A. Pukas, N. Porplytsya, and A. Rot, "Interval model of the efficiency of the functioning of information web resources for services on ecological expertise," *Mathematics*, vol. 8, no. 12, p. 2116, 2020.

[40] M. Dyvak, "Parameters identification method of interval discrete dynamic models of air pollution based on artificial bee colony algorithm," in *Proceedings of the 2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 130–135, Deggendorf, Germany, September 2020.

[41] M. Dyvak, O. Kozak, and A. Pukas, "Interval model for identification of laryngeal nerves," *Przeglad Elektrotechniczny*, vol. 86, no. 1, pp. 139-140, 2010.

[42] M. Dyvak, N. Porplytsya, I. Borivets, and M. Shynkaryk, "Improving the computational implementation of the parametric identification method for interval discrete dynamic

models," in *Proceedings of the Proceedings of the 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 533–536, Lviv, Ukraine, September 2017.

[43] N. Porplytsya and M. Dyvak, "Interval difference operator for the task of identification recurrent laryngeal nerve," in *Proceedings of the 16th International Conference on Computational Problems of Electrical Engineering*, pp. 156–158, Lviv, Ukraine, September 2015.

[44] Y. Kedrin, M. Dyvak, A. Pukas, I. Voytyuk, Y. Maslyiak, and O. Papa, "Features of artificial bee colony based algorithm realization for parametric identification method of the interval discrete dynamic models," in *Proceedings of the 2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 239–245, Deggendorf, Germany, September 2020.

[45] M. Dyvak, A. Melnyk, A. Kovbasistyi, R. Shevchuk, O. Huhul, and V. Tymchyshyn, "Mathematical modeling of the estimation process of functioning efficiency level of information web-resources," in *Proceedings of the 2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 492–496, Deggendorf, Germany, September 2020.

[46] N. Ocheretnyuk, I. Voytyuk, M. Dyvak, and M. Ye, "Features of structure identification the macromodels for nonstationary fields of air pollution from vehicles," in *Proceedings of the Modern Problems of Radio Engineering, Telecommunications and Computer Science*, p. 444, Lviv, Ukraine, May 2012.

[47] M. Dyvak, N. Porplytsya, and Y. Maslyiak, "Modified method of structural identification of interval discrete models of atmospheric pollution by harmful emissions from motor vehicles," in *Advances in Intelligent Systems and Computing IV*, N. Shakhovska and M. M. O. Advances, Eds., vol. 1080, pp. 491–507, Springer, Cham, Switzerland, 2019.

[48] N. Ocheretnyuk, M. Dyvak, T. Dyvak, and I. Voytyuk, "Structure identification of interval difference operator for control the production process of drywall," in *Proceedings of the 12th Int. Conf. on Experience of Designing and Application of CAD Systems in Microelectronics*, pp. 262–264, CADSM), Lviv, UKraine, February 2013.

[49] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," technical report—tr06," Erciyes University, Kayseri, Turkey, 2005.

[50] A. Stadnicki, F. Filip Pietron, and P. Burek, "Towards a modern ontology development environment," *Procedia Computer Science*, vol. 176, pp. 753–762, 2020.

[51] V. N. Dolzhenkov, "Software tools for ontology development," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 935–941, 2020.

[52] R. Pigazzi, C. Confalonieri, M. Rossoni, E. Gariboldi, and G. Colombo, "Ontologies as a tool for design and material engineers," *ASME 2020 International Mechanical Engineering Congress and Exposition*, vol. 6, 2020.

[53] J. M. Patel, "Relational databases and SQL language," *Getting Structured Data from the Internet*, pp. 225–275, 2020.

[54] Y. N. Silva, I. Almeida, and M. Queiroz, "Sql, proceedings of the 47th ACM technical symposium on computing science education - sigcse '16," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 413–418, Association for Computing Machinery, Memphis TN USA, March 2016.

[55] B. Kulik, "A logic programming system based on cortege algebra," *Journal of Computer and Systems Sciences International*, pp. 159–170, 1995.

[56] B. Kulik and A. Fridman, "N-tuple algebra as a generalized theory of relations," , 2021.

[57] A. Melnyk and R. Pasichnyk, "System of semantic classes for test's generation," in *Proceedings of the 2010 International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)*, p. 206, Lviv, UKraine, February 2010.

[58] B. A. Kulik and A. Y. Fridman, "N-tuple algebra as a unifying system to process data and knowledge," *Advances in Computer and Electrical Engineering*, pp. 602–615, 2019.

[59] G. Rossum, *Python reference manual. Technical Report*, CWI (Centre for Mathematics and Computer Science), Amsterdam, Netherlands, 1995.

[60] T. Kohn, G. van Rossum, G. B. Bucher II, Talin, and I. Levkivskyi, "Dynamic pattern matching with Python," in *Proceedings of the 16th ACM SIGPLAN International Symposium on Dynamic Languages (DLS 2020)*, Association for Computing Machinery, New York, NY, USA, pp. 85–98, November 2020.

[61] J. Gosling, B. Joy, G. Steele, and G. Bracha, *Java(TM) language specification*, Addison-Wesley Professional, Boston, MA, USA, 3rd edition, 2005.

[62] D. Saenz, *Advanced Java programming (Java SE 7)*, Virtual Training Company, 2013.

[63] J. Juneau, *Java 9 recipes: a problem-solution approach*, Apress, New York, NY, USA, 3rd edition, 2017.

Hindawi
Complexity
Volume 2022, Article ID 3123475, 15 pages
https://doi.org/10.1155/2022/3123475

Research Article

Integration of Multiple Models with Hybrid Artificial Neural Network-Genetic Algorithm for Soil Cation-Exchange Capacity Prediction

Mahmood Shahabi,[1] Mohammad Ali Ghorbani,[1] Sujay Raghavendra Naganna,[2] Sungwon Kim,[3] Sinan Jasim Hadi,[4] Samed Inyurt,[5] Aitazaz Ahsan Farooque,[6,7] and Zaher Mundher Yaseen[8,9,10]

[1]*Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran*
[2]*Department of Civil Engineering, Siddaganga Institute of Technology, Tumakuru 572103, Karnataka, India*
[3]*Department of Railroad Construction and Safety Engineering, Dongyang University, Yeongju 36040, Republic of Korea*
[4]*Department of Real Estate Development and Management, Faculty of Applied Sciences, Ankara University, Ankara, Turkey*
[5]*Faculty of Engineering and Architecture, Department of Geomatics Engineering, Tokat Gaziosmanpaşa University, Tokat, Turkey*
[6]*Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE C1A4P3, Canada*
[7]*School of Climate Change and Adaptation, University of Prince Edward Island, Charlottetown, PE C1A4P3, Canada*
[8]*New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Nasiriyah 64001, Iraq*
[9]*Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia*
[10]*Adjunct Research Fellow, USQ's Advanced Data Analytics Research Group, School of Mathematics Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia*

Correspondence should be addressed to Zaher Mundher Yaseen; yaseen@alayen.edu.iq

Received 20 February 2022; Revised 23 April 2022; Accepted 5 May 2022; Published 13 June 2022

Academic Editor: Gonzalo Farias

The potential of the soil to hold plant nutrients is governed by the cation-exchange capacity (CEC) of any soil. Estimating soil CEC aids in conventional soil management practices to replenish the soil solution that supports plant growth. In this study, a multiple model integration scheme supervised with a hybrid genetic algorithm-neural network (MM-GANN) was developed and employed to predict the accuracy of soil CEC in Tabriz plain, an arid region of Iran. The standalone models (i.e., artificial neural network (ANN) and extreme learning machine (ELM)) were implemented for incorporation into the MM-GANN. In addition, it was tested to enhance the prediction accuracy of the standalone models. The soil parameters such as clay, silt, pH, carbonate calcium equivalent (CCE), and soil organic matter (OM) were used as model inputs to predict soil CEC. With the use of several evaluation criteria, the results showed that the MM-GANN model involving the predictions of ELM and ANN models calibrated by considering all the soil parameters (e.g., Clay, OM, pH, silt, and CCE) as inputs provided superior soil CEC estimates with a Nash Sutcliffe Efficiency (NSE) = 0.87, Root Mean Square Error (RMSE) = 2.885, Mean Absolute Error (MAE) = 2.249, Mean Absolute Percentage Error (MAPE) = 12.072, and coefficient of determination (R^2) = 0.884. The proposed MM-GANN model is a reliable intelligence-based approach for the assessment of soil quality parameters intended for sustainability and management prospects.

1. Introduction

Cation-exchange capacity (CEC) refers to the extent of soil's capacity to preserve exchangeable cations, the like of which have a direct bearing on the soil fertility triangle [1]. Soil CEC is a sensitive indicator of natural and human-induced perturbations over soil profile and groundwater [2]. Monitoring changes in soil CEC can assist in predicting whether soil quality has degraded, improved, or sustained under diverse agricultural or forestry schemes. In the course of conventional soil management practices to replenish the soil solution that supports plant growth, the negatively charged clay particles and organic substances adsorb and hold on positively charged soil nutrients (e.g., NH_4^+, K^+, Mg^{2+}, Ca^{2+}, etc.) via electrostatic forces [3, 4]. The preferential adsorption of cations is as per the sequence: $Al^{3+} > Ca^{2+} > Mg^{2+} > K^+ = NH_4^+ > Na^+$ [5]. Depending on the soil structure, CEC clearly demonstrates the shrink-swell potential of any soil; a high CEC value (>40 meq/100 g) denotes that a soil structure will recuperate gradually and sometimes can show expansive behavior. In contrast, a soil with low CEC value (<10 meq/100 g) will have a reduced capacity to hold water and end up being acidic rapidly [6]. Soil CEC can fluctuate according to clay percentage, soil pH, ionic strength, soil-to-solution ratio, clay type, and changing organic matter composition. It is sometimes affected by the redistribution of cations (exchange kinetics) in the soil attributed to soil solution buffering and solute transport. CEC also enables the categorization of certain soils including oxisols, vertisols, alfisols, mollisols, and ultisols [7]. In general, the organic matter enriches soils and, usually, clays (except kaolinite) have a high CEC, while sands have no CEC. For agriculture, the preferred value of CEC is >10 meq/100 g for exchange between plant root hairs and soils [8]. The leaching of contaminants into the underlying aquifer system is usually affected by CEC and percent base saturation which are eloquent indices of soil fertility and nutrient retention capacity. In areas of intensive irrigation, the continuous use of inorganic fertilizers (in excess) inundates the soil profile with more nutrients and thereby flushes a plume of contaminants into the groundwater [9]. Hence, in the early stages of agriculture, it is necessary to estimate CEC for determining the supplemental nutrient needs or to remove excess salts which influence soil structure and agricultural productivity. Soil CEC is a sensitive indicator of natural and human-induced perturbations over soil profile and groundwater. Monitoring changes in soil CEC can assist in predicting whether soil quality has degraded, improved, or sustained under diverse agricultural or forestry schemes [10].

Various methods for direct measurement of soil CEC have been reported and extensively discussed in the literature [11–13]. Multiple comparison of CEC estimation techniques is presented by Conradie and Kotze [14]. In addition, there exist several ancillary approaches such as pedotransfer functions (PTF) for estimating CEC based on easily measured soil's physical and chemical properties [15–18]. Several other researchers conducted studies on the functional relationships between CEC, water retention, and particle-size distribution. Lambooy investigated the influence of CEC on the water retention characteristics of soils [19]. Implementing multiple regressions, Parfitt et al. estimated CEC by taking into account soil organic carbon and clay content [20]. Krogh et al. modeled CEC rates of Danish soils by using clay and organic matter content as input variables through multiple linear regression analysis [21]. The actual CEC of agricultural soils was found to be directly related to the estimated charge of clay and organic carbon in the soil mass at the actual pH [22]. Using soil organic matter and noncarbonate clay contents as predictors, Seybold et al. explained the variation in CEC for several soil horizons based on soil pH, mineralogy class, taxonomic family, and CEC-activity class [23]. Fooladmand derived PTFs using multiple linear regression between CEC and soil textural data including sand content, clay content, geometric mean particle size diameter, the soil particle-size distribution, and soil organic matter content [24]. Several PTFs relating soil CEC with soil's sand, silt or clay fractions, and soil organic carbon content were evaluated by Khodaverdiloo et al. taking into account calibration dataset size on the prediction accuracy of soil CEC [25]. These classical pedotransfer function-based approaches often suffer from a high degree of inaccuracy due to spatial scale dependence, nonlinear relationships among variables, and incompetence to handle mixed data [26]. Hence, the motivation of the current state of the art is directed toward a new research era where more intelligent models should be explored in this field.

Recent research studies have focused on improving the estimation accuracy of soil CEC by means of artificial intelligence (AI) techniques. Artificial neural network (ANN) based PTFs have become popular to predict/estimate soil CEC of different soil types under diverse climatic zones [27–31]. Kalkhajeh et al. conducted the accurate prediction of soil CEC using different data-driven models [32]. They compared the performance of multiple linear regressions (MLR), adaptive neurofuzzy inference system (ANFIS), multilayer perceptron (MLP), and radial basis function (RBF) based ANN models for predicting the soil CEC using the bulk density, calcium carbonate, organic carbon, clay, and silt content (%) of the soil as input variables. The MLP model gave the most reliable prediction of soil CEC. A set of AI models along with empirical PTFs were developed and evaluated by Ghorbani et al. [33]; the authors found the most influential soil properties that influence soil CEC through sensitivity analysis. The ANFIS model provided the superior performance to RBF, MLP, MLR, and empirical PTFs while estimating soil CEC. Arthur [2] presented an ANN based methodology for estimating CEC from soil water content at different relative humidity ranges. Relatively few studies utilize a support vector machine (SVM), random forests (RF), genetic expression programming (GEP), multivariate adaptive regression splines (MARS), and a subtractive clustering algorithm based ANFIS for estimating soil CEC using readily measured soil properties as inputs [34–38]. A hybrid model integrating ant colony optimization (ACO) algorithm with ANFIS improved the prediction accuracy of soil CEC accompanied by an optimal choice of input subset which comprised soil properties (e.g., soil organic matter,

clay, silt, pH, and bulk density) [39]. Although there has been noticeable progress in AI implementation in the field of geoscience, the enthusiasm for developing and exploring more reliable intelligent predictive models is still an ongoing research era. In addition, the applications of hybrid AI models have been observed remarkably reported in the literature and for diverse engineering and sciences domains [40–43]. As a result, the inspiration for developing multiple learning intelligent models is investigated here for modeling soil CEC.

Soil CEC is a sensitive indicator of natural and human-induced perturbations over soil profile and groundwater. Monitoring changes in soil CEC can assist in predicting whether soil quality has degraded, improved, or sustained under diverse agricultural or forestry schemes. Hybrid soft computing approaches involving evolutionary algorithms coupled with AI techniques facilitate the development of more sophisticated models with higher prediction accuracy. Hence, in the present study, a hybrid approach involving the multilayer perceptron neural network optimized with a genetic algorithm (GANN) was developed and employed to enhance the prediction efficiency of soil CEC in Tabriz plain, an arid region of Iran. In addition, a multiple model integration scheme supervised with hybrid GANN (MM-GANN) was also simulated and verified to improve the soil CEC prediction efficiency. This multiple model integration scheme supervised with the GANN approach is a unique form of a hybrid model for soil CEC prediction. Standalone MLP artificial neural network (ANN) and extreme learning machine (ELM) models were also implemented for incorporation into the multiple model integration scheme and for reasonable evaluation with MM-GANN model predictions.

2. Theoretical Overview

2.1. Artificial Neural Network (ANN). The multilayer perceptron (MLP), a class of feedforward ANN, is one of the most versatile algorithms that has proven able to simulate highly complex and nonlinear relationships between a set of input variables (predictors) and the output data (predictand) [44]. A multilayer perceptron (MLP) neural network with 1 hidden layer is shown in Figure 1. The network is trained to learn a function, $f(\cdot)$: $P^d \longrightarrow P^o$ on a set of training data, where "d" denotes the number of input dimensions and "o" denotes the number of output dimensions of the model [45]. The Levenberg-Marquardt backpropagation (BP) algorithm fine-tunes the weights and parameters of the MLP network. The network architecture involving the input layer consists of a set of processing units (neurons) $\{p_i | p_1, p_2, \ldots, p_n\}$ signifying the model input features and every hidden layer neuron performs a nonlinear transformation of the inputs from the previous layer via weighted linear summation of inputs $(w_1 p_1 + w_2 p_2 + \cdots + w_n p_n)$. A nonlinear activation function (σ) is then applied to each hidden unit to make a specific topology of weighted links more flexible following the affine transformation [46]. The neurons of the final layer receive connections from hidden layers of the network and are referred to as the output layer that produces a refined output. Some of the commonly used activation functions include hyperbolic tangent (tanh) and sigmoid (logsig) functions. There are no general rules for choosing training algorithms and adjusting associated parameters of the MLP architecture to maximize the efficiency of the network. A good introduction and mathematical concepts of ANN and its applications are provided in the following literature [47–51].

2.2. Extreme Learning Machine (ELM). The extreme learning machine (ELM) model proposed by Huang et al. [52] for a single layer feedforward network (SLFN) has been widely used for the prediction, forecasting, and estimation in many engineering fields [53–55]. Previous research studies have proved the outstanding advantages of the ELM model over the traditional AI techniques [56–58]. In addition, the ELM model can be implemented easily and has improved features such as fast learning speed [59], superior generalization performance [60], and utilization of activation functions (of nondifferentiable form) for training SLFN [52, 61]. Figure 2 portrays the general network structure of an ELM model. For N arbitrary distinct input samples $(X_i, Y_i) \in R^n \times R^n$, the standard SLFN with "L" hidden layer nodes can be described as follows:

$$\sum_{i=1}^{N} \beta_i g(X_i) = \sum_{i=1}^{N} \beta_i g(W_i \cdot X_i + c_i) = Y_k \quad k = 1, 2, 3, \ldots N, \tag{1}$$

where $c_i \in R$ is the assigned bias of the i^{th} hidden node, $w_i \in R$ is the assigned input weight connecting the i^{th} hidden and input layer nodes, β_i isthe weight connecting the i^{th} hidden and output layer nodes, and $g(X_i)$ is the output of the i^{th} hidden layer node with respect to the input X_i. Each input is assigned to the hidden nodes in the ELM model. The output weights can be derived by finding the least square solutions to the linear system. The main difference between the ELM model and traditional AI techniques is that the parameters of the feedforward network including its input weights and the hidden layer biases are randomly selected without any adjustment in the ELM model. For good introduction and mathematical concepts of ELM and its architectures, refer to Huang et al. [62], Martínez-Martínez et al. [63], Wang et al. [64], and Ding et al. [53].

