Hindawi

Research Article

Hybrid Machine Learning Model for Electricity Consumption Prediction Using Random Forest and Artificial Neural Networks

Witwisit Kesornsit[1] and Yaowarat Sirisathitkul [2]

[1]*Government Data Solution Division, Department of Data Solution,*
 Digital Government Development Agency (Public Organization), Bangkok, Thailand
[2]*Department of Computer Engineering and Electronics, School of Engineering and Technology, Walailak University,*
 Nakhon Si Thammarat, Thailand

Correspondence should be addressed to Yaowarat Sirisathitkul; kinsywu@gmail.com

Predicting electricity consumption is notably essential to provide a better management decision and company strategy. This study presents a hybrid machine learning model by integrating dimensionality reduction and feature selection algorithms with a backpropagation neural network (BPNN) to predict electricity consumption in Thailand. The predictive models are developed and tested using an actual dataset with related predictor variables from public sources. An open geospatial data gathered from a real service as well as geographical, climatic, industrial, household information are used to train, evaluate, and validate these models. Machine learning methods such as principal component analysis (PCA), stepwise regression (SWR), and random forest (RF) are used to determine the significant predictor variables. The predictive models are constructed using the BPNN with all available variables as baseline for comparison and selected variables from dimensionality reduction and feature selection methods. Along with creating a predictive model, the most related predictors of energy consumption are also selected. From the comparison, the hybrid model of RF with BPNN consistently outperforms the other models. Thus, the proposed hybrid machine learning model presented from this study can predict electricity consumption for planning and managing the energy demand.

1. Introduction

Due to the pandemic crisis in 2020, total energy demand during 2020–2030 is likely to be higher than the International Energy Agency (IEA) forecast. Electricity consumption has become one of the critical issues in most countries. Therefore, the accurate prediction of electricity consumption has an essential role in achieving efficient energy utilization. Assessment of the electricity consumption in advance will improve operation strategies and management of energy storage system and planning activities for future power plants [1, 2]. One of the largest electricity consumers is the business sector. Generally, socioeconomic and environmental factors contribute to electricity consumption. Socioeconomic factors include industrial and household information, while environmental factors include geographical and climatic information. Determining the significant relation of different factors related to electricity consumption could provide guidelines for electricity authority management to carry out the planning and strategies in an efficient manner.

Over the past several decades, many statistical and computational intelligence methods have been implemented in the fields of prediction. Previous studies are mainly limited to a small dataset of independent variables based on time-series forecasting, regression analysis, and clustering methods [3, 4]. The statistical methods have some restrictions on the linearity, normality, and independence of variables [5]. Computational intelligence methods, such as artificial intelligence algorithms, have been primarily implemented in prediction [6]. Recent research shows that the super computing power is more efficient and effective in

handling and analyzing huge volumes of data. The machine learning algorithms, a subset of artificial intelligence, exhibit superior performance in handling large numbers of data [7].

The majority of research in electricity consumption modeling uses various machine learning methods, in particular artificial neural network (ANN), decision tree, and clustering. Platon et al. [1] performed research on ANN concerning the hourly prediction of electricity consumption. Walker et al. [8] predicted the energy consumption of the building using random forest (RF) and ANN. Pérez-Chacón et al. [9] proposed a methodology for finding patterns of electricity consumption using the k-means. Shi et al. [10] employed echo state network in the prediction of building energy demand. Furthermore, the literature review on machine learning algorithms in energy research has been proposed by Mohandes et al. [11] and Lu et al. [12].

Moreover, the current research of predicting the future electricity demand related to hybrid models combining different two or more machine learning algorithms was reviewed and analyzed by Deb et al. [13] and Mamun et al. [14]. The hybrid models show excellent results because individual algorithms have different advantages. Zekić-Sušac et al. [4] integrated the variable selection with ANN in developing a predictive model for the energy cost of public buildings in Croatia. Zekić-Sušac et al. [15] integrated clustering and ANN to improve the accuracy of modeling energy efficiency. Muralitharan et al. [16] proposed a convolutional neural network based optimization approach for predicting the future energy demand. Pérez-Chacón et al. [17] used a big data time-series and experimental method with decision tree, gradient boosting machine (GBM), pattern sequence-based forecasting, ARIMA, and ANN to forecast the electricity demand. Zekić-Sušac et al. [18] used variable reduction procedures with RPart regression tree, RF, and deep neural networks to construct a predictive model for the energy demand of public buildings in Croatia. Basurto et al. [19] employed a hybrid intelligent system based on ANN and clustering algorithm to predict the solar energy in Spain. Therefore, it is appropriate to exploit machine learning algorithms in electricity consumption prediction.

This study has demonstrated the prediction of electricity consumption. The proposed procedure has several phases: data collection and data preprocessing, dimensionality reduction and feature selection, and prediction. In the data collection and data preprocessing phase, the data is collected from publicly available sources and processed to handle the missing values and outliers. In the dimensionality reduction and feature selection phase, the techniques including principal component analysis (PCA), stepwise regression (SWR), and RF are applied. The prediction phase is implemented by using the backpropagation neural networks (BPNN) algorithm. The selected important predictor variables from PCA, SWR, and RF have been used as the inputs for the BPNN algorithm to predict the electricity consumption. Besides creating a predictive model, the subset of relevant variables is also selected and compared. Six metrics evaluate the effectiveness of the predictive models. The model with the highest accuracy in the test evaluation has been selected.

The other sections of this paper are as follows. Section 2 introduces the machine learning algorithms employed in this study and exploration of the relevant literature. Section 3 outlines the architecture of the proposed predictive models, experimental dataset, and the evaluation method. Section 4 describes the experimental outcome and some statistical results. Finally, Section 5 gives the paper achievement as well as the conclusions and future works.

2. Literature Review

Machine learning is an algorithm to construct empirical models from the dataset and is categorized as data-driven modeling requiring a sufficient quantity of historical data to predict future demand reliably [8]. Machine learning algorithms extract essential information presented in large amounts of the recorded data, thereby achieving better performance and accuracy [7, 20]. Systematic literature reviews of artificial intelligence and machine learning algorithms are provided by Duan et al. [21], Borges et al. [22], and Dwivedi et al. [23]. It is accepted that these techniques bring a significant impact and new research frameworks in industries such as finance, medicine, manufacturing, and various government, public sector, and business domains. For example, the machine learning methodology is employed in the prediction of crime [24], future price of agricultural products [25], natural gas consumption [26], commercial banks performance [27], landslide displacement [28], and seawater evaporation [29]. As may be seen, prediction algorithms have been extensively investigated in several sectors.

Constructing the predictive models by employing a lot of variables is not straightforward in practice. All these variables might not be completely collected in a real-world situation and result in a more complex model. The prominence of dimensionality reduction and feature selection of modeling variables have been broadly revealed in [30]. The PCA is the most common feature extraction method to reduce the dimensionality of large dataset into a small dataset that retains most of the information [31]. The PCA uses the eigenvalue and eigenvector for projecting the high-dimensional dataset on to a lower-dimensional space. It converts a set of correlated variables into a set of principal components. Only the principal components that can sustain the most original variance will be extracted [32, 33]. Dimensionality reduction with the PCA is applied in many domains such as electricity consumption [1], finance [31], engineering [34], and agriculture [35, 36]. The disadvantage of PCA is that the predictor variables become less interpretable and have no corresponding physical meaning and this makes it more challenging to determine the predictor variables that are important in the predictive model [35, 37].