2.3. Hybrid Genetic Algorithm-Neural Network (GANN). Genetic algorithm (GA) belongs to a class of search iterative approaches based on the "Darwinian" theory of natural selection and genetics that provide optimum solutions for combinatorial optimization, heuristic search, or process planning problems [65, 66]. GA implements genetic operators like reproduction, crossover, and mutation for upgradation and search for the best population by imitating the natural evolution process artificially. The genetic algorithm is initiated with individuals, an initial population of possible solutions, with a specified objective (fitness function) wherein every single individual is symbolized using a chromosome, a distinct form of encoding [67]. The chromosomes of a population are nominated for

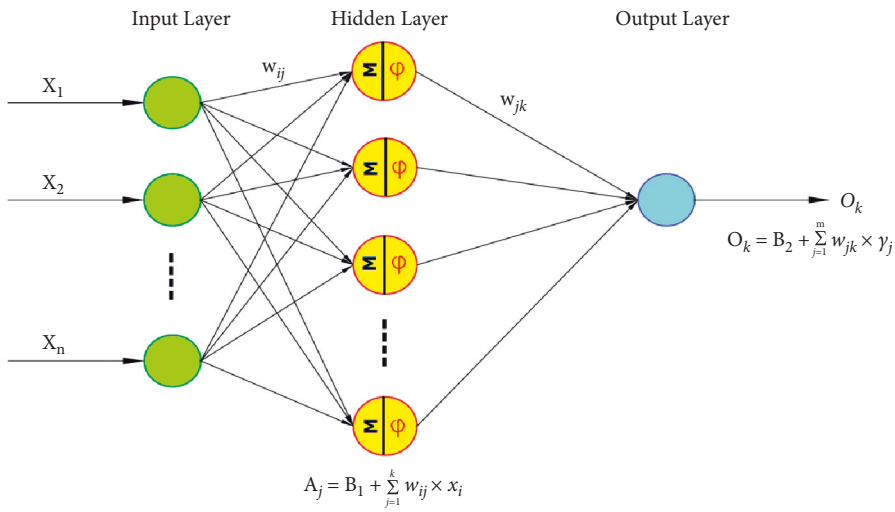


Figure 1: Architecture of MLP network.

reproduction based on the fitness value and the fittest individuals so selected are manipulated using crossover and mutation. The rudimentary idea here is the hope that superior parents can probabilistically produce superior offspring. The offspring of the next generation are generated by applying the GA operators crossover and mutation, upon the selected parents. The iteration process continues until the search converges to the termination criterion [65, 68]. The schematic illustration of the GA cycle is represented in Figure 3. The advantages of GA include (1) rapid convergence to the global optima, (2) superior multidirectional global search even in complex search surfaces, (3) use of probabilistic transition rules, and (4) the not deterministic ones in the search spaces where the gradient information is missing. The training of an MLP network, which is a type of neural network (NN), is somewhat a cyclic process. However, in the present case of the hybrid genetic algorithm-neural network (GANN), the intelligent search technique (GA) allows the user to configure the weight initialization range and the number of hidden layer neurons and update the weights and bias terms of an MLP network. Eventually, GA is used to learn the best hyperparameters for an MLP network. Even though the weights of the MLP network are initialized randomly, GA does not adhere to a simple random walk. Based on the parameter settings, it effectively exploits the information to gamble on fresh search points for anticipated improved performance [69]. GA selects the primary superlative solution with the best fitness values iteratively and recombines it with mutation and crossover operators to introduce offspring into the population. This process continues until the optimal solution with the highest fitness value is found based on any stopping criterion. Thus, the population's most fit MLP network is determined.

2.4. Multiple Model Integration Scheme Supervised with Hybrid GANN Model. The proposed multiple-model integration scheme involves the development of ANN and ELM models individually using input combinations as defined in their model structures. The discrete outputs (predicted series) of individual ANN and ELM models are then unified as inputs for the GANN model to obtain superior soil CEC predictions. The implementation of this multiple-model scheme involves two phases. In the first phase, the best-performing ANN and ELM models are identified by simulating all possible combinations of inputs. Later, in the second phase, the discrete outputs (predicted series) of the best ANN and ELM models are unified as inputs to simulate the GANN model. The GA optimizes the number of hidden layer neurons and updates the weights and bias terms of an ANN. The final output derived from this proposed scheme is referred to as integrated multiple models supervised with a hybrid GANN (MM-GANN) strategy (Figure 4).

3. Case Study and Data Description

The study area (Tabriz plain) considered encompasses an area of 150000 hectares (between 45°25′–46°12′ E, 37°50′–38°20′ N) and is located in the East Azerbaijan province of Iran. The surface topography of the area comprises rugged, mountainous rims, and the study area is sited toward the north-eastern part of Urmia Lake (Figure 5). Tabriz plain is a high-altitude location (1360 m above mean sea level) characterized by cooler, wetter winters and hot summers with a tropical and subtropical steppe climate. The study area never receives greater than 40 mm of rainfall in any of the months, and the annual mean precipitation is around 360.7 mm. The geology of the area includes recent alluvium, fine elastic sediments, and red conglomerate with an alternation of sandstone and red marl. The method of ammonium saturation as mentioned in Chapman [70] was used for the cation-exchange capacity determination. The descriptive statistics of soil CEC and other soil parameters of the study area under consideration are tabulated in Table 1. The spatial distribution of observed soil CEC is presented in Figure 6. The clay and soil organic matter were positively

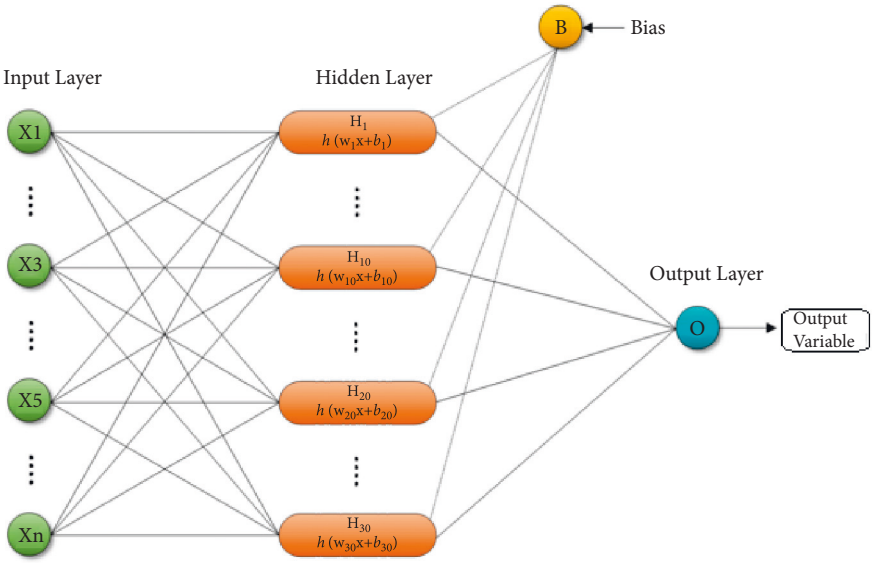


Figure 2: Architecture of ELM model.

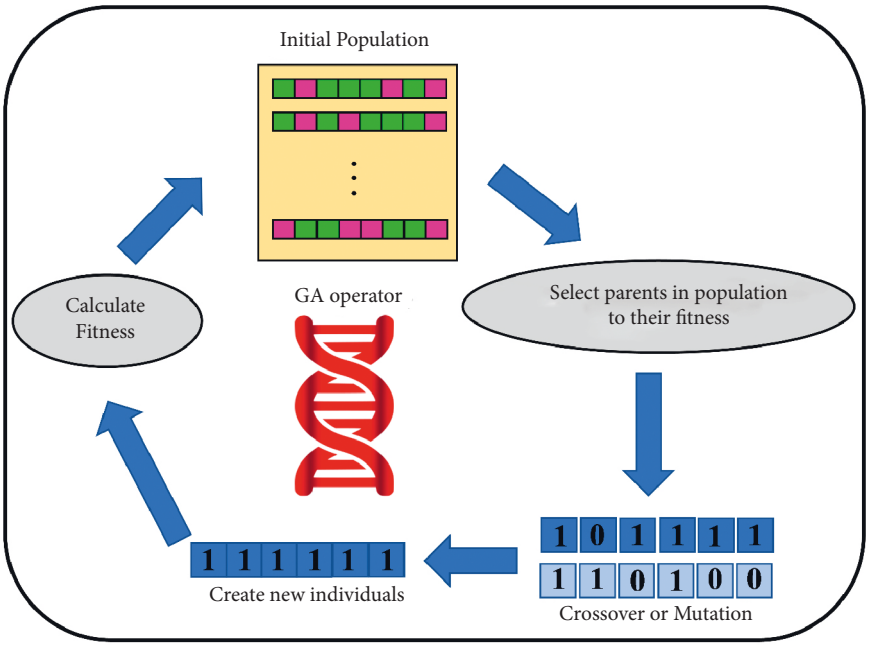


Figure 3: Schematic diagram of GA cycle.

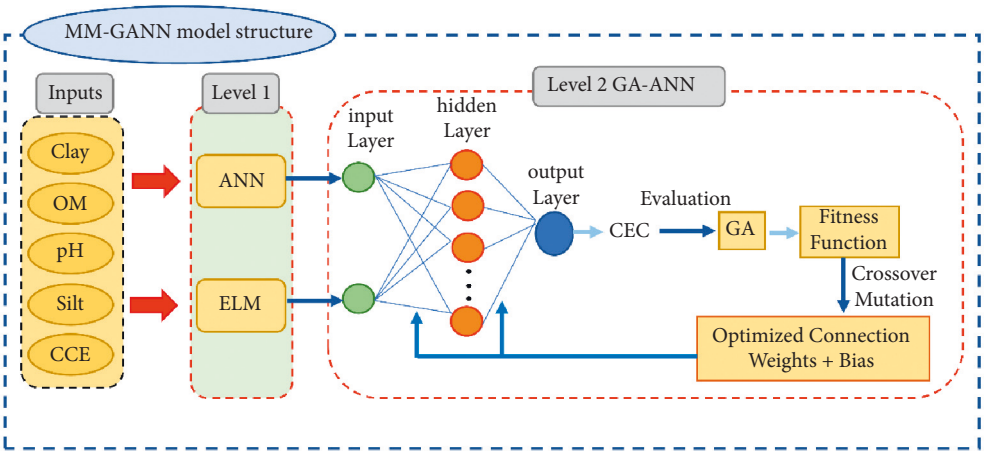


Figure 4: Structure of MM-GANN model.

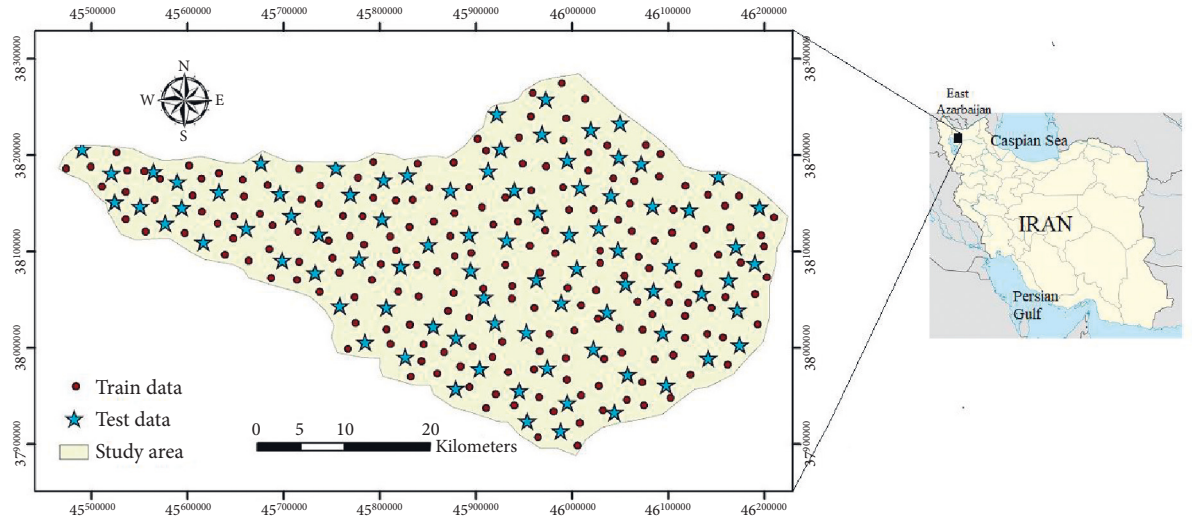


Figure 5: Location of the study area along with sampling points.

Table 1: Descriptive statistics of CEC and the soil parameters at the sampling points (training ($n = 195$) and testing ($n = 85$)).

Data set	Units	min		max		Mean		Standard deviation		Correlation with CEC
		Train	Test	Train	Test	Train	Test	Train	Test	
Clay	%	3.80	4.80	70.5	69.3	37.77	36.42	16.38	16.31	0.630
Silt	%	6.00	2.00	68.0	66.6	39.06	40.98	12.07	13.21	0.117
Sand	%	1.30	3.00	81.8	88.2	23.08	22.58	19.80	18.04	−0.605
OM	%	0.03	0.06	1.64	1.40	0.47	0.48	0.30	0.32	0.534
pH	—	7.00	6.90	8.70	8.20	7.65	7.60	0.35	0.30	0.284
CCE	%	0.55	1.60	26.55	45.0	13.02	13.82	5.00	6.38	0.220
CEC	cmol(+) kg^{-1}	6.20	5.90	45.00	42.0	21.89	21.38	9.23	8.05	—

CCE: carbonate calcium equivalent; OM: organic matter, CEC: cation-exchange capacity.

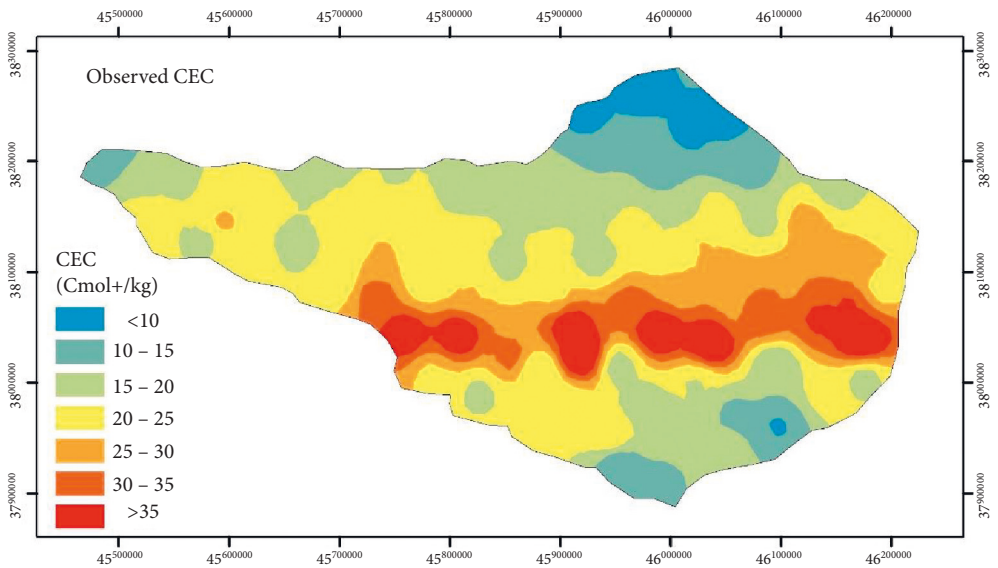


Figure 6: The spatial distribution map of observed soil CEC.

correlated, while the sand was found negatively correlated with soil CEC. The silt, pH, and carbonate calcium equivalent (CCE) parameters were not so significantly correlated with the soil CEC.

4. Modeling Development

Based on different combinations of soil parameters, the framework of model input-output scenarios was set for the development of ANN and ELM models with soil CEC as the output parameter. The input-output scenarios put on trial are listed in Table 2. The performance of the developed models was assessed based on the multiple statistical indices, namely, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash Sutcliffe Efficiency (NSE) [71], Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2).

Table 2: Input-Output structures for model development.

Models	Input combination	Output
Model 1	Clay	CEC
Model 2	Clay + pH	CEC
Model 3	Clay + OM	CEC
Model 4	Clay + OM + pH	CEC
Model 5	Clay + OM + pH + silt	CEC
Model 6	Clay + OM + pH + silt + CCE	CEC

$$\begin{aligned}
\text{Root Mean Square Error (RMSE)} &= \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{n}},\\
\text{Mean Absolute Error (MAE)} &= \frac{\sum_{i=1}^{n}\left|y_i - x_i\right|}{n},\\
\text{Nash Sutcliffe Efficiency (NSE)} &= 1 - \frac{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2},\\
\text{Mean Absolute Percentage Error (MAPE)} &= \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right|,\\
\text{coefficient of determination } R^2 &= \left(\frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2}}\right)^2,
\end{aligned} \tag{2}$$

where x_i is the actual value, y_i is the model estimated value, $\overline{x}$ is the mean of true values, $\overline{y}$ is the mean of the model estimated values, and n is the number of data points.

5. Results and Discussion

5.1. Performance of ANN and ELM Models. The ANN and ELM models were simulated for predicting the soil CEC based on the input-output combinations as mentioned in Table 2. The model structure (input nodes-hidden layer nodes-output nodes) and performance metrics of the ANN model for each input combination are presented in Table 3. In this study, the proposed ANN, ELM, and MM-GANN models were developed using MATLAB interface coding. The input-output scenario involving all the soil parameters (i.e., Clay + OM + pH + silt + CCE) provided the virtuous estimates of soil CEC with an NSE = 0.842. The input-output scenario involving four soil parameters (i.e., clay, OM, Ph, and silt) also offered relatively good soil CEC estimates with an NSE = 0.826. Despite having a significant correlation between clay and soil CEC, the single input-output ANN model (i.e., clay-CEC) failed to provide good soil CEC predictions. The spatial distribution map of ANN predicted soil CEC is presented in Figure 7. The ability of the MLP network to formulate a priori explicit hypotheses about a possible nonlinear relationship among several input variables makes it illustrious from other AI methods.

The performance metrics of ELM models for each input-output scenario are tabulated in Table 4. The scenario involving all the soil parameters (i.e., clay + OM + pH + silt + CCE) provided the virtuous predictions of soil CEC with an NSE = 0.835. The ELM model efficiency was slightly lesser than that of ANN. The ELM model simulated with four inputs (i.e., clay, OM, pH, and silt) had reasonably substandard performance when compared to that of the ANN model with a similar input structure. The spatial distribution map of ELM predicted soil CEC is shown in Figure 8. The scatter plots presented in Figure 9 of the three efficient models display the accounted linear relationship between the observed and estimated soil CEC by ANN and ELM models. According to Figure 9, the ELM outperformed ANN although they have very close performance in terms of the statistical indices (Tables 3 and 4). The ELM is known for its superior learning speed and virtuous generalization performance than the ANN architecture.

Table 3: The performance criteria of the ANN model for six input combinations.

Input combination	Output	Structure	Train					Test				
			MAE	RMSE	MAPE	NSE	R^2	MAE	RMSE	MAPE	NSE	R^2
Model 1	CEC	1-4-1	5.593	6.904	27.784	0.438	0.438	4.599	6.162	23.598	0.407	0.409
Model 2	CEC	2-6-1	3.913	5.186	20.008	0.683	0.683	4.425	6.016	23.007	0.435	0.442
Model 3	CEC	2-5-1	4.571	5.619	24.081	0.627	0.628	4.101	5.326	22.161	0.557	0.577
Model 4	CEC*	3-7-1	2.664	3.361	**13.911**	**0.866**	0.866	3.105	3.818	**16.504**	**0.787**	**0.787**
Model 5	CEC**	4-5-1	2.807	3.496	**15.701**	**0.855**	0.855	2.736	3.338	**15.043**	**0.826**	**0.827**
Model 6	CEC***	5-6-1	2.352	2.927	**12.492**	**0.899**	0.899	2.585	3.177	**13.695**	**0.842**	**0.843**

Note: the unit of MAE and RMSE is cmol (+) kg^{-1}.

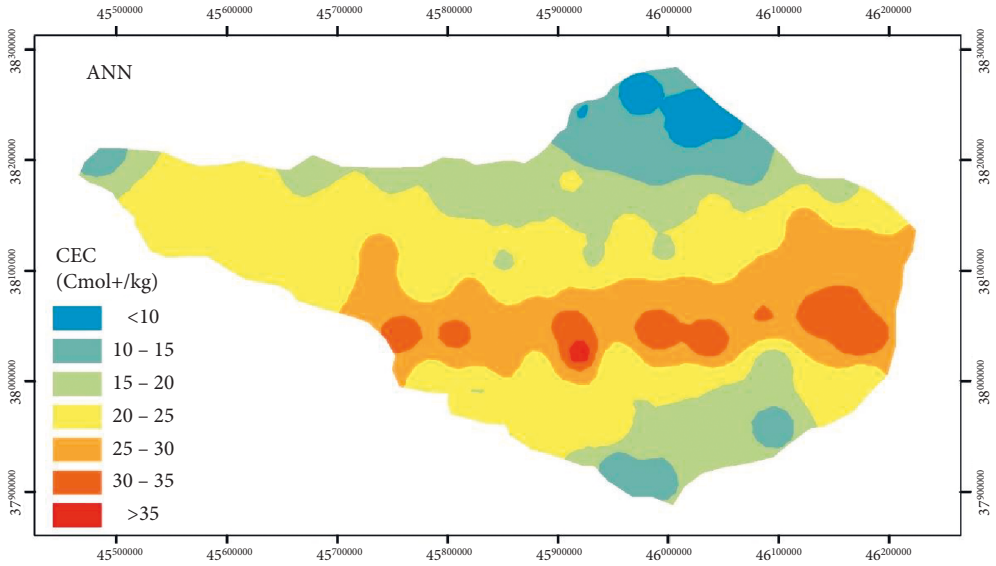


Figure 7: The spatial distribution map of ANN predicted soil CEC (ANN model with 5 inputs).

Table 4: The performance criteria of the ELM model for six input combinations.

Input combination	Output	Train					Test				
		MAE	RMSE	MAPE	NSE	R^2	MAE	RMSE	MAPE	NSE	R^2
Model 1	CEC	5.626	6.966	28.784	0.428	0.428	4.609	6.202	23.698	0.399	0.408
Model 2	CEC	3.965	5.310	20.108	0.667	0.668	4.744	6.384	23.307	0.364	0.432
Model 3	CEC	4.659	5.740	24.181	0.611	0.612	4.261	5.425	22.261	0.540	0.560
Model 4	CEC*	2.541	3.210	**13.711**	**0.878**	0.878	2.977	3.840	**16.554**	**0.769**	**0.804**
Model 5	CEC**	2.942	3.656	**15.801**	**0.842**	0.842	3.064	3.674	**16.043**	**0.789**	**0.808**
Model 6	CEC***	2.019	2.597	**10.844**	**0.920**	0.920	2.463	3.248	**13.640**	**0.835**	**0.858**

Note. The unit of MAE and RMSE is cmol (+) kg^{-1}.

5.2. Performance of MM-GANN Models. The soil CEC estimates of ANN and ELM models were employed as new inputs to the GANN model to predict soil CEC. To select the optimal input combinations in further modeling steps, examples from previous literature were referred to for enhancing the accuracy of models based on the different fields [72–75]. Within this category, it is worth mentioning that only the three highest performed combinations were considered in this hybrid model. The parameters of the genetic algorithm for adjusting the weights and bias terms of the ANN are presented in Table 5. Also, the performance statistics of MM-GANN models are shown in Table 6. The MM-GANN models involving the predictions of ELM and ANN models calibrated by considering all the soil parameters (i.e., clay, OM, pH, silt, and CCE) as inputs provided superior performance with an NSE = 0.87 in the test phase. The ANN parameters of the best MM-GANN model are charted in Table 7. The multiple model integration scheme imparts potential to the hybrid GANN model through the usage of standalone model outputs as inputs to the model. Thus, the hybrid GANN model exploits the previously learned information to improvise the predictive power of the model. The multiple model integration scheme is apparently enhancing the learning process of the hybrid model, where the output of the standalone models is used as relative informative

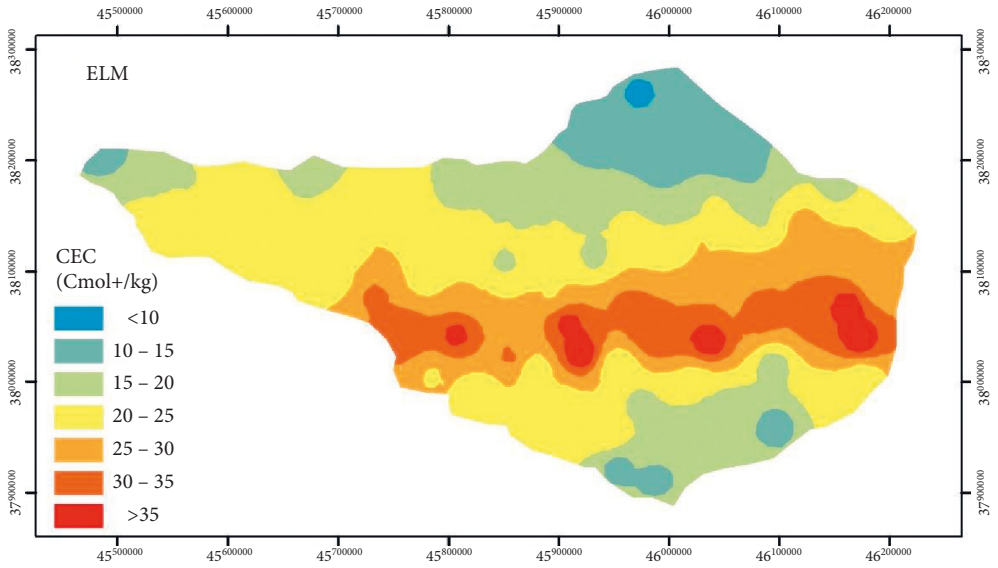


Figure 8: The spatial distribution map of ELM predicted soil CEC (ELM model with 5 inputs).

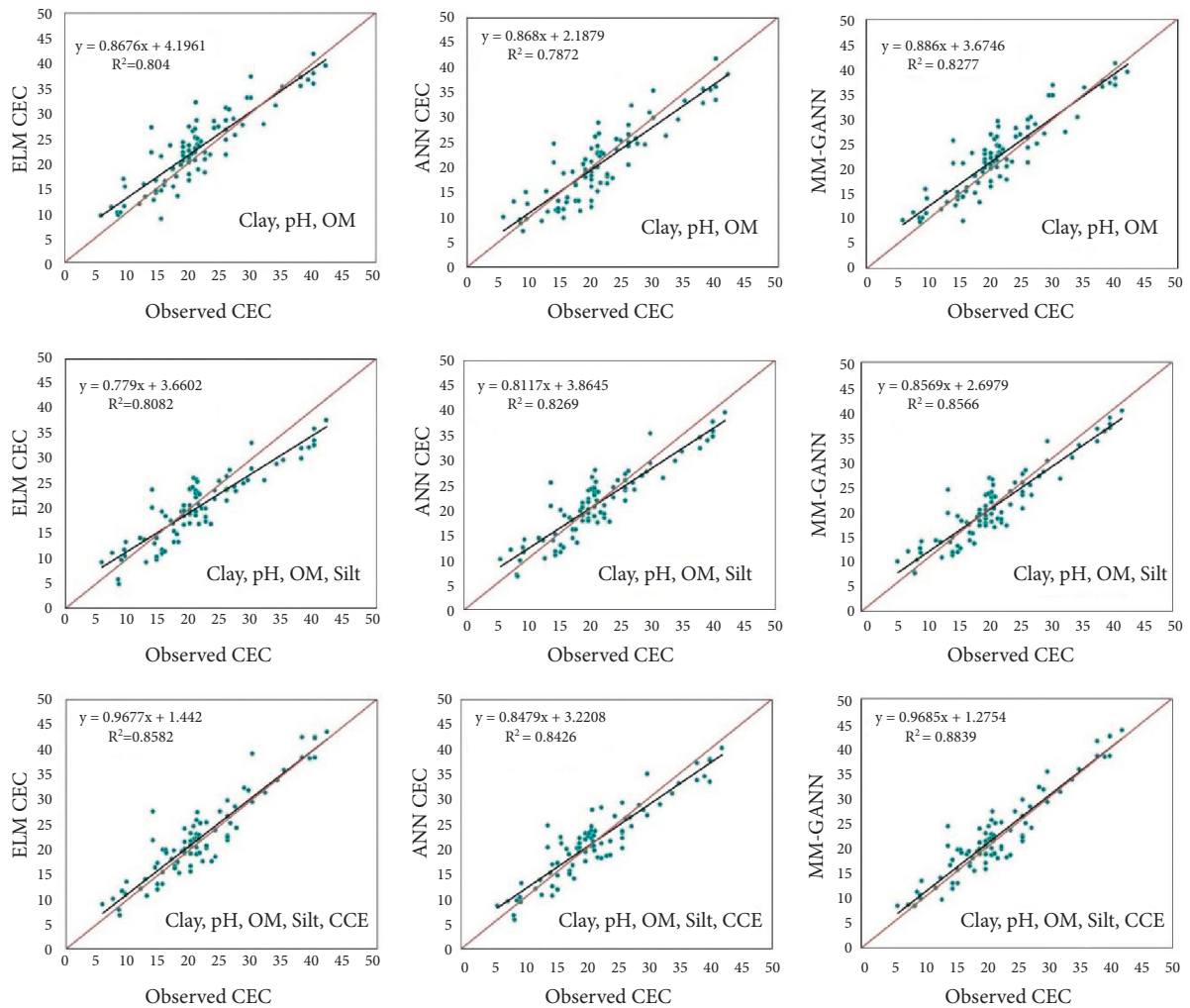


Figure 9: The observed and estimated soil CEC in the test stage: ELM, ANN, and MM-GANN models.

Table 5: Parameters of GA-Toolbox model.

Inputs	Outputs	Population	Generation	Crossover fraction	Mutation operator
ANN weights + bias	Best weights + bias	1000	500	0.7	Gaussian

Table 6: The performance of MM-GANN models for the three best input combinations.

Input combination	Output	ANN-structure	Train					Test				
			MAE	RMSE	MAPE	NSE	R^2	MAE	RMSE	MAPE	NSE	R^2
CEC*(ELM), CEC*(ANN)	CEC	2-5-1	2.465	3.107	**13.620**	**0.886**	0.878	2.816	3.583	**14.802**	**0.800**	**0.828**
CEC**(ELM), CEC**(ANN)	CEC	2-6-1	2.640	3.320	**13.893**	**0.870**	0.866	2.491	3.054	**13.020**	**0.854**	**0.857**
CEC***(ELM), CEC***(ANN)	CEC	2-5-1	1.942	2.486	**10.236**	**0.927**	0.929	2.249	2.885	**12.072**	**0.870**	**0.884**

Note: The unit of MAE and RMSE is cmol (+) kg^{-1}.

Table 7: ANN Parameters of the best MM-GANN model.

Inputs	Output	Structure	Hidden layer function	Output layer function	Hidden layer neurons	Training algorithm
CEC***(ELM), CEC***(ANN)	CEC	2-5-1	Tansig	Linear	5	Trainlm

predictors. The spatial distribution map of MM-GANN predicted soil CEC is presented in Figure 10 which is very much similar to that of the observed soil CEC map. The MM-GANN models developed with the predictions of ELM and ANN models calibrated by considering three and four soil parameters as inputs also offered convincingly good soil CEC predictions with NSE = 0.80 and 0.854, respectively. The scatter plots of MM-GANN models shown in Figure 9 depict the goodness of fit of the model predictions against the actual soil CEC values. In Figure 9, it is evident that the third combination of the MM-GANN model indicated a very close linearly fitted line to the 1 : 1 line, especially for the combination that had all the parameters.

Table 8 compares the performances of the best model of ANN, ELM, and MM-GANN models based on the statistical measures during the training and testing phases. This table shows that the performance accuracy of the hybrid model is higher than the ELM and ANN models, respectively, based on all the criteria values. The Taylor diagrams plotted for the best ANN, ELM, and MM-GANN models are shown in Figure 11. According to the Taylor diagram, it is very much evident that the multiple-model scheme (MM-GANN) offered relatively accurate estimates of soil CEC compared to the ELM and ANN models based on three statistical metrics (RMSD, standard deviation, and correlation coefficient). The MM-GANN model was the closest to the observed/actual data. The point density plots presented in Figure 12 also supported the above statement by exposing the tradeoff between observed soil CEC against the modeled.

5.3. Validation with Published Research Studies. Validating the results of current research with reliable published literature within the context of a similar kind of study area (i.e., semiarid region) is worthwhile. The correlation coefficient (R^2) indices were selected as an indicator of the prediction capability. The best R^2 obtained for MM-GANN, ELM, and ANN models is $R^2 \approx 0.88$, 0.85, and 0.84. In one of the earliest research performed on the soil CEC simulation along the Zayandehroud River in Isfahan, Iran, Amini et al. [27] established two classical ANN algorithms (i.e., feed-forward neural network and generalized regression neural network). The applied models were performed with poor prediction results with $R^2 \approx 0.69$ and 0.66. Another study was conducted by Emamgolizadeh et al. to predict soil CEC on collected soil information from Semnan, Mashahad, and Taybad cities of Iran [35]. The authors developed two new data intelligence models, namely, genetic expression programming (GEP) and multivariate adaptive regression spline (MARS). GEP and MARS models attained an $R^2 \approx 0.80$ and 0.86. Overall, the current study showed a convincing correlation performance over the state-of-the-art research studies.

Although the current research was the solitary approach to develop and assess the multiple model integration scheme supervised with hybrid GANN (MM-GANN), the certified limitations should be addressed in future research. It is evident from tables and figures that the MM-GANN model can improve the prediction accuracy of soil CEC when the inputs involving the predictions of ELM and ANN models calibrated by considering all the soil parameters (e.g., clay, OM, pH, silt, and CCE) are provided. However, one of the disadvantages of the MM-GANN model lies in the selection of the best standalone model for enhancing the prediction accuracy of soil CEC. Therefore, it is recommended to incorporate the prediction results of other data-driven models as the inputs of the MM-GANN model which can enhance the model's performance. In addition, this concept can be expanded and applied to other engineering fields such as structural, hydrologic, water resources, climatic, and different time series prediction/forecasting.

Table 8: Comparing performances of the best models of ANN, ELM, and MM-GANN.

Input combination	Output	Model structure	Train					Test				
			MAE	RMSE	MAPE	NSE	R^2	MAE	RMSE	MAPE	NSE	R^2
ANN model 6	CEC***	5-6-1	2.352	2.927	12.492	0.899	0.899	2.585	3.177	13.695	0.842	0.843
ELM model 6	CEC***		2.019	2.597	10.844	0.920	0.920	2.463	3.248	13.640	0.835	0.858
CEC***(ELM), CEC***(ANN)	CEC	2-5-1	**1.942**	**2.486**	**10.236**	**0.927**	**0.929**	**2.249**	**2.885**	**12.072**	**0.870**	**0.884**

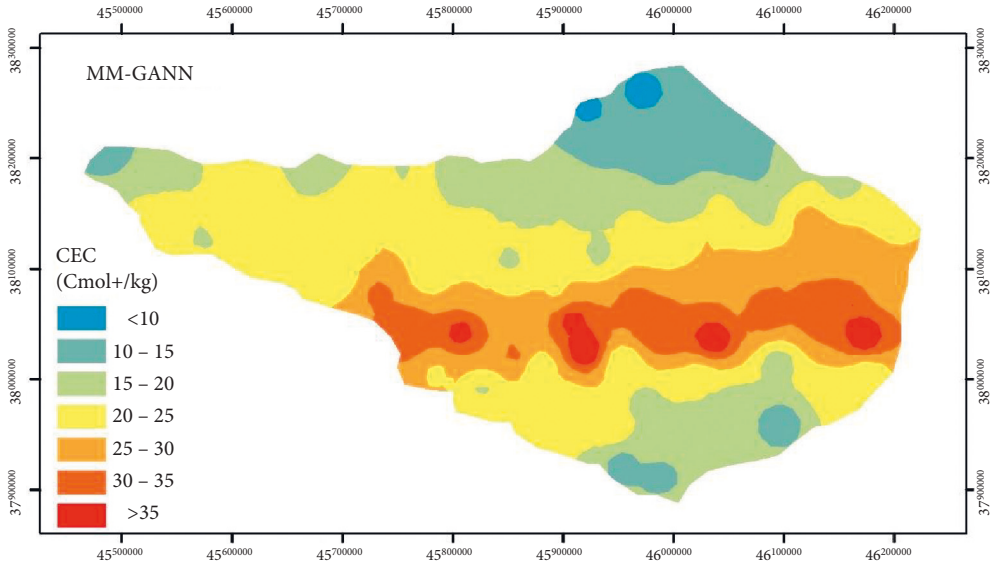


Figure 10: The spatial distribution map of MM-GANN predicted soil CEC (MM-GANN model calibrated with CEC***(ELM) and CEC***(ANN) as inputs).

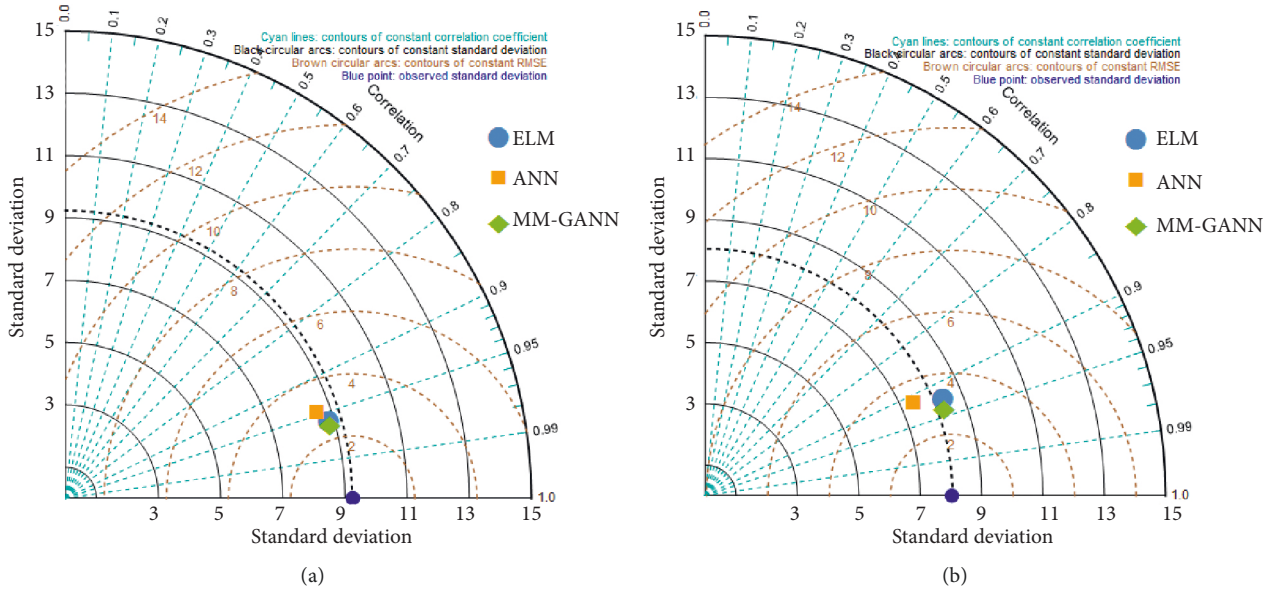


Figure 11: Taylor diagrams of the models at the training (a) and testing (b) phases.