Feature selection is a process to select the features that contain the most useful information while discarding redundant features that contain little to no information. The wrapper feature selection algorithm such as SWR and RF is a method that depends on the accuracy of the subsequent feature selection criterion. The SWR is a statistical method of fitting regression models and has the advantage of evading

collinearity [38]. It defines appropriate variable subsets and evaluates variables priorities [39]. Selection of predictor variables is performed automatically by assessing the relative importance of the variables based on prespecified criteria such as the F-test, the t-test, the adjusted R^2, and the Akaike information criterion (AIC).

The RF is a supervised machine learning algorithm that is effective and efficient for both classification and prediction. It is based on a decision tree algorithm and classified as a bagging ensemble learning method [24, 40]. The RF structure is composed of multiple decision trees, and then each RF tree runs in a parallel manner to each other. During the variable randomization in each iteration, a variable importance index and the Gini index can be given [41]. The final value is evaluated by aggregating the results from all leaves of each tree [35, 42]. The RF is also one of the best algorithms for estimating the importance of variables and is applied in various fields [20, 43–45]. Furthermore, the RF is an excellent prediction algorithm and has the advantages of its generalization and a good balance of error [11, 46, 47]. Some analysts employed both SWR and RF to select the input variables or analyze the importance of variables in domains such as the electronic industry [5], geographical poverty [45], reservoir characterisation [48], and soil carbon [49]. They stated that RF is better than SWR in identifying nonlinear relationships between variables.

The ANN is classified as a supervised learning method and also deployed in the comparison of prediction performance with other machine learning techniques [50–55]. Among the ANN, BPNN is one of the most widely used technique to optimize the feedforward neural network. The backpropagation algorithm is a broadly used technique and a standard method for training the weights in a multilayer feedforward neural network through a chain rule method [56]. The weights of a neural net are appropriately adjusted based on the loss in the previous iteration. Therefore, this results in a lower error rate, making the model more reliable by enhancing a generalization. Researchers have applied BPNN in many classifications and predictions. As an example, BPNN is employed in agricultural product sales [57], crude oil future price [58], and hybrid cement [59] and also deployed in the comparison of prediction performance with other machine learning techniques [50–52].

3. Materials and Methods

The outline of the proposed framework is shown in Figure 1. The hybrid predictive model comprises three stages conducting in sequence. The first stage explores the exploratory data analysis. The second stage uses the PCA, SWR, and RF methods to select suitable predictor variables. The third stage is to establish the predictive model by constructing the BPNN. The developed models are trained with 10-fold cross-validation and evaluated. All models have been implemented and tested on intel®Core™ i7-8550U, CPU @1.80 GHz, 1.99 GHz running, 64 bit Windows 10 operating system with 8 GB RAM. The Scikit-learn machine learning package for the Python programming language is used to implement the models with Python version 3.9.4. Many algorithms
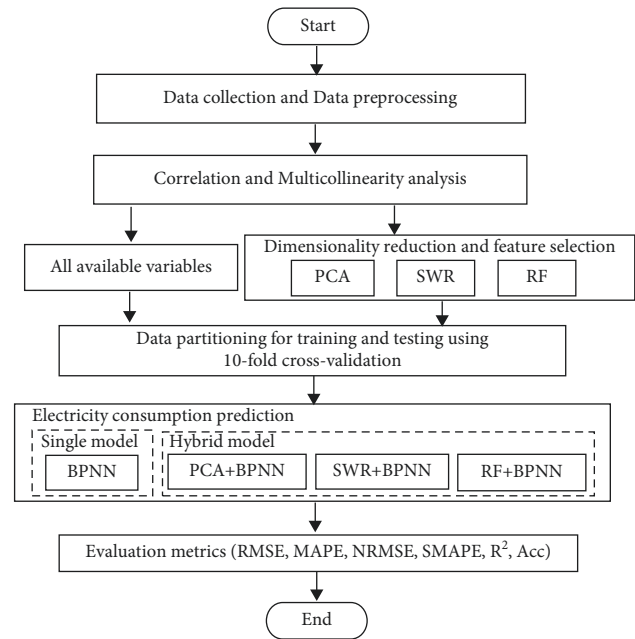

FIGURE 1: The framework of the hybrid machine learning model for electricity consumption prediction.

configuration parameters are set to the defaults of Scikit-learn version 0.23.0, Numpy version 1.20.2, and Pandas version 1.2.3.

3.1. Data Collection. The real data samples were collected from publicly available online sources from the beginning of 2018 to the end of 2019. The dataset contained 884,736 records of monthly electricity consumption. The 21 predictor variables were grouped into five categories, namely, geospatial, geographical, climatic, industrial, and household factors. The electricity consumption and geospatial factor were obtained from the official website of the Provincial Electricity Authority of Thailand [60]. The geographical, industrial, and household factors were obtained from the official website of the National Statistical Office of Thailand [61]. The climatic factor was obtained from the official website of the Thai Meteorological Department [62].

3.2. Data Preprocessing. The geospatial factor represented four categorical variables. Firstly, the electrical substation was the type of electrical distribution substations covering the four regions of Thailand (South, North, Northeast, and Central). Each region has three areas; thus, this variable can take on 12 different values (0–11). Secondly, the business type belonged to eight types of business (0–7): small residential houses, large residential houses, small business, medium business, large business, specific activities, government, and agriculture. Thirdly, the time of use was categorized into three periods: peak day (2: 00 p.m.–7: 00 p.m.), semipeak (5: 00 a.m.–2: 00 p.m., 7: 00 p.m.–12: 00 a.m.), and offpeak (12: 00 a.m.–5: 00 a.m.) as suggested by Yang et al. [63]. Finally, periods in Thailand are grouped into

three seasons: summer (February–May), rainy (June–October), and winter (November–January).

Data quality is a key success in model development since poor data quality can negatively impact model accuracy. To perform the analysis, it is vital to identify outliers which are measurement results far from other values. Hence, they are not a representative of the majority of data. The outliers are the minimum amount of electricity consumption in the range of 0.00–0.09 W from the specific activities, government, and agriculture sectors and are subsequently removed from the dataset. Only 231 records of the available data are considered as the outliers, and the final modeling dataset contained 884,505 records. The geographical, climatic, industrial, and household factors are numerical values on a monthly basis in each province. According to the geospatial factor, each variable is averaged monthly. The 21 predictor variables and one target variable are shown in Table 1 with their descriptive statistics.

3.3. Reduction of Modeling Inputs.

The variable reduction as inputs is a critical issue in a successful predictive model. Theoretically, a model should be built with a small number of relevant inputs to achieve an acceptable level of predictive accuracy [1].

3.3.1. Principle Component Analysis.

All 21 predictor variables are processed by the PCA for finding the principal components in the dimensionality reduction. The StandardScaler function in the Scikit-learn Python library is used to standardize the 21 predictor variables onto a unit scale. Therefore, all the normalized predictor variables have a mean of zero and a standard deviation of one.