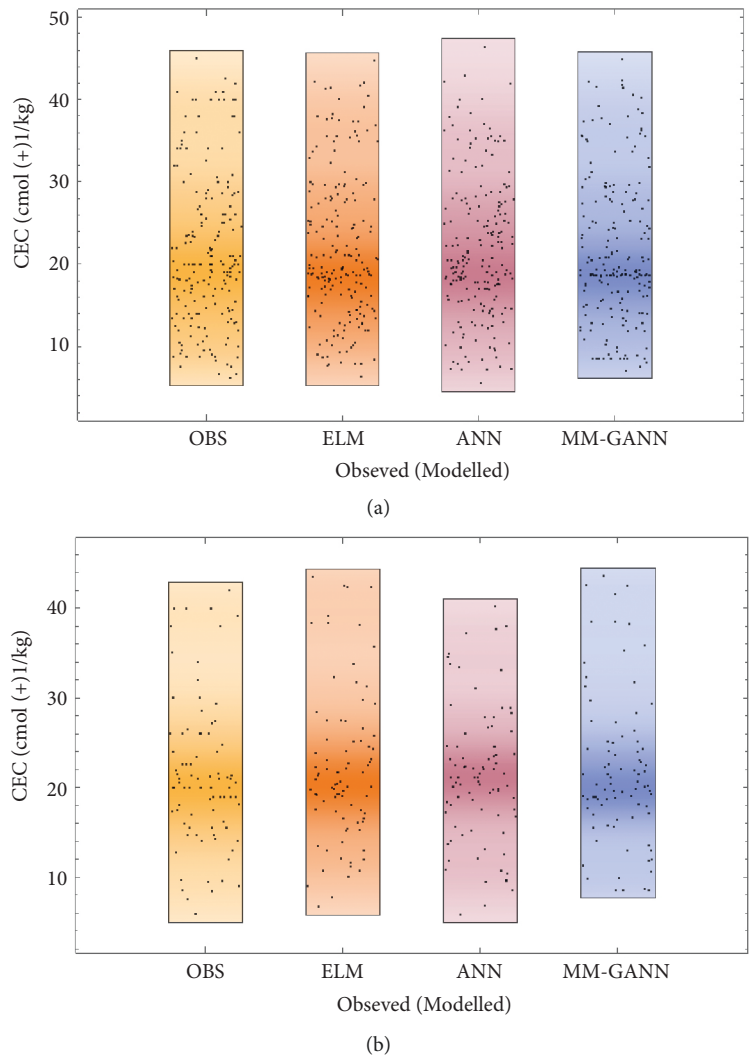


Figure 12: Point density plots of the models at the training (a) and testing (b) phases.

6. Conclusion

Over the past two decades, there is a noticeable demand for soil data assessment with regard to pollution and land degradation. The new era of soil process modeling using data intelligence models has been rapidly boosted. The current study was to develop a hybrid machine intelligence model based on the multimodel genetic algorithm-neural network for soil cation-exchange capacity. Two classical artificial intelligence models, namely, the ANN and ELM, were developed to evaluate their performance in estimating soil CEC along with the proposed hybrid MM-GANN model. Several correlated soil parameters including clay, silt, pH, carbonate calcium equivalent (CCE), and soil organic matter (OM) were used in the form of input attributes to the proposed and the comparable machine intelligence models. In particular, the hybrid MM-GANN model which received the predicted values of ANN and ELM as input attributes performed well in the estimation of soil CEC. Overall, the proposed multiple model integration scheme supervised with hybrid GANN model functions as an efficient pedotransfer function to predict or estimate soil CEC using readily available soil parameters (i.e., clay, OM, pH, silt, and CCE) as input variables. In particular, the conclusions of the current investigation are as follows:

(i) Based on the applied evaluation metrics, the ELM model provided superior CEC estimates than ANN.

(ii) The proposed hybrid MM-GANN model outperforms both standalone ANN and ELM models in terms of all the statistical metrics.

(iii) The proposed integrated hybrid machine intelligence scheme (MM-GANN) proved to be a reliable modeling strategy for modeling the soil cation-exchange capacity of the study area.

Before this end, it is worth stating the possibility for future research. As a fact, soil CEC is influenced by several morphological parameters [76, 77]; thus, integrating a feature selection as a prior modeling phase for the prediction process is highly recommended to be established. In addition, owing to the associated variability with each soil CEC type, it is an ideal proposition to estimate each type individually.

Data Availability

The datasets are available. Data can be shared upon request from the corresponding author.

Disclosure

This article has been previously published as preprint as follows: https://doi.org/10.21203/rs.3.rs-471256/v1 [78].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Wolf, "Cation exchange capacity and anion exchange," in *The fertile triangle :the Interrelationship of Air, Water, and Nutrients in Maximizing Soil Productivity*, pp. 177–188, Food Products Press, New York, NY, USA, 1999.

[2] E. Arthur, "Rapid estimation of cation exchange capacity from soil water content," *European Journal of Soil Science*, vol. 68, no. 3, pp. 365–373, 2017.

[3] Q. Ketterings, S. Reid, and R. Rao, *Cation Exchange Capacity*, Cornell University Cooperative Extension, Agronomy Fact Sheet #22, Ithaca, NY, USA, 2007.

[4] B. M. Tucker, "Displacement OF ammonium IONS for cation exchange capacity measurements," *Journal of Soil Science*, vol. 25, no. 3, 1974.

[5] D. Carroll, "Ion exchange in clays and other minerals," *GSA Bull*, vol. 70, no. 6, pp. 749–779, 1959.

[6] P. J. Thomas, J. C. Baker, and L. W. Zelazny, "An expansive soil index for predicting shrink–swell potential," *Soil Science Society of America Journal*, vol. 64, no. 1, pp. 268–274, 2000.

[7] M. Pansu and J. Gautheyrou, "Cation exchange capacity," in *Handbook of Soil Analysis*, pp. 709–754, Springer Berlin Heidelberg, Berlin, Germany, 2006.

[8] D. B. Mengel, *Fundamentals of Soil Cation Exchange Capacity (CEC)*, Purdue University Cooperative Extension Service, Agronomy Guide AY-238, West Lafayette, Indiana, 1993.

[9] J.-K. Böhlke, "Groundwater recharge and agricultural contamination," *Hydrogeology Journal*, vol. 10, no. 1, pp. 153–179, 2002.

[10] E. Arthur, M. Tuller, P. Moldrup, and L. W. Jonge, "Clay content and mineralogy, organic carbon and cation exchange capacity affect water vapour sorption hysteresis of soil," *European Journal of Soil Science*, vol. 71, no. 2, 2020.

[11] R. Dohrmann, "Cation exchange capacity methodology I: an efficient model for the detection of incorrect cation exchange capacity and exchangeable cation results," *Applied Clay Science*, vol. 34, no. 1–4, pp. 31–37, 2006.

[12] R. Dohrmann, "Cation exchange capacity methodology II: a modified silver–thiourea method," *Applied Clay Science*, vol. 34, no. 1–4, pp. 38–46, 2006.

[13] L. Delavernhe, M. Pilavtepe, and K. Emmerich, "Cation exchange capacity of natural and synthetic hectorite," *Applied Clay Science*, vol. 151, pp. 175–180, 2018.

[14] M. Conradie and W. A. G. Kotze, "Comparison of methods for estimating cation exchange capacity of orchard soils," *South African Journal of Plant and Soil*, vol. 6, no. 2, pp. 136-137, 1989.

[15] M. Khorshidi and N. Lu, "Determination of cation exchange capacity from soil water retention curve," *Journal of Engineering Mechanics*, vol. 143, no. 6, Article ID 04017023, 2017.

[16] K. Liao, S. Xu, and Q. Zhu, "Development of ensemble pedotransfer functions for cation exchange capacity of soils of Qingdao in China," *Soil Use & Management*, vol. 31, no. 4, pp. 483–490, 2015.

[17] L. A. Manrique, C. A. Jones, and P. T. Dyke, "Predicting cation-exchange capacity from soil physical and chemical properties," *Soil Sci. Soc. Am.*vol. 55, no. 3, pp. 787–794, 1991.

[18] S. E. Obalum, Y. Watanabe, C. A. Igwe, M. E. Obi, and T. Wakatsuki, "Improving on the prediction of cation exchange capacity for highly weathered and structurally contrasting tropical soils from their fine-earth fractions," *Communications in Soil Science and Plant Analysis*, vol. 44, no. 12, pp. 1831–1848, 2013.

[19] A. M. Lambooy, "Relationship between cation exchange capacity, clay content and water retention of Highveld soils," *South African Journal of Plant and Soil*, vol. 1, no. 2, pp. 33–38, 1984.

[20] R. L. Parfitt, D. J. Giltrap, and J. S. Whitton, "Contribution of organic matter and clay minerals to the cation exchange capacity of soils," *Communications in Soil Science and Plant Analysis*, vol. 26, no. 9–10, pp. 1343–1355, 1995.

[21] L. Krogh, H. Breuning-madsen, and M. H. Greve, "Cation-exchange capacity pedotransfer functions for Danish soils," *Acta Agriculturae Scandinavica Section B Soil and Plant Science*, vol. 50, no. 1, pp. 1–12, 2000.

[22] P. J. van Erp, V. J. G. Houba, and M. L. van Beusichem, "Actual Cation Exchange Capacity of Agricultural Soils and its relationship with pH and content of organic carbon and clay," *Communications in Soil Science and Plant Analysis*, vol. 32, no. 1–2, pp. 19–31, 2001.

[23] C. A. Seybold, R. B. Grossman, and T. G. Reinsch, "Predicting cation exchange capacity for soil survey using linear models," *Soil Science Society of America Journal*, vol. 69, no. 3, pp. 856–863, 2004.

[24] H. R. Fooladmand, "Estimating cation exchange capacity using soil textural data and soil organic matter content: a case study for the south of Iran," *Archives of Agronomy and Soil Science*, vol. 54, no. 4, pp. 381–386, 2008.

[25] H. Khodaverdiloo, H. Momtaz, and K. Liao, "Performance of soil cation exchange capacity pedotransfer function as affected by the inputs and database size," *Clean - Soil, Air, Water*, vol. 46, no. 3, Article ID 1700670, 2018.

[26] K. V. Looy, J. Bouma, and M. Herbst, J. Koestel, B. Minasny, U. Mishra, C. Montzka, A. Nemes, A. Y. Pachepsky, J. Padarian, M. G. Schaap, B. Tóth, A. Verhoef, J. Vanderborght, M. J. V. D. Ploeg, L. Weihermüller, S. Zacharias, Y. Zhang, H. Vereecken, Pedotransfer functions in earth system science: challenges and perspectives," *Reviews of Geophysics*, vol. 55, no. 4, pp. 1199–1256, 2017.

[27] M. Amini, K. C. Abbaspour, H. Khademi, N. Fathianpour, M. Afyuni, and R. Schulin, "Neural network models to predict cation exchange capacity in arid regions of Iran," *European Journal of Soil Science*, vol. 56, no. 4, pp. 551–559, 2005.

[28] H. Bayat, N. Davatgar, and M. Jalali, "Prediction of CEC using fractal parameters by artificial neural networks," *International Agrophysics*, vol. 28, no. 2, pp. 143–152, 2014.

[29] J. Seyedmohammadi, L. Esmaeelnejad, and H. Ramezanpour, "Determination of a suitable model for prediction of soil cation exchange capacity," *Model. Earth Syst. Environ.* vol. 2, no. 3, p. 156, 2016.

[30] L. Tang, G. Zeng, F. Nourbakhsh, and G. L. Shen, "Artificial neural network approach for predicting cation exchange capacity in soil based on physico-chemical properties," *Environmental Engineering Science*, vol. 26, no. 1, pp. 137–146, 2009.

[31] A. A. Zolfaghari, R. T. Mehrjardi, A. R. Moshki et al., "Using the nonparametric k-nearest neighbor approach for predicting cation exchange capacity," *Geoderma*, vol. 265, pp. 111–119, 2016.

[32] Y. K. Kalkhajeh, R. R. Arshad, H. Amerikhah, and M. Sami, "Comparison of multiple linear regressions and artificial intelligence-based modeling techniques for prediction the soil cation exchange capacity of Aridisols and Entisols in a semi-arid region," *Aust. J. Agric. Eng.*vol. 3, no. 2, p. 39, 2012.

[33] H. Ghorbani, H. Kashi, N. Hafezi Moghadas, and S. Emamgholizadeh, "Estimation of soil cation exchange capacity using multiple regression, artificial neural networks, and adaptive neuro-fuzzy inference system models in golestan province, Iran," *Communications in Soil Science and Plant Analysis*, vol. 46, no. 6, pp. 763–780, 2015.

[34] S. I. C. Akpa, S. U. Ugbaje, T. F. A. Bishop, and I. O. A. Odeh, "Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation," *Soil Use & Management*, vol. 32, no. 4, pp. 644–658, 2016.

[35] S. Emamgolizadeh, S. M. Bateni, D. Shahsavani, T. Ashrafi, and H. Ghorbani, "Estimation of soil cation exchange capacity using genetic expression programming (GEP) and multivariate adaptive regression splines (MARS)," *Journal of Hydrology*, vol. 529, pp. 1590–1600, 2015.

[36] A. A. Jafarzadeh, M. Pal, M. Servati, M. H. FazeliFard, and M. A. Ghorbani, "Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction," *International journal of Environmental Science and Technology*, vol. 13, no. 1, pp. 87–96, 2016.

[37] A. Keshavarzi, F. Sarmadian, J. Shiri, M. Iqbal, R. Tirado-Corbalá, and E.-S. E. Omran, "Application of ANFIS-based subtractive clustering algorithm in soil Cation Exchange Capacity estimation using soil and remotely sensed data," *Measurement*, vol. 95, pp. 173–180, 2017.

[38] K. Liao, S. Xu, J. Wu, Q. Zhu, and L. An, "Using support vector machines to predict cation exchange capacity of different soil horizons in Qingdao City, China," *Journal of Plant Nutrition and Soil Science*, vol. 177, no. 5, pp. 775–782, 2014.

[39] H. Shekofteh, F. Ramazani, and H. Shirani, "Optimal feature selection for predicting soil CEC: comparing the hybrid of ant colony organization algorithm and adaptive network-based fuzzy system with multiple linear regression," *Geoderma*, vol. 298, pp. 27–34, 2017.

[40] A. A. Ewees, A. A. Mohammed, A. Laith, and O. Diego, "Boosting arithmetic optimization algorithm with genetic algorithm operators for feature selection: case study on cox proportional hazards model," *Mathematics*, vol. 9, no. 18, p. 2321, 2021.

[41] O. Maciel, A. Valdivia, D. Oliva, E. Cuevas, D. Zaldívar, and M. Pérez-Cisneros, "A novel hybrid metaheuristic optimization method: hypercube natural aggregation algorithm," *Soft Computing*, vol. 24, pp. 8823–8856, 2020.

[42] H. Tao, A. A. Ewees, A. O. A. Sulttani et al., "Global solar radiation prediction over North Dakota using air temperature: development of novel hybrid intelligence model," *Energy Reports*, vol. 7, pp. 136–157, 2021.

[43] A. A. Bidgoli, H. E. Komleh, and S. J. Mousavirad, "Seminal Quality Prediction Using Optimized Artificial Neural Network with Genetic Algorithm," in *Proceedings ot*, Bursa, Turkey, 2016.

[44] K. Mehrotra, C. Mohan, and S. Ranka, *Elements of Artificial Neural Networks*, The MIT Press, Cambridge, MA, USA, 1996.

[45] S. K. Bhagat, K. Pyrgaki, S. Q. Salih et al., "Prediction of copper ions adsorption by attapulgite adsorbent using tuned-artificial intelligence model," *Chemosphere*, vol. 276, Article ID 130162, 2021.

[46] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 842 pages, Prentice-Hall, Hoboken, NY, USA, 1999.

[47] T. Kohonen, "An introduction to neural computing," *Neural Networks*, vol. 1, no. 1, pp. 3–16, 1988.

[48] G. Dreyfus, *Neural Networks*, Springer-Verlag, Berlin/Heidelberg, Germany, 2005.

[49] V. Kotu and B. Deshpande, *Data Science: Concepts and Practice*, Elsevier, Amsterdam, Netherlands, 2019.

[50] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, CRC Press, Boca Raton, FL, 2018.

[51] F. Cui, Z. A. Al-Sudani, G. S. Hassan, H. A. Afan, S. J. Ahammed, and Z. M. Yaseen, "Boosted artificial intelligence model using improved alpha-guided grey wolf optimizer for groundwater level prediction: comparative study and insight for federated learning technology," *Journal of Hydrology*, vol. 606, Article ID 127384, 2022.

[52] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine : a new learning scheme of feedforward neural networks," *IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 985–990, 2004.

[53] S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: algorithm, theory and applications," *Artificial Intelligence Review*, vol. 44, no. 1, pp. 103–115, 2015.

[54] S. Naganna, P. Deka, M. Ghorbani, S. Biazar, N. Al-Ansari, and Z. Yaseen, "Dew Point Temperature Estimation: Application of Artificial Intelligence Model Integrated with Nature-Inspired Optimization Algorithms," *Water*, vol. 11, no. 4, 2019.

[55] M. Leuenberger and M. Kanevski, "Extreme Learning Machines for spatial environmental data," *Computers & Geosciences*, vol. 85, pp. 64–73, 2015.

[56] B. Deng, X. Zhang, W. Gong, and D. Shang, "An overview of extreme learning machine," in *Proceedings of thin 2019 4th international conference on control, robotics and cybernetics (crc)*, pp. 189–195, Tokyo, Japan, September 2019.

[57] J. X. Chen, "Applications of extreme learning machines," *Computing in Science & Engineering*, vol. 21, no. 5, pp. 4-5, 2019.

[58] J. Alavi, A. A. Ewees, S. Ansari, S. Shahid, and Z. M. Yaseen, "A new insight for real-time wastewater quality prediction using hybridized kernel-based extreme learning machines with advanced optimization algorithms," *Environmental Science & Pollution Research*, vol. 29, pp. 20496–20516, 2022.

[59] H. T. Huynh, Y. Won, and J.-J. Kim, "An improvement of extreme learning machine for compact single-hidden-layer feedforward neural networks," *International Journal of Neural Systems*, vol. 18, no. 05, pp. 433–441, 2008.