3.3.2. Stepwise Regression.

By dropping the correlated predictor variables, the 15 predictor variables are retained. All remaining variables are entered or removed from the regression equation of the SWR model one by one. When each predictor variable is entered, a selection is adopted based on the AIC to remove redundant variables. This process is repeated until no significant predictor variable is entered into or removed from the regression equation.

3.3.3. Random Forest.

The RF is used for evaluating the importance of predictor variables. The RF model is implemented using an ensemble of 1000 trees, and the number of trees was determined by trial and error. A typical split criterion is the mean square error (MSE) between the target and predicted output in a node. The 8-maximum depth of the tree was used in model construction.

3.4. BPNN Predictive Model Development.

Before developing the models, the data are divided into training and testing subsamples. The training subsample is used for constructing the model, while the testing subsample is used to determine the model efficiency. The sample data presented in Table 2 indicates the number of samples in the training group (70% of samples) and the testing group (30% of samples). This division ratio is recommended by Zekić-Sušac et al. [4, 18]. In the training stage, the 10-fold cross-validation is also applied to reduce the overfitting problem and provides more reliable and unbiased models. The training dataset is divided randomly into 10-fold, and all models are evaluated 10 times. The cross-validation technique is implemented because it will make the model more reliable for new unseen data [33]. Lastly, the final assessment of predictive accuracy is evaluated with the remaining unseen 30% of data in the testing groups.

A feedforward ANN and backpropagation training method are chosen for developing the predictive model for electricity consumption. The result of this predictive model is the value of the target variable, which indicates the forecasted electricity consumption. The six BPNN models are developed using all available variables and selected variables previously reduced by PCA, SWR, and RF methods. This study tests the architectures of BPNN with two or three hidden layers, as suggested by Zekić-Sušac et al. [4, 18]. The structure of six BPNN predictive models is indicated in Table 3. All six BPNN models have only one node in output layer, that is, the electricity consumption value obtained from the predictive model. The rectified linear unit function (ReLU) is utilized as the activation function to define the output of that node. The BPNN is entered with the training subsample, and the learning rating is 0.001. The stopping criterion for the training process is set where either the epochs reach 1,000 or the training goal is reached.

3.5. Performance Evaluation Metrics.

For assessing the performance of all predictive models, the following statistical indicators have been computed: coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and predictive accuracy (Acc) according to Walker et al. [8], Pérez-Chacón et al. [17], Qiao et al. [26], Chen et al. [43], and Li et al. [46]. In a model with lower RMSE and MAPE, higher R^2 indicates better accuracy. These metrics can be formulated as follows:

$$R^2 = \left[\frac{\sum_{i=1}^{N} (y_{ti} - \overline{y_t})(y_{ci} - \overline{y_c})}{\sqrt{\sum_{i=1}^{N} (y_{ti} - \overline{y_t})^2 \times \sum_{i=1}^{N} (y_{ci} - \overline{y_c})^2}} \right],$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_{ti} - y_{ci})^2}{N}},$$

$$\text{MAPE} = 100 \frac{1}{N} \sum_{i=1}^{N} \frac{|y_{ti} - y_{ci}|}{y_{ti}},$$

$$\text{Acc} = 1 - \text{abs} \left(\text{mean} \frac{y_c - y_t}{y_t} \right).$$

(1)

The other evaluation metrics are symmetric mean absolute percentage error (SMAPE) and normalized root means square error (NRMSE) according to Zekić-Sušac et al. [4, 18] and Janković et al. [53]. A model with lower NRMSE

TABLE 1: Descriptive statistics of variables for the electricity consumption prediction.

Group of variables	Variable description	Variable code	Descriptive statistics
Geospatial	Electrical substation (0–2 = south area, 1–3 3–5 = north area, 1–3 6–8 = northeast area, 1–3 9–11 = central area 1–3)	Electrical_Substation	Categorical, 0 = 8.35%, 1 = 8.35, 2 = 8.35%, 3 = 12.52%, 4 = 8.35%, 5 = 8.35%, 6 = 8.40%, 7 = 8.35%, 8 = 8.22%, 9 = 8.35%, 10 = 4.07%, 11 = 8.35%
	Business type (0–1 = small, large residential houses, 2–4 = small, medium, large business, 5 = specific activities, 6 = government, 7 = agriculture)	Usage_Type	Categorical, 0 = 13.02%, 1 = 12.91%, 2 = 12.05%, 3 = 12.20%, 4 = 12.00%, 5 = 12.00%, 6 = 12.78%, 7 = 13.04%
	Time of use (0 = peakday, 1 = semipeak, 2 = offpeak)	TOU	Categorical, 0 = 12.50%, 1 = 66.67%, 2 = 20.83%
	Thailand season period (0 = summer, 1 = rainy, 2 = winter)	Season	Categorical, 0 = 24.50%, 1 = 42.94%, 2 = 32.56%
Geographical	Total number of populations	Population_N	Numerical, min = 2810387.00, max = 7871210.00, mean = 4965500.09, st. dev = 1641653.19
	Total surface area	Area	Numerical, min = 22423.00, max = 72806.35, mean = 42012.25, st. dev = 16950.41
	Ratio of people and area per sq. km	Population_Ratio	Numerical, min = 445.00, max = 1860.00, mean = 1003.87, st. dev = 369.42
Climatic	Mean station pressure	Mean_Station_Pressure	Numerical, min = 975.23, max = 1010.86, mean = 999.44, st. dev = 9.18
	Mean msl pressure	Mean_MSL_Pressure	Numerical, min = 1004.46, max = 1014.20, mean = 1008.85, st. dev = 2.50
	Mean maximum temperature	Mean_Maximum_Temperature	Numerical, min = 27.92, max = 36.62, mean = 32.15, st. dev = 1.71
	Mean minimum temperature	Mean_Minimum_Temperature	Numerical, min = 14.00, max = 25.36, mean = 22.24, st. dev = 2.56
	Mean drybulb temperature	Mean_Drybulb_Temperature	Numerical, min = 20.46, max = 29.87, mean = 26.59, st. dev = 1.84
	Mean relative humidity	Mean_Relative_Humidity	Numerical, min = 57.00, max = 85.73, mean = 75.56, st. dev = 6.36
	Total rainfall	Total_Rainfall	Numerical, min = 11.00, max = 4841.00, mean = 1264.52, st. dev = 1096.77
Industrial	Total number of industrial labors	Industrial_Labor	Numerical, min = 1471.00, max = 21368.00, mean = 7087.11, st. dev = 6573.51
	Total number of industrial plants	Industrial_Plant	Numerical, min = 101.00, max = 727.00, mean = 263.75, st. dev = 167.60
	Total number of agriculturists	Agriculturists	Numerical, min = 197885.00, max = 1412997.00, mean = 628382.35, st. dev = 375015.78
Household	Average expenditure per household	Expenditure	Numerical, min = 63795.00, max = 168442.00, mean = 112928.45, st. dev = 28137.06
	Average income per household	Income	Numerical, min = 81571.00, max = 206231.00, mean = 143155.35, st. dev = 34841.15
	Total number of households	Household	Numerical, min = 861.00, max = 1984.00, mean = 1479.48, st. dev = 318.94
	Average liabilities per household	Liabilities	Numerical, min = 383606.00, max = 1536859.00, mean = 1005455.20, st. dev = 316864.28
Target variable	Electricity consumption	Demand_KW	Numerical, min = 0.10, max = 8731298.26, mean = 167920.67, st. dev = 358252.18

TABLE 2: Sampling procedures used for modeling.

Sample	No. of cases and % in the subsample	10-fold cross-validation
Training	619,154 cases (70%)	Train: 557,238 cases (90% of learning subsample); cross-validation: 61,915 cases (10% of learning subsample)
Test	265,352 cases (30%)	
Total	884,505 cases (100%)	

TABLE 3: The parameter structure of the BPNN predictive models.

Predictive model	No. of nodes in input layer	No. of hidden layers	No. of nodes in hidden layer
BPNN	21	3	12, 8, 2
PCA # 1 + BPNN	9	2	8, 2
PCA # 2 + BPNN	13	2	8, 2
SWR + BPNN	12	2	8, 2
RF # 1 + BPNN	18	3	12, 8, 2
RF # 2 + BPNN	11	2	8, 2

and SMAPE indicates greater accuracy in prediction. These metrics can be formulated as follows:

$$\text{SMAPE} = 100 \frac{1}{N} \sum_{i=1}^{N} \frac{|y_t - y_c|}{|y_t| + |y_c|},$$

$$\text{NRMSE} = \frac{\text{RMSE}}{y_{t\max} - y_{t\min}} = \frac{\sqrt{\left(\sum_{i=0}^{N}(y_{ti} - y_{ci})^2/N\right)}}{y_{t\max} - y_{t\min}}.$$
(2)

All parameters are explained as follows: y_t is the target (real) output, y_c is the calculated output, $\overline{y_t}$ is an average of the target output, and $\overline{y_c}$ is an average of the calculated output and N is the total number of measurements.

4. Results and Discussion

4.1. Correlation and Multicollinearity Analysis. The Pearson correlation coefficient (r_p) is used to analyze the correlation between two numerical variables, and Spearman's rank correlation coefficient (r_s) is utilized to evaluate the correlation in categorical variables. The correlation coefficients indicated the strength, and direction of association between all electricity consumption variables is shown in Figure 2.

For the multicollinearity analysis, the VIF of all predictor variables is also computed. The presence of multicollinearity implies that the variable provides redundant information contained in other variables [3]. The available variables in the modeling stage are determined by considering the correlation and VIF between the pair variables. Therefore, six predictor variables, namely, *Electricity_Substation, Season, Population_N, Mean_Minimum_Temperature, Agriculturists,* and *Expenditure,* are removed from the 21 predictor variables. As a result, the 15 remaining predictor variables are used as the input of SWR.

4.2. Selecting Predictor Variables

4.2.1. Principle Component Analysis. The 21 predictor variables are sent as the input of PCA. Since PCA works on numerical variables, four categorical variables are converted into numerical variables using the one-hot encoding technique. This study uses two experiments, namely, PCA # 1 and PCA # 2, by using the cumulative contribution rate of principal components as shown in Figure 3. For PCA # 1 and PCA # 2, the principal components with the cumulative contribution rate reach 95% and 99%, respectively. Therefore, the first 9 and 13 principal components are considered to be significant for PCA # 1 and PCA # 2 and used as variables in the predictive modeling stage.

4.2.2. Stepwise Regression. The 15 predictor variables from the correlation and multicollinearity analysis stage are used as the input for the SWR. Since the *Usage_Type* and *TOU* variables are categorical variables, these variables are set as dummy variables to trick the SWR algorithm into correctly analyzing variables. Consequently, the input variables in the SWR estimation parameter consist of 22 variables. Variables selected by SWR are summarized in Table 4, with the VIF metrics calculated per variable. According to the VIF value, the 11 selected variables are essential in modeling the electricity consumption. Geospatial variables are highly correlated with electricity consumption, among which *TOU* is the most important variable, followed by *Usage_Type*.

4.2.3. Random Forest. The importance of 21 predictor variables by the RF model is listed in Table 5. From the results, the four most important variables according to the percentage increase in mean squared error (%IncMSE) are *Usage_Type, Area, Mean_Station_Pressure,* and *Industrial_Labor*. The variables selected by RF have two experiments: RF # 1 with 18 selected predictor variables having importance over 0.01 and RF # 2 with 11 variables having the importance over 0.02.