[60] A. Akusok, *Extreme Learning Machines: Novel Extensions and Application to Big Data*, University of Iowa, Iowa City, IA, USA, 2016.

[61] Z. M. Yaseen, M. Ali, A. Sharafati, N. Al-Ansari, and S. Shahid, "Forecasting standardized precipitation index using

data intelligence models: regional investigation of Bangladesh," *Scientific Reports*, vol. 11, no. 1, 2021.
[62] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *Int. J. Mach. Learn. Cybern.*vol. 2, no. 2, pp. 107–122, 2011.
[63] J. M. Martínez-Martínez, P. Escandell-Montero, E. Soria-Olivas, J. D. Martín-Guerrero, R. Magdalena-Benedito, and J. Gómez-Sanchis, "Regularized extreme learning machine for regression problems," *Neurocomputing*, vol. 74, no. 17, pp. 3716–3721, 2011.
[64] Y. Wang, F. Cao, and Y. Yuan, "A study on effectiveness of extreme learning machine," *Neurocomputing*, vol. 74, no. 16, pp. 2483–2490, 2011.
[65] D. E. G. A. J. H. Holland, *GUEST EDITORIAL Genetic Algorithms and Machine Learning*, pp. 95–99, Springer, New York, NY, USA, 1988.
[66] J. H. Holland, "Genetic algorithms," *Scientific American*, vol. 267, no. 1, pp. 66–72, 1992.
[67] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," *Foundations of genetic algorithms*, vol. 1, pp. 69–93, 1991.
[68] A. H. Wright, "Genetic algorithms for real parameter optimization," in *Foundations of genetic algorithms*vol. 1, , pp. 205–218, Elsevier, 1991.
[69] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-We, Boston, MA, 1989.
[70] H. D. Chapman, "Cation-exchange capacity," *Methods soil Anal. Part 2 Chem. Microbiol. Prop.*vol. 9, pp. 891–901, 1965.
[71] Z. M. Yaseen, "An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions," *Chemosphere*, vol. 277, Article ID 130126, 2021.
[72] S. Kim and H. S. Kim, "Neural Networks and Genetic Algorithm Approach for Nonlinear Evaporation and Evapotranspiration Modeling," *Journal of Hydrology*, vol. 351, pp. 299–317, 2008.
[73] S. Kim, S. Seo, M. Rezaie-Balf, O. Kisi, M. A. Ghorbani, and V. P. Singh, "Evaluation of Daily Solar Radiation Flux using Soft Computing Approaches based on Different Meteorological Information: Peninsula vs Continent," *Theoretical and Applied Climatology*, vol. 137, pp. 693–712, 2019.
[74] S. Kim and V. P. Singh, "Modeling Daily Soil Temperature using Data-Driven models and Spatial Distribution," *Theoretical and Applied Climatology*, vol. 118, pp. 465–479, 2014.
[75] S. Kim, V. P. Singh, C. J. Lee, and Y. Seo, "Modeling the Physical Dynamics of Daily Dew Point Temperature using Soft Computing Techniques," *KSCE Journal of Civil Engineering*, vol. 19, pp. 1930–1940, 2015.
[76] A. Sharma, D. C. Weindorf, D. D. Wang, and S. Chakraborty, "Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC)," *Geoderma*, vol. 239, pp. 130–134, 2015.
[77] K. H. Tan and P. S. Dowling, "Effect of organic matter on CEC due to permanent and variable charges in selected temperate region soils," *Geoderma*, vol. 32, no. 2, pp. 89–101, 1984.
[78] M. Shahabi, M. I. Ghorbani, S. R. Naganna et al., "Modeling Soil Cation Exchange Capacity in Arid Region of Iran," *Application of Novel Hybrid Intelligence Paradigm*, 2021.

Hindawi
Complexity
Volume 2022, Article ID 6312579, 7 pages
https://doi.org/10.1155/2022/6312579

WILEY | Hindawi

Research Article

An Aggregating Prediction Model for Management Decision Analysis

Jianhong Guo,[1,2] **Che-Jung Chang**,[1,2] **Yingyi Huang**,[1,2] **and Xiaotian Zhang**[1]

[1]*TSL Business School, Quanzhou Normal University, No. 398, Donghai Street, Quanzhou, Fujian 362000, China*
[2]*Fujian University Engineering Research Center of Cloud Computing, Internet of Things and E-Commerce Intelligence, No. 398, Donghai Street, Quanzhou, Fujian 362000, China*

Correspondence should be addressed to Che-Jung Chang; r3795102@nckualumni.org.tw

Received 3 March 2022; Revised 26 April 2022; Accepted 28 April 2022; Published 23 May 2022

Academic Editor: Andrea Murari

Facing an increasingly competitive market, enterprises need correct decisions to solve operational problems in a timely manner to maintain their competitive advantages. In this context, insufficient information may lead to an overfitting phenomenon in general mathematical modeling methods, making it difficult to ensure good analytical performance. Therefore, it is important for enterprises to be able to effectively analyze and make predictions using small data sets. Although various approaches have been developed to solve the problem of prediction, their application is often limited by insufficient observations. To further enforce the effectiveness of data uncertainty processing, this study proposed an aggregating prediction model for management decision analysis using small data sets. Compared with six popular approaches, the results from the experiments show that the proposed method can effectively deal with the small data set prediction problem and is thus an appropriate decision analysis tool for managers.

1. Introduction

Decision analysis is one of the most important tasks for managers [1]. Effective decision-making helps managers solve operational problems in a timely manner, which is vital to remain viable in an increasingly competitive market. Uncontrollable factors and uncertain events may lead to invalid decision-making and affect business performance. Predictive analysis can help managers grasp possible future trends and reduce the impact of uncertainty on personal judgments [2], thereby assisting enterprises to make better decisions for their business.

Decisions that require immediate responses can be difficult for managers. In order to process information and effectively run an operation, managers must grasp the situation in real time through a limited number of observations [3]. An example of this is analyzing the occurrence of a new disease. If the government can make the right decisions as soon as possible, the potential harm and impact of new diseases on people's health will be reduced. Decisions should be made in a timely manner to prevent infected people from spreading the disease. A prompt response thereby adds the management value. Building prediction models using small data sets, therefore, has a significant practical value.

Popular prediction approaches are roughly divided into three categories: time series analysis, relational models, and data mining techniques [4]. Time series analysis considers the continuous trend of data, which only needs historical data to predict the future demand [5]. It has been widely adopted to solve various prediction problems; however, it typically requires a large number of observations to obtain better forecasting performance. The relational models are used to explore the causal connection between the independent variables and the dependent variables to predict the possible outputs of the dependent variables [6]. The accuracy of the prediction depends on whether the selected independent variables can properly explain

the dependent variables. Data mining techniques use algorithms to polish useful hidden information from the collected data [7]. These techniques can obtain favorable predictions through an effective learning process; however, the prediction results depend on the amount of training data and the representativeness of the training set in the population [8].

The approaches discussed above vary in their practical applications. Therefore, before building a model, data analysis must be performed to determine which approach is suitable for the collected data [9]. The modeling pretest requires a sufficient number of samples; otherwise, the evaluation may fail. This limitation makes these approaches unsuitable for prediction using small data sets [10]. One positive example is the prediction problem on the electronic commerce (e-commerce) transaction volumes in China. To plan an appropriate development strategy, it must be drafted based on new information [11]. Using observations with updated information to build a model could reflect the true situation [12]; therefore, it is valuable to use a limited number of updated observations to make predictions [13].

This research proposed a modeling procedure based on a grey system theory for developing an aggregating prediction model that combines various approaches instead of determining the most suitable prediction model. The perspective of model integration is used to solve the small data set prediction problem, with the aim to improve the stability of the prediction results by combining the advantages of various approaches. Specifically, the proposed method is designed as a two-stage modeling procedure. First, four methods are assessed through grey incidence analysis to determine whether the trend of the real series can be reflected. Second, a robust compound prediction model for small samples is built by the weighted average method. In addition, a pretest is performed to evaluate the feasibility of the proposed method before trend forecasting. Results from these experiments show that the proposed method produces favorable predictions under small data sets to solve the encountered problem. Because the proposed method reduces the decision risk, it is considered a practical tool for small data set prediction.

The remainder of this article is organized as follows: Section 2 describes the proposed method. Section 3 presents the data analysis and comparison among the various approaches. Finally, the conclusion is discussed in Section 4.

2. Methodology

This study aims at solving the prediction problems when the available samples are limited. Although popular prediction approaches (such as statistical methods, data mining, and artificial neural networks) have acceptable performance in normal applications [14–16], they are not directly applicable to small data set analyses due to limited information. Therefore, this study proposed a modeling procedure to integrate the advantages of various popular approaches for this specific problem. This section details the concepts and steps of the proposed method.

2.1. Conceptual Design. Popular forecasting approaches usually have their own limitations and scope of applications [7, 17]. Therefore, it is necessary to conduct a pretest through data analysis to determine which approach is more appropriate before formal trend forecasting. A robust forecasting technique is very important for effectively grasping future trends [18]. For this reason, this research proposed a relatively robust compound prediction model based on grey system theory, which is called the grey-based aggregating model (GAM).

The proposed method accumulates the advantages of various approaches by applying the viewpoint of compounding models, thereby improving the prediction performance. The proposed method is a two-stage modeling procedure, i.e., four popular methods are used to obtain the basic predicted values, and then, the proposed method is used to obtained grey-based weights to identify the final predicted values.

2.2. Basis of Aggregating Model. Four fundamental prediction techniques are selected as the basis of the proposed model here, namely, grey model (GM), linear regression (LR), backpropagation neural network (BPNN), and support vector regression (SVR). GM is an important technique for managing insufficient information, which is easy to implement and can bring accurate predictions under small data sets [9]. LR is a commonly used numerical prediction method due to its implementation being not complicated, and it can produce good results when data follow a linear trend [17]. BPNN is widely used in nonlinear data analysis, which is a modeling method with excellent learning ability [7]. SVR is a statistical learning method that can overcome the difficulty of nonparametric prediction with limited samples [7].

Because the above methods have their own specific conditions and the components of demand are inherently complex, it is difficult to determine which method is most suitable for demand forecasting. Therefore, this research does not recommend using a single model for analysis. Instead, the study tries to retain the advantages of the above four models to form a new compound model.

The aggregating model aims at achieving relatively robust trend prediction. Therefore, the risk of choosing the wrong model will be controlled, thereby helping managers to better cope with uncertainty. In summary, this paper proposes GAM to solve the problems encountered in demand forecasting.

2.3. Grey Incidence Analysis. The degree of grey incidence (DGI) is a pretesting measurement index commonly used in grey system theory. DGI is a technique for evaluating the fitness of a prediction model regardless of whether the sample size is large or small [19]. The basic idea of DGI is to use the geometrical similarity of the series curves to determine the relationship between two series. The more similar the curves , the greater incidence exists between the series and vice versa [20].

Theoretically, residual analysis must be applied after building a model to investigate whether the model performs acceptably [21, 22]. However, an error-based objective function is usually selected to optimize the fitting status of a developed model in the learning phase of the modeling process. If similar error indicators are repeatedly used to analyze the data, it would lead to specific deviations and nonobjective results in the pretesting stage. To avoid this phenomenon, it is a feasible option to adopt the DGI to evaluate the fitting status during the modeling process instead of using any error-based index for pretesting.

There are many kinds of general DGI. Although their development principles are different, they can effectively measure the geometric similarity between the two series. To facilitate the calculation and application, this paper adopted the similitude degree of grey incidence (SDGI) [9] as the measurement index for evaluating different prediction methods. The SDGI is suitable for evaluating the fitting status of a given model to real data, which calculates the relational similarity between two series based on their geometrical similarity. A higher SDGI value indicates that the two series are geometrically similar, and the variance of the prediction error is stable. This implies that such a method is more robust and should be given a higher weight. Therefore, the order of SDGI values is used to determine the weights of different prediction methods. The detailed steps of SDGI are described as follows:

Step 0: give two paired series with n periods $X_0 = \{x_0(1), x_0(2), \ldots, x_0(n)\}$ and $X_i = \{x_i(1), x_i(2), \ldots, x_i(n)\}$.

Step 1: use equation (1) to perform the zero-starting point operator to form two new series, $X_0^0 = \{x_0^0(1), x_0^0(2), \ldots, x_0^0(n)\}$ and $X_i^0 = \{x_i^0(1), x_i^0(2), \ldots, x_i^0(n)\}$.

$$x_l^0(k) = x_l(k) - x_l(1) k = 1, 2, ..., n. \quad (1)$$

Step 2: use equation (2) to subtract X_0^0 from X_i^0 to obtain a difference series $X_{0i}^0 = \{x_{0i}^0(1), x_{0i}^0(2), \ldots, x_{0i}^0(n)\}$.

$$x_{0i}^0(k) = x_0^0(k) - x_i^0(k) k = 1, 2, ..., n. \quad (2)$$

Step 3: use equation (3) to sum the area of the difference series.

$$s_{0i} = \frac{1}{2} \sum_{k=2}^{n} \left[x_{0i}^0(k-1) + x_{0i}^0(k) \right]. \quad (3)$$

Step 4: use equation (4) to calculate the SDGI ε_{0i}.

$$\varepsilon_{0i} = \frac{1}{1 + s_{0i}}. \quad (4)$$

2.4. Rank-Sum Weighting Method. To obtain the final aggregating model, different forecasting models need to be combined and the most common way is to achieve it by the weighted average method. How to determine a reasonable weight for each method is something that must be overcome at this stage. To facilitate the application of the proposed method, this paper used a heuristic weighting method to set the importance of each method. The adopted method is the rank-sum weighting method (RSWM), and its calculation formula is equation (5), where m is the total number of prediction methods and R_j is the rank of the methods. The method ranked first has the highest reference value, and its numerator is exactly equal to the total number of methods m; conversely, the method ranked last has the lowest reference value, and its numerator is exactly equal to 1.

$$w_j = \frac{m + 1 - R_j}{\sum_{j=1}^{m} R_j}. \quad (5)$$

2.5. Modeling Procedure of the Proposed GAM. The GAM is applied to combine the final prediction outputs for grasping the development trend of China's e-commerce. The detailed processes of GAM are as follows:

Step 0: give an initial series with n period $X_0 = \{x_0(1), x_0(2), \ldots, x_0(n)\}$.

Step 1: apply X_0 as the training samples to establish forecasting models based on BPNN, GM, LR, and SVR and then obtain the fitted series of each established model; they are $X_{GM} = \{x_{GM}(1), x_{GM}(2), \ldots, x_{GM}(n)\}$, $X_{LR} = \{x_{LR}(1), x_{LR}(2), \ldots, x_{LR}(n)\}$, $X_{BPNN} = \{x_{BPNN}(1), x_{BPNN}(2), \ldots, x_{BPNN}(n)\}$, and $X_{SVR} = \{x_{SVR}(1), x_{SVR}(2), \ldots, x_{SVR}(n)\}$

Step 2: calculate the SDGI (Section 2.2) for each paired series between the fitted values and initial values to obtain ε_{GM}, ε_{LR}, ε_{BPNN}, and ε_{SVR}

Step 3: sort the SDGI of each method from largest to smallest and get the ranking value of each method; they are R_{GM}, R_{LR}, R_{BPNN}, and R_{SVR}

Step 4: use RSWM to determine the weights of each method and obtain w_{GM}, w_{LR}, w_{BPNN}, and w_{SVR} as the final model weights

Step 5: apply the weighted average method to aggregate the final forecast value

$$\hat{x} = w_{GM}\hat{x}_{GM} + w_{LR}\hat{x}_{LR} + w_{BPNN}\hat{x}_{BPNN} + w_{SVR}\hat{x}_{SVR}. \quad (6)$$

2.6. Feasibility Measurement. An effective prediction model must bring accurate forecasting results. It is therefore necessary to evaluate prediction methods using an error-based index, and only methods that pass inspection can be used for future trend forecasting. In this study, the mean absolute percentage error (MAPE) is selected to determine the modeling performance in the pretesting phase. The MAPE can assist managers in assessing the possible risks of using different forecasting tools. Equation. (7) is the calculation formula of the MAPE, where y_i and $\hat{y}_i$ are the predicted and actual values, respectively.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \times 100. \quad (7)$$

In the pretesting stage, the prediction results of the proposed GAM are compared with those of six popular prediction techniques to confirm whether the proposed method can provide more robust prediction results. The machine learning software used here is Weka 3.6.11, and the models are built with default parameter settings. The GM chooses the typical first-order one-variable grey model.

2.7. Rolling Framework. Time series forecasting emphasizes the immediateness of information, and the rolling framework is a process that allows data to be metabolized for achieving this purpose. For example, four given pieces of data $\{x_0(1), x_0(2), x_0(3), x_0(4)\}$ are used to predict the next value $\widehat{x}_0(5)$ with the prediction techniques. After the prediction is acquired, the newly predicted output is added to the data set to replace the oldest datum $x_0(1)$. Subsequently, the updated data set $\{x_0(2), x_0(3), x_0(4), \widehat{x}_0(5)\}$ is used to obtain the next predicted value $\widehat{x}^{(0)}(6)$. The process is repeated until all desired predicted values are found.

3. Experimental Results

The effectiveness and applicability of the proposed GAM are validated using a real case in the following sections.

3.1. Data Description and Experimental Design. This experiment verifies the effectiveness of the proposed GAM in processing China's e-commerce transaction volume forecast. The study included data collected from the National Bureau of Statistics of China on the total amount of e-commerce transactions. The data set contains ten-period annual observations ranging from 2011 to 2020 (Table 1). The unit of measurement in this table is trillion Chinese Yuan.

E-commerce is one of the current main business modes based on network communication technology, which is an important driving force for the integration and development of the physical transactions and the digital economy [23]. Under the impact of coronavirus disease 2019 (COVID-19), e-commerce has quickly become an indispensable part of people's lives [24]. E-commerce allows for economic and commercial activities with reduced interpersonal contact leading to the decrease in transmission of viruses [25]. In the postpandemic era, the causal relationship between the development of e-commerce and economic growth is obvious.

To maintain the momentum of economic development, the support of the e-commerce operating environment is necessary [26]. The formulation of development policy is therefore critical. It not only guides the operation of industries but also affects the consumption habits of people [27]. The work of creating an e-commerce operating environment (for example, formulating laws and regulations, determining industry standards, cultivating talents, and building logistics facilities) usually requires long-term efforts to achieve an acceptable result [28], and improper policy directions could bring substantially negative effects [29]. An adequate e-commerce transaction volume prediction is a prerequisite for formulating an effective development policy as it reduces the possibility of errors in policy planning [2]. Therefore, an accurate prediction for e-commerce transaction volumes has important practical significance for governments [26]. Effectively determining the trend of e-commerce transaction volumes helps the government draft an industrial development strategy, which is crucial for economic recovery after COVID-19. In order to reflect the current situation, appropriate e-commerce development policies should be based on updated and relevant information.

In the experiment, four data points are used each time to build a model for predicting the next output. That is, 2015's predicted value is inferred from the model built based on the data from 2011 to 2014. In the pretesting stage, the four techniques mentioned above are first used to obtain a total of 24 models and 24 predicted values. Next, these predicted values are used to determine the weights required to generate the aggregating model. The final predicted value is obtained by the weighted average method.