4.3. Prediction Evaluation. The results of six machine learning models are shown in Table 6. The predictive performance is evaluated using six metrics, namely, RMSE, MAPE, NRMSE, SMAPE, R^2, and Acc. The single model is developed with all available 21 predictor variables (BPNN).
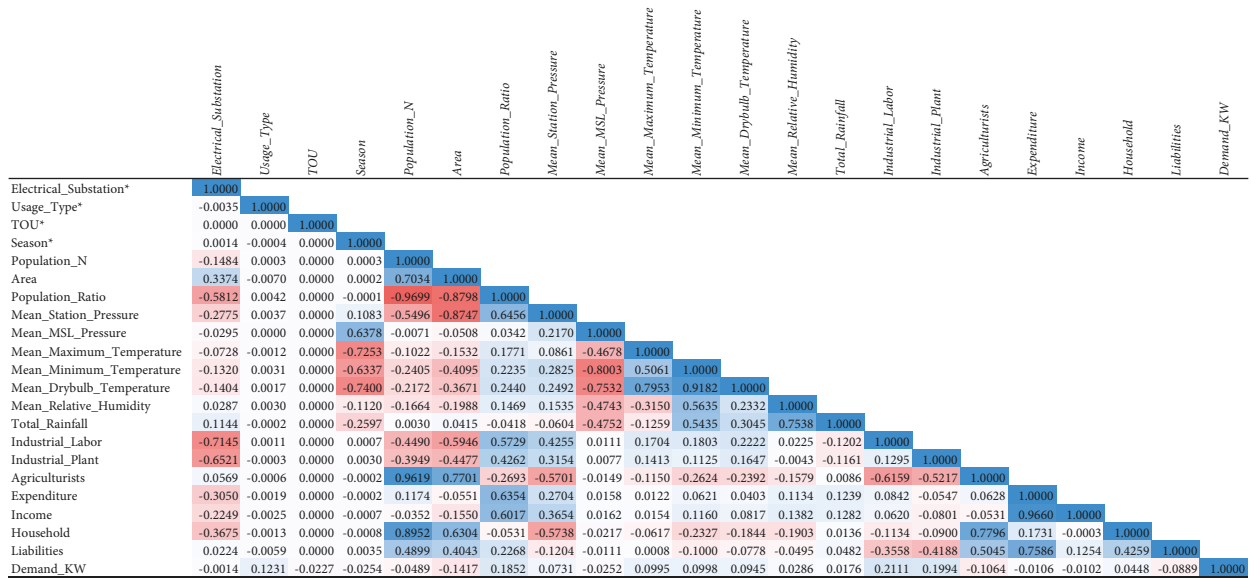
	Electrical_Substation	Usage_Type	TOU	Season	Population_N	Area	Population_Ratio	Mean_Station_Pressure	Mean_MSL_Pressure	Mean_Maximum_Temperature	Mean_Minimum_Temperature	Mean_Drybulb_Temperature	Mean_Relative_Humidity	Total_Rainfall	Industrial_Labor	Industrial_Plant	Agriculturists	Expenditure	Income	Household	Liabilities	Demand_KW
Electrical_Substation*	1.0000																					
Usage_Type*	-0.0035	1.0000																				
TOU*	0.0000	0.0000	1.0000																			
Season*	0.0014	-0.0004	0.0000	1.0000																		
Population_N	-0.1484	0.0003	0.0000	0.0003	1.0000																	
Area	0.3374	-0.0070	0.0000	0.0002	0.7034	1.0000																
Population_Ratio	-0.5812	0.0042	0.0000	-0.0001	-0.9699	-0.8798	1.0000															
Mean_Station_Pressure	-0.2775	0.0037	0.0000	0.1083	-0.5496	-0.8747	0.6456	1.0000														
Mean_MSL_Pressure	-0.0295	0.0000	0.0000	0.6378	-0.0071	-0.0508	0.0342	0.2170	1.0000													
Mean_Maximum_Temperature	-0.0728	-0.0012	0.0000	-0.7253	-0.1022	-0.1532	0.1771	0.0861	-0.4678	1.0000												
Mean_Minimum_Temperature	-0.1320	0.0031	0.0000	-0.6337	-0.2405	-0.4095	0.2235	0.2825	-0.8003	0.5061	1.0000											
Mean_Drybulb_Temperature	-0.1404	0.0017	0.0000	-0.7400	-0.2172	-0.3671	0.2440	0.2492	-0.7532	0.7953	0.9182	1.0000										
Mean_Relative_Humidity	0.0287	0.0030	0.0000	-0.1120	-0.1664	-0.1988	0.1469	0.1535	-0.4743	-0.3150	0.5635	0.2332	1.0000									
Total_Rainfall	0.1144	-0.0002	0.0000	-0.2597	0.0030	0.0415	-0.0418	-0.0604	-0.4752	-0.1259	0.5435	0.3045	0.7538	1.0000								
Industrial_Labor	-0.7145	0.0011	0.0000	0.0007	-0.4490	-0.5946	0.5729	0.4255	0.0111	0.1704	0.1803	0.2222	0.0225	-0.1202	1.0000							
Industrial_Plant	-0.6521	-0.0003	0.0000	0.0030	-0.3949	-0.4477	0.4262	0.3154	0.0077	0.1125	0.1125	0.1647	-0.0043	-0.1161	0.1295	1.0000						
Agriculturists	0.0569	-0.0006	0.0000	-0.0002	0.9619	0.7701	-0.2693	-0.5701	-0.0149	-0.1150	-0.2624	-0.2392	-0.1579	0.0086	-0.6159	-0.5217	1.0000					
Expenditure	-0.3050	-0.0019	0.0000	-0.0002	0.1174	-0.0551	0.6354	0.2704	0.0158	0.0122	0.0621	0.0403	0.1134	0.1239	0.0842	-0.0547	0.0628	1.0000				
Income	-0.2249	-0.0025	0.0000	-0.0007	-0.0352	-0.1550	0.6017	0.3654	0.0162	0.0154	0.1160	0.0817	0.1382	0.1282	0.0620	-0.0801	-0.0531	0.9660	1.0000			
Household	-0.3675	-0.0013	0.0000	-0.0008	0.8952	0.6304	-0.0531	-0.5738	-0.0217	-0.0617	-0.2327	-0.1844	-0.1903	0.0136	-0.1134	-0.0900	0.7796	0.1731	-0.0003	1.0000		
Liabilities	0.0224	-0.0059	0.0000	0.0035	0.4899	0.4043	0.2268	-0.1204	-0.0111	0.0008	-0.1000	-0.0778	-0.0495	0.0482	-0.3558	-0.4188	0.5045	0.7586	0.1254	0.4259	1.0000	
Demand_KW	-0.0014	0.1231	-0.0227	-0.0254	-0.0489	-0.1417	0.1852	0.0731	-0.0252	0.0995	0.0998	0.0945	0.0286	0.0176	0.2111	0.1994	-0.1064	-0.0106	-0.0102	0.0448	-0.0889	1.0000

FIGURE 2: Correlation matrix between the electricity consumption variables (*Spearman's rank correlation (r_s)).
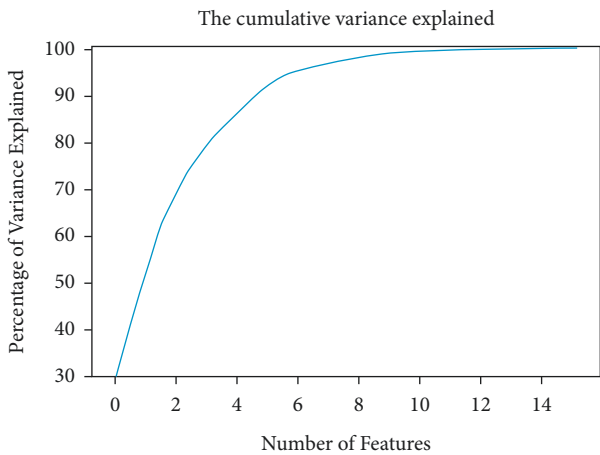

FIGURE 3: The cumulative contribution rate of principal components.

The hybrid models are constructed with the principal components extracted by the PCA method (PCA # 1 + BPNN and PCA # 2 + BPNN) and with the reduced variables selected by the SWR and RF methods (SWR + BPNN, RF # 1 + BPNN, and RF # 2 + BPNN).

It can be seen from Table 6 that the most successful is the RF # 1 + BPNN with 18 selected variables, producing R^2 of 0.9932 and Acc of 0.9926. The RF # 2 + BPNN has a slightly lower R^2 of 0.9630 and Acc of 0.9923, while the BPNN has R^2 of 0.9638 and Acc of 0.9931. In summary, the hybrid model is the most effective in electricity consumption prediction, as Deb et al. [13] and Mamun et al. [14] show.

In the SMAPE assessment, using all available variables (BPNN) results in a SMAPE of 21.0117%. The PCA # 1 + BPNN and PCA # 2 + BPNN produce the higher SMAPE of 28.0458% and 22.0321%, respectively, while the lowest SMAPE is obtained by the SWR + BPNN (21.0112%). Although the accuracy of the most successful model is not satisfactory (SMAPE is 21.0112%), comparing to Walker et al. [8] and Zekić -Sušac et al. [18], whose models obtained an accuracy below 20%, it is greater than the model obtained by Zekić -Sušac et al. [4] (SMAPE is 22.3555%).

In the case of using the dimensionality reduction method, PCA # 1 + BPNN produced RMSE of 0.0601 and the PCA # 2 + BPNN produced RMSE of 0.0405. This means that the RMSE value decreases with the number of principal components increasing as stated by Zhang [34] and in accordance with a decrease in the number of principal components leading to a decrease in predictive accuracy [34]. The error values show a slight improvement in relation to BPNN only when the 13 principal components are used (RMSE = 0.0405 in PCA # 2 + BPNN comparing to RMSE = 0.0415 in BPNN). At the same time, the performance of other hybrid models does not improve in terms of RMSE.

The RF selects the significant modeling inputs among 21 available inputs. As a result, 18 significant variables and 11 significant variables are selected and used in the model of RF # 1 + BPNN and RF # 2 + BPNN, respectively. The RF # 1 + BPNN model slightly outperforms the RF # 2 + BPNN. This means that using more relevant variables mainly presents superior prediction accuracy [35]. Moreover, using a smaller number of variables does not make a significant difference of the model prediction performance while minimizing the computational cost. A comprehensive comparison shows that most accuracy comes from the RF # 1 + BPNN (0.9926), followed by RF # 2 + BPNN (0.9923) and BPNN (0.9631). This result suggests that the hybrid model using fewer inputs is slightly more accurate. This also suggests that irrelevant variables bring meaningless information to the input dataset and add unnecessary variability and noise and hinder the ability to accurately model [1]. The RF can model nonlinearity relations between variables and has high accuracy, as investigated by Richardson et al. [49]. The results of this study indicate that the proposed hybrid

TABLE 4: Feature importance based on the SWR model.

Group of variables	Variable code		Coefficients	SE coef	t-value	p-value	VIF
		Agriculture	$-3.724E+04$	$1.144E+03$	-32.556	$<2E-16$	1.7382
		Government	$-3.755E+04$	$1.146E+03$	-32.763	$<2E-16$	1.7332
		House < 150	$4.953E+04$	$1.169E+03$	42.378	$<2E-16$	1.6988
	Usage_Type (dummy variable)	House > 150	$2.475E+05$	$1.164E+03$	212.582	$<2E-16$	1.7033
Geospatial		Large	$5.338E+05$	$1.170E+03$	456.169	$<2E-16$	1.6971
		Medium	$1.781E+05$	$1.170E+03$	152.206	$<2E-16$	1.6966
		Small	$1.029E+05$	$1.149E+03$	89.538	$<2E-16$	1.7275
	TOU (dummy variable)	Semipeak	$1.687E+04$	$7.327E+02$	23.031	$<2E-16$	1.4001
		Peakday	$2.737E+04$	$1.044E+03$	26.210	$<2E-16$	1.4000
Geographical	Population_Ratio		$-1.330E+02$	$1.417E+00$	-93.872	$<2E-16$	3.2135
	Mean_MSL_Pressure		$1.585E+04$	$2.453E+02$	64.603	$<2E-16$	4.4071
Climatic	Mean_Maximum_ temperature		$2.438E+04$	$3.329E+02$	73.236	$<2E-16$	3.7922
	Mean_Relative_ humidity		$7.057E+03$	$1.073E+02$	65.774	$<2E-16$	5.4626
	Total_Rainfall		$1.114E+01$	$4.604E$-01	24.193	$<2E-16$	2.9923
Industrial	Industrial_Labor		$1.626E+01$	$1.460E$-01	111.316	$<2E-16$	9.8164
	Industrial_Plant		$-3.204E+01$	$5.090E+00$	-6.294	$3.09E-10$	8.5413
Household	Household		$1.059E+02$	$1.144E+00$	92.641	$<2E-16$	1.5609
	Liabilities		$3.041E-02$	$1.403E-03$	21.665	$<2E-16$	2.3207

TABLE 5: Feature importance based on the RF model.

Variable code	Variable importance (% IncMSE)
Usage_Type	**0.4000**
Area	**0.1600**
Mean_Station_Pressure	**0.0800**
Industrial_Labor	**0.0800**
Household	**0.0400**
Mean_Relative_Humidity	**0.0300**
Mean_Maximum_Temperature	**0.0200**
Mean_Minimum_Temperature	**0.0200**
Mean_Drybulb_Temperature	**0.0200**
Expenditure	**0.0200**
Income	**0.0200**
Season	**0.0100**
Mean_MSL_Pressure	**0.0100**
Total_Rainfall	**0.0100**
Agriculturists	**0.0100**
Industrial_Plant	**0.0100**
Liabilities	**0.0100**
Population_Ratio	**0.0100**
TOU	0.0000
Population_N	0.0000
Electrical_Substation	0.0000

RF + BPNN model performs excellently in terms of predictive accuracy.

By comparing the hybrid model results using dimensionality reduction and feature selection, the RF # 1 + BPNN model gives the remarkable and insightful performance. In this study, the results of two experiments of PCA + BPNN models do not achieve good performance. This is likely due to the correlation and multicollinearity among predictor variables. The SWR + BPNN seems to have a comparable performance with a BPNN but is lower than the RF + BPNN. Smith [38] suggested that the SWR is less effective in the case of a large number of possible predictor variables. The results indicated that the RF has higher accuracy than the SWR and is more powerful in identifying the nonlinear relationships

between target and predictor variables, as reported by Liu et al. [45].

The results of the experiments suggest that the PCA and SWR cause a loss of predictive accuracy. In addition, the predictive models developed by using the RF-selected models outperform the models using all available variables. It reveals that the RF has sufficient complexity to reduce a high-dimensional dataset.

Another focus of this study is to investigate the abilities of the feature selection methods. The selected variables by SWR, RF#1, and RF#2 are listed in Table 7. The selected variables describe geospatial, geographical, climatic, industrial, and household factor. This could be interpreted that all groups of variables play an essential role in predicting electricity consumption. There are some differences between the selected variable in each variable group. Even if the redundant and irrelevance variables are discarded, the correlation between the rest of the variables still exists.

The identically identified variables by SWR, RF # 1 and RF # 2 are *Usage_Type, Mean_Maximum_Temperature, Mean_Relative_Humidity, Industrial_Labor*, and *Household*. Therefore, these variables are considered to be the critical variables in explaining electricity consumption. Among the industrial factor group of variables, all three models have selected *Industrial_Labor*, while the SWR and RF#1 have selected *Industrial_Plant*. The largest number of selected variables represents the climatic factor group of variables. This result confirms the findings of the previous study by Walker et al. [8], Deb et al. [13], and Zekić -Sušac et al. [18]. Among all variables of the climatic factor, the temperature is reported to be the most significant variable influenced by the electricity consumption [3].

The hybrid predictive model in this study successfully integrates the existing machine learning algorithms and the selection of predictor variables in the electricity consumption is demonstrated. The predictive model is constructed from the actual electricity consumption data during 2018–2019 from the Provincial Electricity Authority of

TABLE 6: Performance of proposed predictive models on the test sample.

Predictive model	No. of nodes in input layer	RMSE	MAPE	NRMSE	SMAPE	R^2	Acc
BPNN	21	0.0415	8.1999	**0.0458**	21.0117	0.9638	0.9631
PCA # 1 + BPNN	9	0.0601	10.5201	0.0596	28.0458	0.6910	0.7921
PCA # 2 + BPNN	13	**0.0405**	8.5201	0.0499	22.0321	0.8913	0.8908
SWR + BPNN	12	0.0454	**8.1921**	0.0488	**21.0112**	0.9492	0.9487
RF # 1 + BPNN	18	0.0416	8.1992	0.0470	21.0114	**0.9932**	**0.9926**
RF # 2 + BPNN	11	0.0419	8.1994	0.0475	21.0117	0.9930	0.9923

TABLE 7: Comparison of predictor variables selected by SWR and RF feature selection methods.