3.2. Modeling Example of the Proposed GAM. This section explains the modeling details of the proposed GAM. First, the four prediction techniques of GM, LR, BPNN, and SVR built six models and obtained six corresponding predicted values (columns 3 to 6 of Table 2). Second, the SDGI between the actual value series and each predicted value series was calculated, and $\varepsilon_{GM} = 0.1355$, $\varepsilon_{LR} = 0.0473$, $\varepsilon_{BPNN} = 0.2830$, and $\varepsilon_{SVR} = 0.0299$ could be obtained. Third, the rank of each prediction technique was determined according to the SDGI, that is, $R_{GM} = 2$, $R_{LR} = 3$, $R_{BPNN} = 1$, and $R_{SVR} = 4$. Fourth, the weight of each method was obtained by the RSWM, which are $w_{GM} = 0.3$, $w_{LR} = 0.2$, $w_{BPNN} = 0.4$, and $w_{SVR} = 0.1$, respectively. Last, through the weighted average method, the final GAM was combined as $GAM = 0.3 \times GM + 0.2 \times LR + 0.4 \times BPNN + 0.1 \times SVR$, and the corresponding predicted values were obtained by this final model (column 7 of Table 2).

3.3. Comparisons. In the pretest stage, the prediction results of GAM were compared with those obtained using the six popular methods. These methods are GM, LR, BPNN, SVR, radial basis function network (RBFN), and Gaussian process regression (GPR). The MAPE of the proposed GAM is 2.84% (Table 3), which is the best performing one of these methods. Its value is less than 5% and falls within the level of highly accurate forecasting (Table 4 [30]), indicating that the proposed method is appropriate for use to predict China's e-commerce transaction volume. In addition, compared with four basis prediction techniques: GM, LR, BPNN, and SVR, the GAM has higher prediction accuracy. This shows that the proposed method can improve the prediction performance and bring a favorable forecast.

3.4. Future Trend of China's E-Commerce Transaction Volume. To grasp the future trend of China's e-commerce transaction volume, this study integrates the rolling

Table 1: China's e-commerce transaction volume (unit: trillion Chinese Yuan).

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Volume	6.09	8.11	10.4	16.39	21.79	26.10	29.16	31.63	34.81	37.21

Table 2: Actual and predicted values.

Year	Actual values	Predicted values				
		GM	LR	BPNN	SVR	GAM
2015	21.79	22.795	18.545	20.512	14.741	20.226
2016	26.10	30.709	25.930	23.618	24.629	26.309
2017	29.16	32.803	31.795	27.746	30.924	30.391
2018	31.63	33.832	34.015	30.092	32.819	32.271
2019	34.81	34.894	35.315	32.339	34.375	33.904
2020	37.21	37.940	37.575	36.452	36.614	37.139

Table 3: Prediction performances among various approaches.

Approaches	MAPE (%)
GM	7.32
LR	5.76
BPNN	5.70
SVR	8.44
RBFN	19.51
GPR	23.77
GAM	2.84

Table 4: MAPE criteria.

MAPE	Prediction power
<10%	Highly accurate prediction
10–20%	Good prediction
20–50%	Reasonable prediction
>50%	Inaccurate prediction

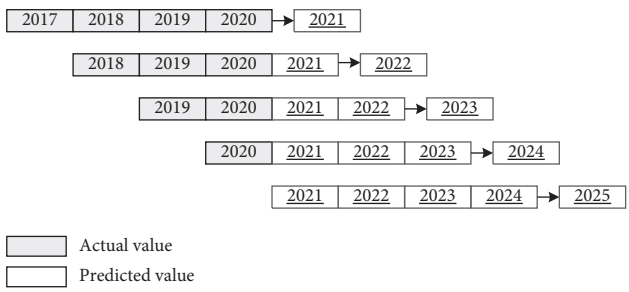


Figure 1: A schematic diagram of the rolling framework.

framework into the proposed GAM to improve its trend prediction performance. A schematic diagram of the rolling framework used in this study is shown in Figure 1.

Table 5 shows the predicted values of e-commerce transaction volume in China for the next five years, obtained using the proposed GAM with the rolling framework. According to this prediction, China's e-commerce transaction volumes will show a steady upward trend from 2021 to 2025.

By 2025, e-commerce transaction volumes in China are expected to be 45.566 trillion Chinese Yuan, which reflects an increase of approximately 22% relative to the transaction volumes in 2020. Under this development trend, the Chinese government should continue to invest in improving the e-commerce operating environment. The improvement of the quality of e-commerce transactions will not only increase people's consumer satisfaction but also help the country's economic growth.

Table 5: Prediction of China's e-commerce transaction volume (unit: trillion Chinese Yuan).

Year	2021	2022	2023	2024	2025
Predicted values	39.412	41.280	42.922	44.334	45.556

4. Conclusion and Discussion

To maintain an operational advantage in a highly competitive environment, enterprises must respond quickly to business problems, and it is essential to be able to make timely and correct decisions. Decision contexts often involve many uncontrollable factors and uncertain events. To overcome this unmanageable uncertainty, businesses must employ the right analytical techniques. Predictive techniques can help managers grasp future trends, mitigate the effects of uncertainty, and lead to meaningful decisions. In a variety of management situations, due to cost and time considerations, it is often impossible to obtain sufficient information, making decisions that require immediate response a more difficult task for managers. If managers can grasp the situation in real time through a limited amount of observation and carry out appropriate processing, effective operation management can be obtained. Therefore, building prediction models under small data sets have significant practical value.

Although popular prediction techniques have acceptable performance in normal applications, they are not suitable for prediction problems with insufficient information from small data sets. Grey system theory is a technique used for small data set analysis [31], and its research scope involves the problems encountered in this paper. Therefore, a modeling procedure based on grey system theory is proposed to integrate the advantages of various approaches to this specific problem and then obtains a more robust prediction output. Through the verification of China's e-commerce transaction volume, the results from this experiment exhibit that the proposed GAM can produce

favorable predictions with a MAPE as low as 2.84%. These results imply that the proposed method is useful for decision analysis with limited data. The proposed procedure outperforms the single popular methods in the experiment. Furthermore, the results obtained using data mining and statistical learning-based methods (such as LR, BPNN, SVR, RBFN, and GPR) are not superior to the proposed method, which may be due to small sample sizes. These approaches typically require a sufficient training data set to prevent overfitting and obtain robust models. If the training data set is large enough, the prediction performance of LR, BPNN, SVR, RBFN, and GPR should improve. Finally, the prediction shows that China's e-commerce transaction volumes are steadily increasing, indicating that the investment of resources to improve the e-commerce operating environment is in line with the direction and interests of China's economic development.

The proposed GAM can improve the stability of the forecasting results by aggregating the advantages of several models and is considered a feasible tool when only small data sets are available. In the future, introducing heuristic methods into the modeling procedure to improve prediction accuracy is a valuable research direction. In addition, the proposed method can be combined with some data preprocessing methods (for example, the virtual sample generating technique) to further improve its ability to handle problems with small data sets. Moreover, using more training samples to confirm the predictive power of the proposed GAM may be a worthwhile research direction. Finally, the proposed method should be used in other fields, such as finance, industry, engineering, and energy, to prove its reliability, validity, and practical value.

Data Availability

The data used in the experiment are listed in this article; anyone can use these data by citing this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Social Science Planning Project of Fujian Province (China) under Grant FJ2019B099, the Natural Science Foundation of Fujian Province (China) under Grant 2021J01326, and the Science and Technology Planning Project of Quanzhou city (China) under Grant 2019C096R.

References

[1] R. T. Clemen and T. Reilly, *Making Hard Decisions with Decisiontools*, Cengage Learning, Ohih, United States, 2013.

[2] J. Guo, C.-J. Chang, Y. Huang, and K.-P. Yu, "A fuzzy-decomposition grey modeling procedure for management decision analysis," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6670196, 6 pages, 2021.

[3] C.-J. Chang, D.-C. Li, C.-C. Chen, and W.-C. Chen, "Extrapolation-based grey model for small-data-set forecasting Economic Computation & Economic Cybernetics Studies & Research," *Economic Computation & Economic Cybernetics Studies & Research*, vol. 53, no. 1, pp. 171–182, 2019.

[4] C.-J. Chang, G. Li, S.-Q. Zhang, and K.-P. Yu, "Employing a fuzzy-based grey modeling procedure to forecast China's sulfur dioxide emissions," *International Journal of Environmental Research and Public Health*, vol. 16, no. 14, p. 2504, 2019.

[5] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[6] J. F. Hair, *Multivariate Data Analysis: A Global Perspective*, Pearson Education, London, England, 2010.

[7] I. H. Witten, E. Frank, M. A. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier Science, Amsterdam, Netherlands, 2016.

[8] Y. S. Lin and T. I. Tsai, "Using virtual data effects to stabilize pilot run neural network modeling," *Journal of Grey System*, vol. 26, no. no. 2, pp. 84–94, 2014.

[9] S. Liu, Y. Yang, and J. Forrest, *Grey Data Analysis: Methods, Models and Applications*, Springer, Singapore, 2016.

[10] D.-C. Li, Q.-S. Shi, and H.-Y. Chen, "Building robust models for small data containing nominal inputs and continuous outputs based on possibility distributions," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2805–2822, 2019.

[11] C.-J. Chang, C.-C. Chen, W.-L. Dai, and G. Li, "A new grey prediction model considering the data gap compensation," *Grey Systems: Theory and Application*, vol. 11, no. 4, pp. 650–663, 2021.

[12] Bo Zeng, Y. Tan, H. Xu, J. Quan, L. Wang, and X. Zhou, "Forecasting the electricity consumption of commercial sector in Hong Kong using a novel grey dynamic prediction model," *Journal of Grey System*, vol. 30, no. 1, pp. 159–174, 2018.

[13] Y.-S. Lin and D.-C. Li, "The generalized-trend-diffusion modeling algorithm for small data sets in the early stages of manufacturing systems," *European Journal of Operational Research*, vol. 207, no. 1, pp. 121–130, 2010.

[14] L. Feng, "Data analysis and prediction modeling based on deep learning in E-commerce," *Scientific Programming*, vol. 2022, Article ID 1041741, 12 pages, 2022.

[15] B., M. Mazhar, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," *Complexity*, vol. 2021, Article ID 5525271, 10 pages, 2021.

[16] J. Artin, V. Amin, M. Ahmadi, A. Sathish, P. Kumar, and A. Sharifi, "Presentation of a novel method for prediction of traffic with climate condition based on ensemble learning of neural architecture search (nas) and linear regression," *Complexity*, vol. 2021, Article ID 8500572, 13 pages, 2021.

[17] D. A. Lind, W. G. Marchal, and S. A. Wathen, *Basic Statistics for Business and Economics*, McGraw-Hill Education, New York, New York State, 2013.

[18] C. Zor and F. Çebi, "Demand prediction in health sector using fuzzy grey forecasting," *Journal of Enterprise Information Management*, vol. 31, no. 6, pp. 937–949, 2018.

[19] C. J. Chang, D. C. Li, W. L. Dai, and C. C. Chen, "Utilizing an adaptive grey model for short-term time series forecasting: a case study of wafer-level packaging," *Mathematical Problems in Engineering*, vol. 2013, Article ID 526806, 6 pages, 2013.

[20] S. F. Liu and Y. Lin, *Grey Systems: Theory and Applications*, Springer-Verlag, Berlin, Germany, 1st ed edition, 2010.

[21] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527–529, 1993.

[22] J. T. Yokuma and J. S. Armstrong, "Beyond accuracy: comparison of criteria used to select forecasting methods," *International Journal of Forecasting*, vol. 11, no. 4, pp. 591–597, 1995.

[23] H. Pan and H. Zhou, "Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce," *Electronic Commerce Research*, vol. 20, no. 2, pp. 297–320, 2020.

[24] S. Gu, B. Ślusarczyk, S. Hajizada, I. Kovalyova, and A. Sakhbieva, "Impact of the covid-19 pandemic on online consumer purchasing behavior," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 6, pp. 2263–2281, 2021.

[25] A. A. Vărzaru and C. George Bocean, "A two-stage sem–artificial neural network analysis of mobile commerce and its drivers," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 6, pp. 2304–2318, 2021.

[26] G. Schneider, *Electronic Commerce*, Cengage Learning, United States, 2016.

[27] R. B. Chase and F. R. Jacobs, *Operations and Supply Chain Management: The Core*, McGraw-Hill Education, NY, USA, 2016.

[28] R. M. Bădîrcea, A. G. Manta, N. M. Florea, J. Popescu, F. L. Manta, and S. Puiu, "E-commerce and the factors affecting its development in the age of digital technology: empirical evidence at Eu-27 level," *Sustainability*, vol. 14, Article ID 101, 1 page, 2022.

[29] C.-J. Chang, G. Li, J. Guo, and K.-P. Yu, "Data-driven forecasting model for small data sets," *Economic Computation & Economic Cybernetics Studies & Research*, vol. 54, no. 4, pp. 217–229, 2020.

[30] S. A. DeLurgio, *Forecasting Principles and Applications*, Irwin/McGraw-Hill, NY, USA, 1998.

[31] J. Liu and P. Gao, "Energy consumption predication in China based on the modified fractional grey prediction model," *Journal of Mathematics*, vol. 2021, Article ID 2477964, 7 pages, 2021.

Hindawi
Complexity
Volume 2022, Article ID 6156210, 10 pages
https://doi.org/10.1155/2022/6156210

Research Article

Prediction of Rockburst Intensity Grade in Deep Underground Excavation Using Adaptive Boosting Classifier

Mahmood Ahmad,[1,2] **Herda Yati Katman**,[3] **Ramez A. Al-Mansob**,[1] **Feezan Ahmad**,[4] **Muhammad Safdar**,[5] **and Arnold C. Alguno**[6]

[1]*Department of Civil Engineering, Faculty of Engineering, International Islamic University Malaysia, Jalan Gombak, Selangor 50728, Malaysia*
[2]*Department of Civil Engineering, University of Engineering and Technology Peshawar (Bannu Campus), Bannu 28100, Pakistan*
[3]*Institue of Energy Infrastructure, Universiti Tenaga Nasional, Putrajaya Campus, Jalan IKRAM-UNITEN, Kajang 43000, Malaysia*
[4]*State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian 116024, China*
[5]*Earthquake Engineering Center, University of Engineering and Technology Peshawar, Peshawar 25000, Pakistan*
[6]*Department of Physics, Mindanao State University-Iligan Institute of Technology, Iligan City 9200, Philippines*

Correspondence should be addressed to Herda Yati Katman; herda@uniten.edu.my

Received 17 February 2022; Accepted 4 April 2022; Published 5 May 2022

Academic Editor: Teddy Craciunescu

Rockburst phenomenon is the primary cause of many fatalities and accidents during deep underground projects constructions. As a result, its prediction at the early design stages plays a significant role in improving safety. The article describes a newly developed model to predict rockburst intensity grade using Adaptive Boosting (AdaBoost) classifier. A database including 165 rockburst case histories was collected from across the world to achieve a comprehensive representation, in which four key influencing factors such as maximum tangential stress of the excavation boundary, uniaxial compressive strength of rock, tensile rock strength, and elastic energy index were selected as the input variables, and the rockburst intensity grade was selected as the output. The output of the AdaBoost model is evaluated using statistical parameters including accuracy and Cohen's kappa index. The applications for the aforementioned approach for predicting the rockburst intensity grade are compared and discussed. Finally, two real-world applications are used to verify the proposed AdaBoost model. It is found that the prediction results are consistent with the actual conditions of the subsequent construction.

1. Introduction

In underground rock engineering, a rockburst is a type of dynamic geological disaster. It is a dynamic instability phenomenon that occurs when a rock mass or geological structure is subjected to high stress or is in a state of limit equilibrium. A rockburst, which still draws a lot of interest today, has a significant impact on rock stability in deep underground conditions [1–3]. Because rockbursts occur suddenly and intensely, they frequently result in harm, including death of workers, equipment damage, and even significant interruption and loss of revenue in underground deep excavation. Various strategies for controlling rockbursts have been proposed, such as temporary and permanent rock support structures; however, these efforts are ineffective since the severity of rockbursts is difficult to predict precisely. To record and evaluate the rockburst occurrences, different monitoring systems, such as a microseismic system, were used [4]. The rockburst intensity is recorded by the microseismic monitoring system after the rockburst occurs and thus cannot predict the rockburst in advance. The tendency and intensity of rockburst were contrastively analyzed by Chen and Guo [5] using the strain energy index model of rockburst. As a result, estimating and predicting rockburst intensity is critical for a safe and cost-effective deep underground excavation or mining in burst-prone soils before it occurs.

Machine learning (ML) algorithms have been widely used to tackle real-world problems in the last ten years, particularly in civil engineering. ML algorithms have been successfully used to a variety of real situations, paving the way for several promising opportunities in civil engineering and other domains such as environmental [6], geotechnical and geological [7–20], and other sciences [21–24] including rockburst hazards prediction [25–27]. Furthermore, a variety of machine learning methods have been used, for example, Support Vector Machine (SVM) [28], Artificial Neural Networks (ANNs) [29], Distance Discriminant Analysis (DDA) [30], Bayes Discriminant Analysis (BDA) [31], and Fisher Linear Discriminant Analysis (LDA) [32], and some systems are based upon hybrid (Zhou et al. [33]; Adoko et al. [34]; Liu et al. [35]) or ensemble (Ge and Feng [36]; Dong et al. [37]) analyzing long-term prediction of rockburst. These studies provided new concepts and ways for predicting rockbursts. However, each of the methods listed above has its own set of benefits and drawbacks. Understanding, predicting, and controlling rock bursts still pose a considerable challenge for underground engineering. Furthermore, the number of data and the type of ML algorithms have an influence on the accuracy of rockburst intensity prediction. As a result, developing a high-performing and time-saving ensemble classifier for a larger dataset is critical. Many researchers have increasingly implemented the AdaBoost-based method for prediction problems such as rock mass class and soil classification as a vital means in recent years [38, 39]. For classification, prediction, and recognition issues, the AdaBoost methodology is widely regarded as the most successful and reliable artificial intelligence method. The article aims to add the following described contributions to this field: (1) A machine learning classifier for rockburst prediction based on case histories data is proposed. (2) The performance of AdaBoost is compared with other classifiers to confirm that the algorithm has superior or at par classification precision. (3) The effectiveness and feasibility in engineering practice applications and real-world examples are analyzed to predict rockburst intensity grade.

The rest of this paper is arranged as follows. The second section introduces the selection of indicators and the data collection, AdaBoost algorithm, and performance measures. The establishment of the algorithm model is described in the third section. In the fourth section, results are discussed, and the proposed algorithm is compared with the developed empirical criteria, widely used models, and two real-world applications are used to verify the proposed model. Finally, the conclusions are drawn in the fifth section.

2. Materials and Methods

2.1. Dataset. A total of 165 cases of rockburst events reported in the literature were collected to build a dataset [33, 40]. The maximum tangential stress of the excavation boundary (σ_θ), the uniaxial compressive strength of rock (σ_c), the tensile rock strength (σ_t), and the elastic energy index (*Wet*) are selected as input parameters in this study by referring to the previous research [41, 42] and rockburst intensity as the output. These input variables are commonly applied in rockburst classification and can provide fundamental understandings about rockburst occurrence in underground conditions. σ_c, σ_t, and *Wet* were obtained by rock mechanics experiments, and σ_θ was calculated according to the stress of the surrounding rock. Through field observation and evaluation, the rockburst grade was obtained. According to rock failure properties, the output parameter, i.e., rockburst intensity, contains four different classes, namely, no, moderate, strong, and violent, which are indicated by 1, 2, 3, and 4, respectively, as shown in Table 1 [40].