Group of variables	SWR	RF # 1	RF # 2
Geospatial	**Usage_Type** TOU	**Usage_Type** Season	**Usage_Type**
Geographical	Population_Ratio	Area Population_Ratio	Area
Climatic	Mean_MSL_Pressure **Mean_Maximum_Temperature** **Mean_Relative_Humidity** Total_Rainfall	Mean_Station_Pressure **Mean_Maximum_Temperature** **Mean_Relative_Humidity** Mean_Minimum_Temperature Mean_Drybulb _Temperature Mean_MSL_Pressure Total_Rainfall	Mean_Station_Pressure **Mean_Maximum_Temperature** **Mean_Relative_Humidity** Mean_Minimum_Temperature Mean_Drybulb_ temperature
Industrial	**Industrial_Labor** Industrial_Plant	**Industrial_Labor** Industrial_Plant	**Industrial_Labor**
Household	**Household** Liabilities	**Household** Expenditure Income Agriculturists Liabilities	**Household** Expenditure Income

Thailand. The derived hybrid machine learning model can assist in determining key variables to support or make decisions on the management of electricity consumption. Based on a successful demonstration with authentic data, the Provincial Electricity Authority can deploy this hybrid predictive model to achieve efficient energy demand planning and management. Future studies to improve the predictive accuracy may include machine learning techniques for feature selection or deep learning.

5. Conclusions

This study investigates the machine learning algorithms under different models to predict electricity consumption. The most effective prediction will likely come from using a more relevant dataset. The number of predictor variables is an important factor, directly related to the performance and accuracy of the predictive models. Some influential predictor variables must be selected to achieve a robust and accurate predictive model and reduce the computation time. The PCA, SWR, and RF algorithms are employed to select the predictor variables according to their importance. This study successfully selected 11 predictor variables out of the original 21 predictor variables by the RF. Both SWR and RF select the same set of 5 predictor variables, and the others are slightly different selections between models.

The BPNN with 21 predictor variables is conducted as a benchmark for performance comparison in this study. The

hybrid models are constructed by combining the PCA, SWR, and RF with the BPNN. The 10-fold cross-validation technique is also employed to ensure the unbiased, reliable, and accurate predictive model. This performance is compared in line with the RMSE, MAPE, NRMSE, SMAPE, R^2, and accuracy metrics. Comparison results confirm that the RF is superior to the PCA and SWR. Based on the experiments, the integrated RF with the BPNN algorithm could effectively improve the accuracy of the electricity consumption prediction.

Data Availability

The data used to support the findings of this study are available upon request from the corresponding author.

Conflicts of Interest

All authors state that there are no conflicts of interest.

Authors' Contributions

WK contributed to conceptualization, methodology, writing the original draft, visualization, validation, and investigation. YS contributed to conceptualization, supervision, writing the original draft, review and editing, validation, investigation, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

References

[1] R. Platon, V. R. Dehkordi, and J. Martel, "Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis," *Energy and Buildings*, vol. 92, pp. 10–18, 2015.

[2] K. Kavaklioglu, "Principal components based robust vector autoregression prediction of Turkey's electricity consumption," *Energy Systems*, vol. 10, no. 4, pp. 889–910, 2019.

[3] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," *Applied Energy*, vol. 140, pp. 385–394, 2015.

[4] M. Zekić-Sušac, A. Has, and M. Knežević, "Predicting energy cost of public buildings by artificial neural networks, CART, and random forest," *Neurocomputing*, vol. 439, pp. 223–233, 2021.

[5] F.-H. Chen and H. Howard, "An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree," *Soft Computing*, vol. 20, no. 5, pp. 1945–1960, 2016.

[6] A.-D. Pham, N.-T. Ngo, Q.-T. Nguyen, and N.-S. Truong, "Hybrid machine learning for predicting strength of sustainable concrete," *Soft Computing*, vol. 24, no. 19, pp. 14965–14980, 2020.

[7] T. R. V. Anandharajan, G. A. Hariharan, K. K. Vignajeth, R. Jijendiran, and Kushmita, "Weather monitoring using artificial intelligence," in *Proceedings of the 2nd International Conference on Computational Intelligence & Networks*, pp. 106–111, Bhubaneswar, India, January 2016.

[8] S. Walker, W. Khan, K. Katic, W. Maassen, and W. Zeiler, "Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings," *Energy and Buildings*, vol. 209, Article ID 109705, 2020.

[9] R. Pérez-Chacón, J. Luna-Romera, A. Troncoso, F. Martínez-Álvarez, and J. Riquelme, "Big data analytics for discovering electricity consumption patterns in smart cities," *Energies*, vol. 11, no. 3, p. 683, 2018.

[10] G. Shi, D. Liu, and Q. Wei, "Energy consumption prediction of office buildings based on echo state networks," *Neurocomputing*, vol. 216, pp. 478–488, 2016.

[11] S. R. Mohandes, X. Zhang, and A. Mahdiyar, "A comprehensive review on the application of artificial neural networks in building energy analysis," *Neurocomputing*, vol. 340, pp. 55–75, 2019.

[12] H. Lu, X. Ma, M. Ma, and S. Zhu, "Energy price prediction using data-driven models: a decade review," *Computer Science Review*, vol. 39, Article ID 100356, 2021.

[13] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902–924, 2017.

[14] A. A. Mamun, M. Sohel, N. Mohammad, M. S. Haque Sunny, D. R. Dipta, and E. Hossain, "A comprehensive review of the load forecasting techniques using single and hybrid predictive models," *IEEE Access*, vol. 8, pp. 134911–134939, 2020.

[15] M. Zekić-Sušac, R. Scitovski, and A. Has, "Cluster analysis and artificial neural networks in predicting energy efficiency of public buildings as a cost-saving approach," *Croatian Review of Economic, Business and Social Statistics*, vol. 4, no. 2, pp. 57–66, 2018.

[16] K. Muralitharan, R. Sakthivel, and R. Vishnuvarthan, "Neural network based optimization approach for energy demand prediction in smart grid," *Neurocomputing*, vol. 273, pp. 199–208, 2018.

[17] R. Pérez-Chacón, G. Asencio-Cortés, F. Martínez-Álvarez, and A. Troncoso, "Big data time series forecasting based on pattern sequence similarity and its application to the electricity demand," *Information Sciences*, vol. 540, pp. 160–174, 2020.

[18] M. Zekić-Sušac, S. Mitrović, and A. Has, "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities," *International Journal of Information Management*, vol. 58, Article ID 102074, 2021.

[19] N. Basurto, Á. Arroyo, R. Vega, H. Quintián, J. L. Calvo-Rolle, and Á. Herrero, "A hybrid intelligent system to forecast solar energy production," *Computers & Electrical Engineering*, vol. 78, pp. 373–387, 2019.

[20] C. El Mazgualdi, T. Masrour, I. El Hassani, and A. Khdoudi, "Machine learning for KPIs prediction: a case study of the overall equipment effectiveness within the automotive industry," *Soft Computing*, vol. 25, no. 4, pp. 2891–2909, 2021.

[21] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data - evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, pp. 63–71, 2019.