Figure 1 shows a boxplot of each affecting parameter for the four rockburst levels. As shown in Figure 1, the rockburst hazard intensity grades are associated with each attribute. Table 2 contains an overview of the case histories, as well as parameter statistics. The following is a brief summary of various input parameters.

2.1.1. Maximum Tangential Stress of the Surrounding Rock. The maximum tangential stress is frequently used to determine the angle at which a rock fractures [43]. For example, Ryder [44] determined that the fault-slip and shear fracture modes had a significant role in African metal mines in his investigation of the influence of excess shear stress on rockburst–prone circumstances, whereas Qian et al. [45] proposed two rock burst dynamic failure modes: one strain mode resulting from the rock failure and one sliding mode caused by the fault-slip and shear fracture events. Qian et al. [45] also analyzed two rockburst accidents in coal mines in China, stating that the instability due to rockburst occurrence could also be classified as fault-slip and shear fracture modes. As a result, past studies show that the maximum tangential stress has a significant impact on the incidence of shear fracture instabilities in tunnels, making it an important parameter for rockburst prediction. It is also an often used parameter in the data set.

2.1.2. Uniaxial Compressive and Tensile Strength. Other characteristics that can influence rockburst include uniaxial compressive strength (UCS) and uniaxial tensile strength (UTS), both of which have been used in the past. UCS and UTS values are widely known parameters for rockburst hazards prediction modeling.

2.1.3. Elastic Energy Index. The proportion of residual strain energy that dissipated during a single loading-unloading cycle under uniaxial compression is defined by the elastic energy index, *Wet* [46, 47]. This parameter is related to the rockburst hazards, and Wang et al. [48] developed a rockburst prediction criterion based on *Wet*. The *Wet* values can be easily obtained through laboratory tests and direct (double-hole method) or indirect (rebound method) in situ evaluations.

2.2. AdaBoost Algorithm. The AdaBoost algorithm, short for Adaptive Boosting, is a boosting approach used in machine learning as an ensemble method that uses decision trees as

Table 1: Grading criteria of rockburst intensity.

Rockburst grade	No rockburst (1)	Moderate rockburst (2)	Strong rockburst (3)	Violent rockburst (4)
σ_θ/σ_c	<0.3	0.3–0.5	0.5–0.7	>0.7
σ_c/σ_t	>40	26.7–40	14.5–26.7	<14.5
W_{et}	>5	3.5–5.0	2.0–3.5	<2.0

Note. σ_θ/σ_c = stress concentration factor; σ_c/σ_t = rock brittleness.

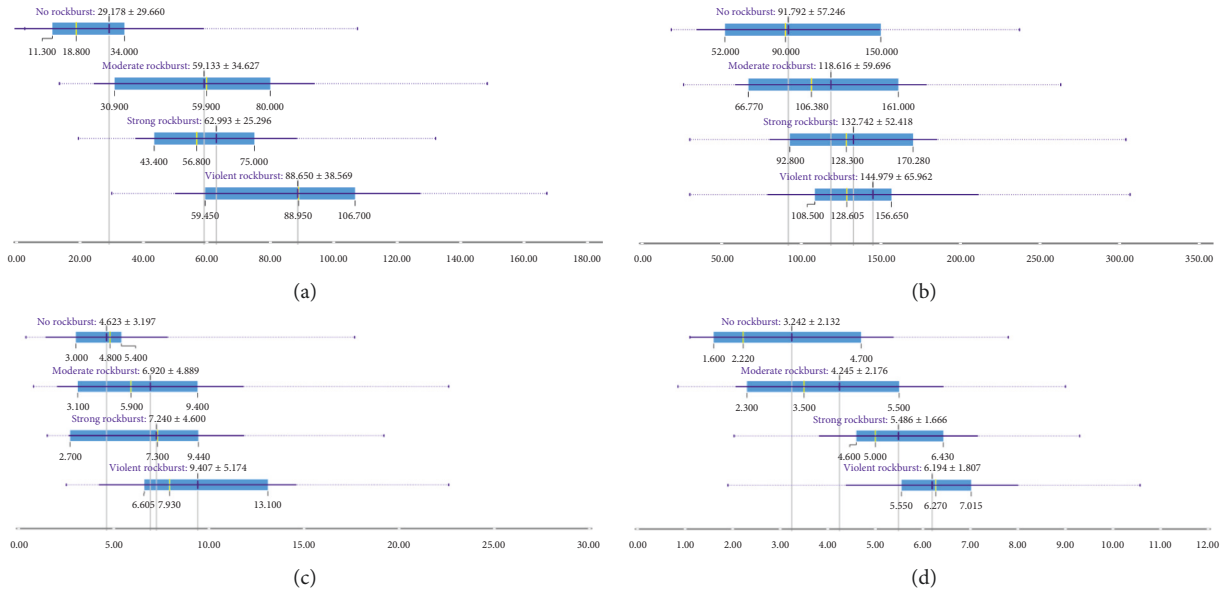


Figure 1: Boxplot of each influencing parameter to corresponding rockburst intensity grade. (a) σ_θ, (b) σ_c, (c) σ_t, and (d) W_{et}.

Table 2: Inputs and output statistics of the present study.

Rockburst rank	Tangential stress, σ_θ (MPa)	Uniaxial compression strength, σ_c (MPa)	Uniaxial tensile strength, σ_t (MPa)	Elastic energy index, w_{et}
3	90	170	11.3	9
2	90	220	7.4	7.3
2	62.6	165	9.4	9
...	...	...	...	...
3	89	236	8.3	5
3	98.6	120	6.5	3.8
4	108.4	140	8	5
Mean	59.988	123.443	7.03	4.861
Standard error	2.811	4.71	0.37	0.17
Standard deviation	36.108	60.498	4.79	2.185
Sample variance	1303.799	3660.031	22.98	4.773
Skewness	0.785	0.621	1.22	0.194
Minimum	2.6	18.32	0.38	0.85
Maximum	167.2	306.58	22.6	10.57
Count	165	165	165	165

the main classifier. It is called Adaptive Boosting as the weights are reassigned to each instance, with higher weights assigned to incorrectly classified instances. Freund and Schapire's AdaBoost is the most widely used version of the boosting algorithm [49], making maximum use of a classifier by improving its accuracy. It is a simple learning approach that creates a strong classifier from a small number of efficient but weak classifiers (see Figure 2). The goal is to combine the weak classifiers to improve their performance. As a result, the final robust classifier generated a data set for a model that can predict the class of a new observation. AdaBoost improves the classification efficiency of a simple learning algorithm by combining sets of weak classifiers to build a more robust classifier. In the language of boosting algorithms, the simple learning algorithm is known as a weak learner, and it selects a small, effective set of weak

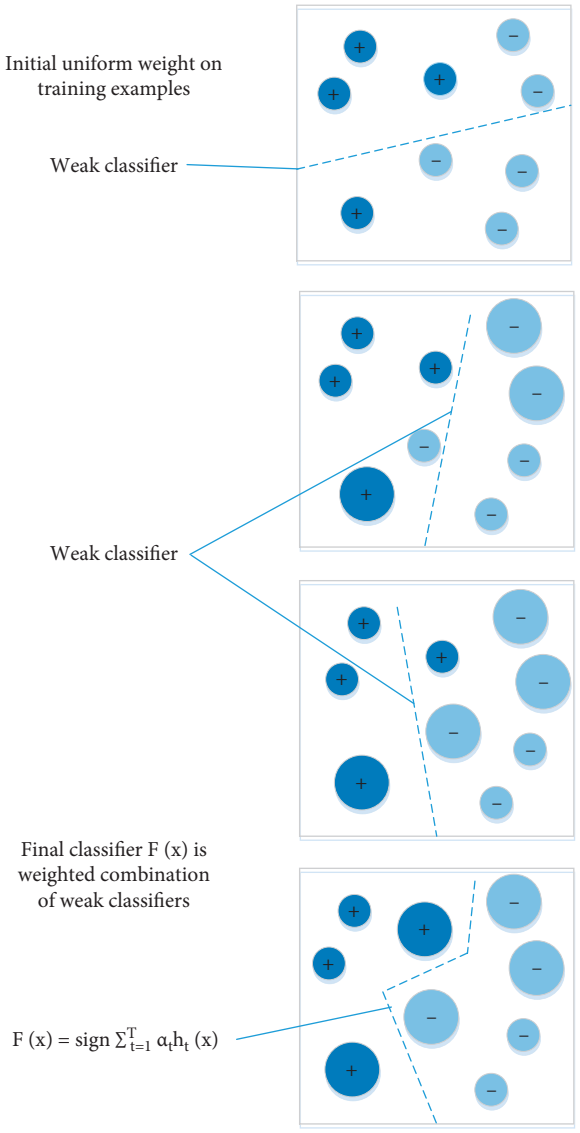


Figure 2: AdaBoost builds a strong classifier from a set of weak classifiers, as shown in this simple example.

classifiers with the lowest classification error from a wide number of potential features. The weak learner does not categorise the training data well even using the best classification function. To enhance the weak learner, it is necessary to solve a series of learning challenges. After the first learning cycle, the instances are reweighted to highlight those that were inaccurately categorised by the previous weak classifier. The final robust classifier uses a weighted combination of the weak classifiers to determine the best threshold classification function for each feature.

Algorithm 1 [50] shows the AdaBoost technique used to solve a prediction problem.

2.3. Performance Metric. In this study, the classical methods for model evaluation are used. The accuracy (*ACC*) and Cohen's kappa index were used to evaluate rockburst classification. A confusion matrix is commonly used as a standard for evaluating the performance of a classification model on training and testing datasets with known true values.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mm} \end{bmatrix}, \tag{1}$$

where m represents the number of rockburst levels, ×11 is the number of features accurately predicted for the class m, and xmm denotes the number of class features categorised to class n. Based on the confusion matrix, *ACC* and Cohen's kappa index are determined by (2) and (3), respectively.

$$ACC = \left(\frac{1}{n} \sum_{i=1}^{m} x_{ii} \right) \times 100\%, \tag{2}$$

$$Kappa = \frac{n \sum_{i=1}^{m} x_{ii} - \sum_{i=1}^{m} (x_{i+} \times x_{+i})}{n^2 - \sum_{i=1}^{m} (x_{i+} \times x_{+i})}. \tag{3}$$

A kappa value of less than 0.4 indicates poor agreement, while a value of 0.4 and above indicates good agreement [51, 52]. The ideal condition of a good model should have high *ACC* and kappa values simultaneously.

3. Model Development

The proposed model for predicting rockburst intensity grade was developed using Orange software. The model structure was based on an input matrix (x) defined by $x = \{\sigma_\theta, \sigma_c, \sigma_t, W_{et}\}$ that provided the predictor variables, while the target variable (y) is rockburst intensity grade. During every modeling step, the most critical task is to identify the appropriate size of the training and testing datasets. The way data is split into training and research sets has a substantial impact on data mining results [53]. The main goal of the statistical analysis was to ensure that the statistical properties of the subsets were as similar as possible, and thus they represented the same statistical population. The dataset was divided into 137 (83%) training cases and 28 (17%) test cases and was kept the same as that of Zhao and Chen [41] owing to fairly evaluating the predictive performance of the proposed AdaBoost model in this work. The AdaBoost model was tuned to optimize the rockburst intensity grade prediction using a trial and error method. Figure 3 depicts the prediction model's construction.

Most ML algorithms have hyperparameters that need to be tuned [54]. The optimization method attempts to find the appropriate parameters for the AdaBoost model in order to achieve the best prediction accuracy. Some critical hyperparameters in the AdaBoost model are tuned in this study, as shown in Table 3. The definitions of these hyperparameters are also clarified in Table 3. First, the search range of different hyperparameters values is specified randomly and then adjusted throughout the trials until the best fitness metrics shown in Table 3 were reached.

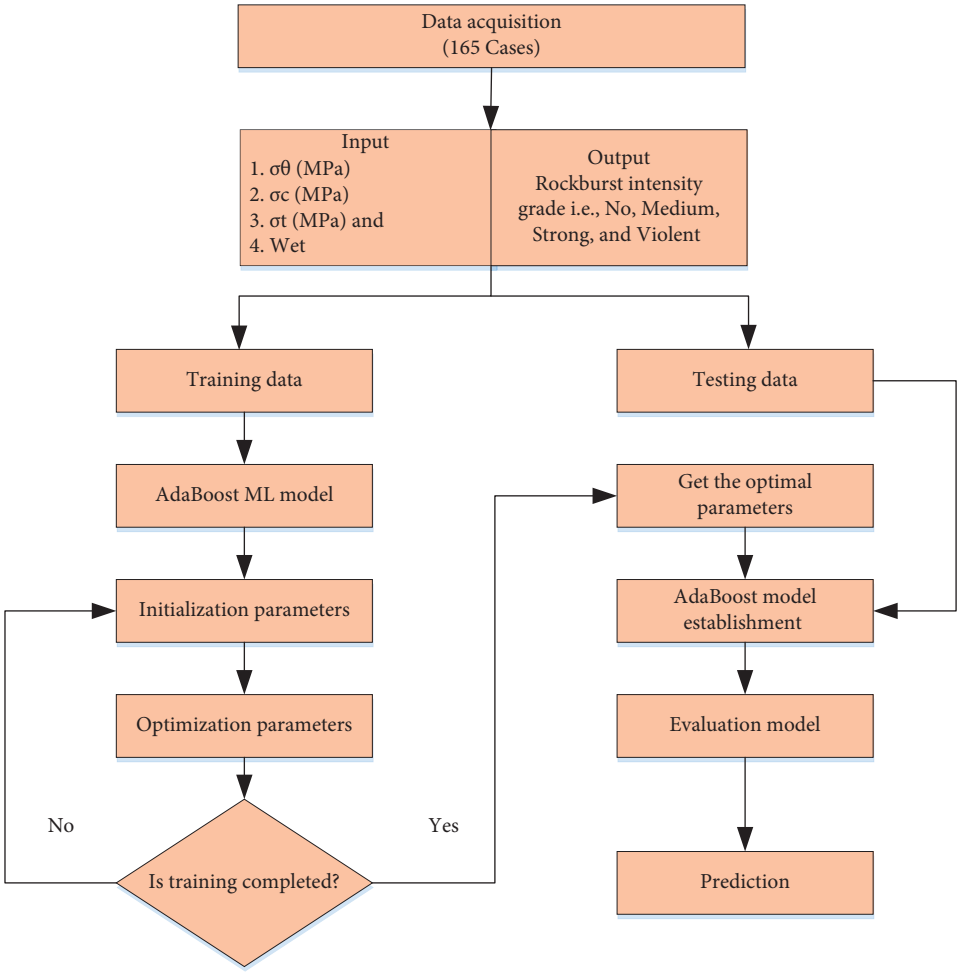


Figure 3: Flowchart of AdaBoost model for prediction of rockburst intensity grade.

Input: Dataset, $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;
The weight of each sample in the training set constitutes a weight vector, Z_t;
The base learning algorithm, L;
Number of learning rounds, T;
Process:
$D_1(i) = 1/m$; % Initialize the weight distribution
For $t = 1, 2, \ldots, T$;
For $h_t = L(D, D_t)$; % Train a base learner, h_t from D using distribution D_t
$\varepsilon_t = Pr_{i \sim D_i}[h_t(x_i \neq y_i)]$; % Measure the error of h_t
$\alpha_t = 1/2\ln 1 - \varepsilon_t/\varepsilon_t$; % Determine the weight of h_t
$D_{t+1}(i) = D_t(i)/Z_t \times \begin{cases} \exp(-\alpha_t) \text{ if } h_t(x_i) = y_i \\ \exp(\alpha_t) \text{ if } h_t(x_i) \neq y_i \end{cases}$;
% Update the distribution, where Z_t is a normalized factor with enables D_{t+1} to be a distribution end.
Output: $F(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x)$

Algorithm 1: AdaBoost.

4. Results and Discussion

4.1. Comparison of the AdaBoost with Baseline Models. The performance of the AdaBoost model was evaluated with ANN, convolutional neural network (CNN), J48, and random tree (RT) models. The AdaBoost model prediction result was the same as the performance of CNN and RT models, having an accuracy of 100% (Table 4), and was found better than the ANN and J48 models. The ANN, CNN, and J48 model's accuracy achieved 89.286%, 100%,

Table 3: Hyperparameter optimization results.

Algorithm	Hyperparameter	Explanation	Optimal value/function
AdaBoost	Number of estimators	Number of trees	0.5
	Learning rate	It determines to what extent the newly acquired information will override the old information	0.75
	Boosting algorithm	Updates base estimator's weight with probability estimates or classification results	SAMME.R
	Regression loss function	Linear/square/exponential	Exponential

Table 4: Performance metrics of each model for test data.

Method	*ACC* (%)	Kappa
Russenes criterion [55]	42.867	0.222
Rock brittleness coefficient criterion [48]	53.571	0.352
Elastic energy index [46]	39.286	0.138
ANN [41]	89.286	0.856
CNN [41]	**100**	**1.000**
Random tree [42]	**100**	**1.000**
J48 [42]	92.857	0.904
AdaBoost (present study)	**100**	**1.000**

Note. Bold values indicate the highest value for each model.

Table 5: Comparison of prediction results by different methods for the testing samples.

S. No.	σ_θ (MPa)	σ_c (MPa)	σ_t (MPa)	W_{et}	Actual	Russenes criterion [55]	Rock brittleness coefficient criterion [48]	Elastic energy index [46]	ANN [41]	CNN [41]	RT [42]	J48 [42]	AdaBoost (present study)
1	34	150	5.4	7.8	1	2	2	3	1	1	1	1	1
2	60.7	111.5	7.86	6.16	4	3	4	3	4	4	4	4	4
3	54.2	134	9.09	7.08	3	3	3	3	3	3	3	3	3
4	70.3	129	8.73	6.43	3	3	3	3	3	3	3	3	3
5	35	133.4	9.3	2.9	2	2	4	2	2	2	2	2	2
6	157.3	91.23	6.92	6.27	4	4	4	3	4	4	4	4	4
7	148.4	66.77	3.81	5.08	2	4	3	3	2	2	2	2	2
8	132.1	51.5	2.47	4.63	3	4	3	2	2	3	3	3	3
9	127.9	35.82	1.24	3.67	2	4	2	2	2	2	2	2	2
10	107.5	21.5	0.6	2.29	1	4	2	2	1	1	1	1	1
11	96.41	18.32	0.38	1.87	1	4	1	1	1	1	1	1	1
12	167.2	110.3	8.36	6.83	4	4	4	3	4	4	4	4	4
13	38.2	53	3.9	1.6	1	4	4	1	1	1	1	1	1
14	11.3	90	4.8	3.6	1	1	3	2	1	1	1	1	1
15	92	263	10.7	8	2	3	3	3	2	2	2	2	2
16	62.4	235	9.5	9	4	2	3	3	4	4	4	3	4
17	43.4	136.5	7.2	5.6	4	3	3	3	3	4	4	4	4
18	11	105	4.9	4.7	1	1	3	2	1	1	1	1	1
19	90	170	11.3	9	3	3	3	3	3	3	3	3	3
20	90	220	7.4	7.3	2	3	2	3	2	2	2	2	2
21	62.6	165	9.4	9	2	3	3	3	2	2	2	2	2
22	55.4	176	7.3	9.3	3	3	3	3	4	3	3	3	3
23	30	88.7	3.7	6.6	3	3	3	3	3	3	3	3	3
24	48.75	180	8.3	5	3	2	3	3	3	3	3	3	3
25	80	180	6.7	5.5	2	3	2	3	2	2	2	2	2
26	89	236	8.3	5	3	3	2	3	3	3	3	3	3
27	98.6	120	6.5	3.8	3	4	3	2	3	3	3	3	3
28	108.4	140	8	5	4	4	3	3	4	4	4	4	4

and 92.857%, respectively. Furthermore, we compared the results (summarized in Table 4) with conventional empirical models, such as the rock brittleness coefficient criterion, elastic energy index, and Russenes criterion. The AdaBoost model performed better than the empirical models. The calculated results of AdaBoost, ANN, CNN, J48, RT, and

Table 6: Applications in real-world Rockburst prediction of the proposed AdaBoost model.

Project	σ_θ (MPa)	σ_c (MPa)	σ_t (MPa)	W_{et}	Actual condition	Prediction
Duoxiongla tunnel	87.3	137.7	9.62	7.14	Strong	Strong
Anlu tunnel	17.02	85.09	1.3	6.14	Moderate	Moderate
	16.7	83.5	1.3	6.53	Moderate	Moderate
	17.35	86.77	1.3	3.22	Moderate	Moderate
	16.87	80.33	1.3	6.92	Moderate	Moderate

conventional empirical models, such as the rock brittleness coefficient criterion, elastic energy index, and Russenes criterion, are listed in Table 5. Obviously, the predicted rank of 28 samples was in excellent agreement with the actual rank, and all samples were classified correctly. The comparison analysis confirmed that the proposed AdaBoost model achieved a better performance than the other machine learning classifiers which can effectively mine the relationship between rockburst and its influence factors.

The proposed AdaBoost model was compared with the findings of the previous studies. Zhou et al. [56] compared the performance of 10 machine learning algorithms to analyze rockburst events, including 246 cases, considering seven input variables. Lin et al. [57] investigated rockburst events using machine learning models considering 246 rockburst cases having six input variables. The accuracy performances of the RF model developed by Zhou et al. [56] and Lin et al. [57] were 0.73 and 0.61, respectively. Although both models developed by Zhou et al. [56] and Lin et al. [57] considered seven and six input variables, they still provide lower prediction accuracy compared to the developed model.

4.2. Applications in Real-World Rockburst Prediction. Two real-world examples are analyzed using our proposed AdaBoost-based rockburst prediction model to study the effectiveness and feasibility in engineering practice applications. Five rockburst events in two different tunnel projects were predicted by the AdaBoost model. The field data were collected from available literature, including the Duoxiongla tunnel [58] and Anlu tunnel [59]. The prediction outcomes are summarized in Table 6, indicating that the rock burst intensity was predicted correctly for all cases. The prediction results in the real-world rockburst prediction cases are basically consistent with this strong-to-moderate intensity grading. This study proves that the AdaBoost model is a robust alternative tool for the rockburst intensity grade assessment, and it can be successfully applied in various geotechnical engineering projects.

5. Limitations and Future Works

The proposed approach obtains desirable prediction results, although some limitations should be addressed in the future.

(1) The dataset is relatively small and unbalanced. The prediction performance of ML algorithms is heavily affected by the number and quality of dataset. Generally, if the dataset is small, the generalization and reliability of model would be influenced, although AdaBoost algorithm works well with small datasets. Furthermore, the suggested model is open to further development, and the accumulation of more data will lead to a much better prediction capacity. It is important to note that the validity of the proposed model is limited by the data ranges used to train the model.

(2) Other variables may have an effect on the prediction outcomes. Numerous factors influence the risk of a rockburst, including rock properties, energy, excavation depth, and support structure, among others. Although the four indicators used in this study can define the required conditions for rockburst hazard assessment to some degree, some other indicators, such as the buried depth of the tunnel, failure duration time, and energy-based burst potential index, may also have an impact on rockburst hazard. As a consequence, it is crucial to look into the effects of these variables on the prediction outcomes.

6. Conclusions

In this paper, the AdaBoost classifier's application was investigated to evaluate the rockburst phenomenon. The predictive variables for the AdaBoost model included the main effective parameters on rockburst, i.e., σ_θ, σ_c, σ_t, and W_{et}. The model was developed and tested using Orange software based on a database including 165 rockburst case histories. The main conclusion points are summarized below:

(1) The comparison of proposed model efficiency and previously developed empirical criteria revealed that the AdaBoost model is remarkably better than empirical criteria with accuracy and kappa value obtained as 100% and 1.00, respectively.

(2) The proposed approach was compared with other machine learning-based models in the literature. The comparison results have shown that the prediction accuracy of the proposed model is as adequate as other techniques such as CNN and RT models.

(3) Two real-world rockburst examples are used to verify the proposed model's accuracy and effectiveness. It can be concluded that the AdaBoost classifier is a feasible and efficient tool for the classification of rockburst intensity grades. The proposed model can be applied in the initial stages of underground projects and the rockburst phenomenon can be assessed by an acceptable accuracy, which can reduce casualties due to rockburst.

Data Availability

The data that support the findings of this study are openly available in [33, 40].

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

M.A. and R.A.A.-M. conceptualized the study; M.A., F.A., and H.Y.K. were responsible for methodology; M.A. was responsible for software, performed formal analysis, and prepared the original draft; M.A., R.A.A.-M., and F.A. validated the data; M.A., F.A., and M.S. investigated the data; M.A., M.S., and F.A. reviewed and edited the manuscript; H.Y.K. was responsible for resources and was involved in visualization, funding acquisition, and project administration; R.A.A.-M. carried out study supervision. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors gratefully acknowledge Grant no. J510050002-IC-6 Boldrefresh 2025-Centre of Excellence.

References

[1] P. Kaiser, D. Tannant, and D. McCreath, *Canadian Rockburst Support Hand-Book, Geomechanics Research Centre*, Laurentian University, Sudbury, Ontario, 1996.

[2] W. Ortlepp, "RaSiM comes of age—a review of the contribution to the understanding and control of mine rockbursts," in *Proceedings of the Sixth International Symposium on Rockburst and Seismicity in Mines*, pp. 9–11, Perth, Western Australia, 2005.

[3] M. Wu, Y. Ye, Q. Wang, and N. Hu, "Development of rockburst research: a comprehensive review," *Applied Sciences*, vol. 12, no. 3, p. 974, 2022.

[4] C.-P. Lu, L.-M. Dou, B. Liu, Y.-S. Xie, and H.-S. Liu, "Microseismic low-frequency precursor effect of bursting failure of coal and rock," *Journal of Applied Geophysics*, vol. 79, pp. 55–63, 2012.

[5] L. Chen and L. Guo, "Discussions on the complete strain energy characteristics of deep granite and assessment of rockburst tendency," *Shock and Vibration*, vol. 2020, pp. 1–9, Article ID 8825505, 2020.

[6] A. Froemelt, D. J. Dürrenmatt, and S. Hellweg, "Using data mining to assess environmental impacts of household consumption behaviors," *Environmental Science & Technology*, vol. 52, no. 15, pp. 8467–8478, 2018.

[7] A. Mahmood, X.-W. Tang, J.-N. Qiu, W.-J. Gu, and A. Feezan, "A hybrid approach for evaluating CPT-based seismic soil liquefaction potential using Bayesian belief networks," *Journal of Central South University*, vol. 27, no. 2, pp. 500–516, 2020.

[8] M. Ahmad, X.-W. Tang, J.-N. Qiu, and F. Ahmad, "Evaluating seismic soil liquefaction potential using bayesian belief network and C4.5 decision tree approaches," *Applied Sciences*, vol. 9, no. 20, p. 4226, 2019.

[9] M. Ahmad, X. Tang, J. Qiu, F. Ahmad, and W. Gu, "LLDV-a comprehensive framework for assessing the effects of liquefaction land damage potential," in *Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 527–533, Dalian, China, November 2019.

[10] M. Ahmad, X.-W. Tang, J.-N. Qiu, F. Ahmad, and W.-J. Gu, "A step forward towards a comprehensive framework for assessing liquefaction land damage vulnerability: exploration from historical data," *Frontiers of Structural and Civil Engineering*, vol. 14, no. 6, pp. 1476–1491, 2020.

[11] M. Ahmad, X. Tang, and F. Ahmad, "Evaluation of liquefaction-induced settlement using random forest and REP tree models: taking pohang earthquake as a case of illustration," in *Natural Hazards-Impacts*, Adjustments & Resilience, IntechOpen, 2020.

[12] M. Ahmad, N. A. Al-Shayea, X.-W. Tang, A. Jamal, and F. Ahmad, "Predicting the pillar stability of underground mines with random trees and C4.5 decision trees," *Applied Sciences*, vol. 10, no. 18, p. 6486, 2020.

[13] M. Ahmad, P. Kamiński, P. Olczak et al., "Development of prediction models for shear strength of rockfill material using machine learning techniques," *Applied Sciences*, vol. 11, no. 13, p. 6167, 2021.

[14] A. M. Noori, R. Mikaeil, M. Mokhtarian, S. S. Haghshenas, and M. Foroughi, "Feasibility of intelligent models for prediction of utilization factor of TBM," *Geotechnical & Geological Engineering*, vol. 38, no. 3, pp. 3125–3143, 2020.

[15] A. Dormishi, M. Ataei, R. Mikaeil, R. Khalokakaei, and S. S. Haghshenas, "Evaluation of gang saws' performance in the carbonate rock cutting process using feasibility of intelligent approaches," *Engineering Science and Technology, an International Journal*, vol. 22, no. 3, pp. 990–1000, 2019.

[16] R. Mikaeil, S. S. Haghshenas, and S. H. Hoseinie, "Rock penetrability classification using artificial bee colony (ABC) algorithm and self-organizing map," *Geotechnical & Geological Engineering*, vol. 36, pp. 1309–1318, 2018.

[17] R. Mikaeil, S. S. Haghshenas, Y. Ozcelik, and H. H. Gharehgheshlagh, "Performance evaluation of adaptive neuro-fuzzy inference system and group method of data handling-type neural network for estimating wear rate of diamond wire saw," *Geotechnical & Geological Engineering*, vol. 36, no. 6, pp. 3779–3791, 2018.

[18] E. Momeni, R. Nazir, D. Jahed Armaghani, and H. Maizir, "Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN," *Measurement*, vol. 57, pp. 122–131, 2014.

[19] C. Xie, H. Nguyen, Y. Choi, and D. Jahed Armaghani, "Optimized functional linked neural network for predicting diaphragm wall deflection induced by braced excavations in clays," *Geoscience Frontiers*, vol. 13, no. 2, p. 101313, 2022.

[20] D. J. Armaghani, E. T. Mohamad, M. S. Narayanasamy, N. Narita, and S. Yagiz, "Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition," *Tunnelling and Underground Space Technology*, vol. 63, pp. 29–43, 2017.

[21] G. Guido, S. S. Haghshenas, S. S. Haghshenas, A. Vitale, V. Gallelli, and V. Astarita, "Development of a binary classification model to assess safety in transportation systems using GMDH-type neural network algorithm," *Sustainability*, vol. 12, no. 17, p. 6735, 2020.

[22] A. Fiorini Morosini, S. Shaffiee Haghshenas, D. Y. Choi, and Z. W. Geem, "Sensitivity analysis for performance evaluation of a real water distribution system by a pressure driven analysis approach and artificial intelligence method," *Water*, vol. 13, no. 8, p. 1116, 2021.

[23] P. G. Asteris, P. B. Lourenço, P. C. Roussis et al., "Revealing the nature of metakaolin-based concrete materials using artificial intelligence techniques," *Construction and Building Materials*, vol. 322, p. 126500, 2022.

[24] M. Hajihassani, D. Jahed Armaghani, H. Sohaei, E. Tonnizam Mohamad, and A. Marto, "Prediction of airblast-overpressure induced by blasting using a hybrid artificial neural network and particle swarm optimization," *Applied Acoustics*, vol. 80, pp. 57–67, 2014.

[25] D. Guo, H. Chen, L. Tang, Z. Chen, and P. Samui, "Assessment of rockburst risk using multivariate adaptive regression splines and deep forest model," *Acta Geotechnica*, vol. 14, pp. 1–23, 2021.

[26] D. Li, Z. Liu, D. J. Armaghani, P. Xiao, and J. Zhou, "Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments," *Scientific Reports*, vol. 12, no. 1, pp. 1844–1923, 2022.

[27] D. Papadopoulos and A. Benardos, "Enhancing machine learning algorithms to assess rock burst phenomena," *Geotechnical & Geological Engineering*, vol. 39, no. 8, pp. 5787–5809, 2021.

[28] Z. Hong-Bo, "Classification of rockburst using support vector machine," *Rock and Soil Mechanics*, vol. 26, pp. 642–644, 2005.

[29] D. Chen, X. Feng, C. Yang, B. Chen, S. Qiu, and D. Xu, "Neural network estimation of rockburst damage severity based on engineering cases," *Rock Characterisation, Modelling and Engineering Design Methods*, vol. 6, pp. 457–462, 2013.

[30] F. Gong and X. Li, "A distance discriminant analysis method for prediction of possibility and classification of rockburst and its application," *Yanshilixue Yu Gongcheng Xuebao/Chinese Journal of Rock Mechanics and Engineering*, vol. 26, pp. 1012–1018, 2007.

[31] F. Gong, X. Li, and W. Zhang, "Rockburst prediction of underground engineering based on Bayes discriminant analysis method," *Rock and Soil Mechanics*, vol. 31, pp. 370–377, 2010.

[32] J. Zhou, X.-z. Shi, L. Dong, H.-y. Hu, and H.-y. Wang, "Fisher discriminant analysis model and its application for prediction of classification of rockburst in deep-buried long tunnel," *Journal of Coal Science and Engineering*, vol. 16, no. 2, pp. 144–149, 2010.

[33] J. Zhou, X. Li, and X. Shi, "Long-term prediction model of rockburst in underground openings using heuristic algorithms and support vector machines," *Safety Science*, vol. 50, no. 4, pp. 629–644, 2012.

[34] A. C. Adoko, C. Gokceoglu, L. Wu, and Q. J. Zuo, "Knowledge-based and data-driven fuzzy modeling for rockburst prediction," *International Journal of Rock Mechanics and Mining Sciences*, vol. 61, pp. 86–95, 2013.

[35] Z. Liu, J. Shao, W. Xu, and Y. Meng, "Prediction of rock burst classification using the technique of cloud models with attribution weight," *Natural Hazards*, vol. 68, no. 2, pp. 549–568, 2013.

[36] Q. Ge and X. Feng, "Classification and prediction of rockburst using AdaBoost combination learning method," *ROCK AND SOIL MECHANICS-WUHAN-*, vol. 29, p. 943, 2008.

[37] L.-j. Dong, X.-b. Li, and K. Peng, "Prediction of rockburst classification using Random Forest," *Transactions of Nonferrous Metals Society of China*, vol. 23, no. 2, pp. 472–477, 2013.

[38] Q. Liu, X. Wang, X. Huang, and X. Yin, "Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data," *Tunnelling and Underground Space Technology*, vol. 106, p. 103595, 2020.

[39] B. T. Pham, M. D. Nguyen, T. Nguyen-Thoi et al., "A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling," *Transportation Geotechnics*, vol. 27, p. 100508, 2021.

[40] Y. Pu, D. B. Apel, and B. Lingga, "Rockburst prediction in kimberlite using decision tree with incomplete data," *Journal of Sustainable Mining*, vol. 17, no. 3, pp. 158–165, 2018.

[41] H. Zhao and B. Chen, "Data-Driven model for rockburst prediction," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–14, Article ID 5735496, 2020.

[42] M. Ahmad, J.-L. Hu, M. Hadzima-Nyarko et al., "Rockburst hazard prediction in underground projects using two intelligent classification techniques: a comparative study," *Symmetry*, vol. 13, no. 4, p. 632, 2021.

[43] M. R. M. Aliha and M. R. Ayatollahi, "Analysis of fracture initiation angle in some cracked ceramics using the generalized maximum tangential stress criterion," *International Journal of Solids and Structures*, vol. 49, no. 13, pp. 1877–1883, 2012.

[44] J. Ryder, "Excess shear stress in the assessment of geologically hazardous situations," *Journal of the South African Institute of Mining and Metallurgy*, vol. 88, pp. 27–39, 1988.

[45] Q. Qian and L. R. S. Y. Huang, "Enhanced resistance to blast fungus in rice (Oryza sativa L.) by expressing the ribosome-inactivating protein alpha-momorcharin," *Plant Science*, vol. 217-218, pp. 1–7, 2014.

[46] A. Kidybiński, "Bursting liability indices of coal," *In Proceedings of International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, vol. 18, pp. 295–304, 1981.

[47] S. P. Singh, "Burst energy release index," *Rock Mechanics and Rock Engineering*, vol. 21, no. 2, pp. 149–155, 1988.

[48] Y. Wang, W. Li, and Q. Li, "Fuzzy estimation method of rockburst prediction," *Chinese Journal of Rock Mechanics and Engineering*, vol. 17, pp. 493–501, 1998.

[49] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, p. 1612, 1999.

[50] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications*, vol. 38, no. 1, pp. 223–230, 2011.

[51] Y. Sakiyama, H. Yuki, T. Moriya et al., "Predicting human liver microsomal stability with machine learning techniques," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 6, pp. 907–915, 2008.

[52] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, vol. 33, no. 2, pp. 363–374, 1977.

[53] M. Rezania, A. Faramarzi, and A. A. Javadi, "An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 142–153, 2011.

[54] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms," *Mathematics*, vol. 8, no. 5, p. 765, 2020.

[55] B. Russenes, *Analysis of Rock Spalling for Tunnels in Steep valley Sides*, Norwegian Institute of Technology, 1974.

[56] J. Zhou, X. Li, and H. S. Mitri, "Classification of rockburst in underground projects: comparison of ten supervised learning methods," *Journal of Computing in Civil Engineering*, vol. 30, no. 5, p. 04016003, 2016.

[57] Y. Lin, K. Zhou, and J. Li, "Application of cloud model in rock burst prediction and performance comparison with three machine learning algorithms," *IEEE Access*, vol. 6, pp. 30958–30968, 2018.

[58] Z. Tang, J. C. Ding, and X.-X. Zhai, "Effects of resveratrol on the expression of molecules related to the mTOR signaling pathway in pathological scar fibroblasts," *Giornale Italiano di Dermatologia e Venereologia*, vol. 155, no. 2, pp. 161–167, 2020.

[59] Y. Zhou and T. Wang, "PNN-based rockburst prediction model and its applications," *Earth Sciences Research Journal*, vol. 21, no. 3, pp. 141–146, 2017.