[22] A. F. S. Borges, F. J. B. Laurindo, M. M. Spínola, R. F. Gonçalves, and C. A. Mattos, "The strategic use of artificial intelligence in the digital era: systematic literature review and future research directions," *International Journal of Information Management*, vol. 57, Article ID 102225, 2021.

[23] Y. K. Dwivedi, L. Hughes, E. Ismagilova et al., "Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, Article ID 101994, 2021.

[24] I. C. Obagbuwa and A. P. Abidoye, "South Africa crime visualization, trends analysis, and prediction using machine learning linear regression technique," *Applied Computational Intelligence and Soft Computing*, vol. 2021, Article ID 5537902, 14 pages, 2021.

[25] H. Zhao, "Futures price prediction of agricultural products based on machine learning," *Neural Computing & Applications*, vol. 33, no. 3, pp. 837–850, 2021.

[26] W. Qiao, Z. Yang, Z. Kang, and Z. Pan, "Short-term natural gas consumption prediction based on Volterra adaptive filter and improved whale optimization algorithm," *Engineering Applications of Artificial Intelligence*, vol. 87, Article ID 103323, 2020.

[27] S. O. Atiku and I. C. Obagbuwa, "Machine learning classification techniques for detecting the impact of human resources outcomes on commercial banks performance," *Applied Computational Intelligence and Soft Computing*, vol. 2021, Article ID 7747907, 16 pages, 2021.

[28] S. H. Li, L. Zhu, Y. Wu, and X. Q. Lei, "A novel grey multivariate model for forecasting landslide displacement," *Engineering Applications of Artificial Intelligence*, vol. 103, Article ID 104297, 2021.

[29] F. Hu, X. Xu, J. Liang, C. Yang, M. Huang, and Q. Su, "Machine learning-based seawater concentration pathway prediction," *Computers & Electrical Engineering*, vol. 94, Article ID 107336, 2021.

[30] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: a new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[31] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 374, no. 2065, Article ID 20150202, 2016.

[32] J. Nobre and R. F. Neves, "Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets," *Expert Systems with Applications*, vol. 125, pp. 181–194, 2019.

[33] M. Roccetti, G. Delnevo, L. Casini, and S. Mirri, "An alternative approach to dimension reduction for pareto distributed data: a case study," *Journal of Big Data*, vol. 8, no. 1, p. 39, 2021.

[34] P. Zhang, "A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model," *Applied Soft Computing*, vol. 85, Article ID 105859, 2019.

[35] K. Aditya Shastry and H. A. Sanjay, "Hybrid prediction strategy to predict agricultural information," *Applied Soft Computing*, vol. 98, Article ID 106811, 2021.

[36] A. Radhika and M. S. Masood, "Effective dimensionality reduction by using soft computing method in data mining techniques," *Soft Computing*, vol. 25, no. 6, pp. 4643–4651, 2021.

[37] G. Wang, Y. Zhao, J. Zhang, and Y. Ning, "A novel end-to-end feature selection and diagnosis method for rotating machinery," *Sensors*, vol. 21, no. 6, p. 2056, 2021.

[38] G. Smith, "Step away from stepwise," *Journal of Big Data*, vol. 5, no. 1, p. 32, 2018.

[39] S.-F. Huang and C.-H. Cheng, "GMADM-based attributes selection method in developing prediction model," *Quality and Quantity*, vol. 47, no. 6, pp. 3335–3347, 2013.

[40] K. Gao, Y. Yang, T. Zhang, A. Li, and X. Qu, "Extrapolation-enhanced model for travel decision making: an ensemble machine learning approach considering behavioral theory," *Knowledge-Based Systems*, vol. 218, Article ID 106882, 2021.

[41] J. M. Cadenas, M. C. Garrido, R. Martínez, and P. P. Bonissone, "Extending information processing in a fuzzy random forest ensemble," *Soft Computing*, vol. 16, no. 5, pp. 845–861, 2012.

[42] U. Ahmed, H. Waqas, and M. T. Afzal, "Pre-production box-office success quotient forecasting," *Soft Computing*, vol. 24, no. 9, pp. 6635–6653, 2020.

[43] W. Chen, H. Xu, L. Jia, and Y. Gao, "Machine learning model for bitcoin exchange rate prediction using economic and technology determinants," *International Journal of Forecasting*, vol. 37, no. 1, pp. 28–43, 2021.

[44] J. Beninger, A. Hamilton-Wright, H. E. K. Walker, and L. M. Trick, "Machine learning techniques to identify mind-wandering and predict hazard response time in fully immersive driving simulation," *Soft Computing*, vol. 25, no. 2, pp. 1239–1247, 2021.

[45] M. Liu, S. Hu, Y. Ge, G. B. M. Heuvelink, Z. Ren, and X. Huang, "Using multiple linear regression and random forests to identify spatial poverty determinants in rural China," *Spatial Statistics*, vol. 42, Article ID 100461, 2021.

[46] C. Li, Y. Tao, W. Ao, S. Yang, and Y. Bai, "Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition," *Energy*, vol. 165, pp. 1220–1227, 2018.

[47] M. Arora, S. Dhawan, and K. Singh, "Improved performance of machine learning algorithms for prognosis of cervical cancer," *Advances in Computational Design*, vol. 6, pp. 191–205, 2021.

[48] D. A. Otchere, T. O. A. Ganat, R. Gholami, and M. Lawal, "A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction," *Journal of Natural Gas Science and Engineering*, vol. 91, Article ID 103962, 2021.

[49] H. J. Richardson, D. J. Hill, D. R. Denesiuk, and L. H. Fraser, "A comparison of geographic datasets and field measurements to model soil carbon using random forests and stepwise regressions (British Columbia, Canada)," *GIScience and Remote Sensing*, vol. 54, no. 4, pp. 573–591, 2017.

[50] S. E. Seker and I. Ocak, "Performance prediction of road-headers using ensemble machine learning techniques," *Neural Computing & Applications*, vol. 31, no. 4, pp. 1103–1116, 2019.

[51] S. Dami and M. Yahaghizadeh, "Predicting cardiovascular events with deep learning approach in the context of the internet of things," *Neural Computing & Applications*, vol. 33, no. 13, pp. 7979–7996, 2021.

[52] Q. Li, Z. Wu, R. Ling, L. Feng, and K. Liu, "Multi-reservoir echo state computing for solar irradiance prediction: a fast yet efficient deep learning approach," *Applied Soft Computing*, vol. 95, Article ID 106481, 2020.

[53] R. Janković, I. Mihajlović, N. Štrbac, and A. Amelio, "Machine learning models for ecological footprint prediction based on energy parameters," *Neural Computing & Applications*, vol. 33, no. 12, pp. 7073–7087, 2021.

[54] Á. Baran, S. Lerch, M. El Ayari, and S. Baran, "Machine learning for total cloud cover prediction," *Neural Computing & Applications*, vol. 33, no. 7, pp. 2605–2620, 2021.

[55] K. Thinakaran, R. Rajasekar, K. Santhi, and M. Nalini, "Predicting the 2-dimensional airfoil by using machine learning methods," *Advances in Computational Design*, vol. 5, pp. 291–304, 2020.

[56] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Massachusetts, MA, USA, 2016.

[57] S. Mu, Y. Wang, F. Wang, and L. Ogiela, "Transformative computing for products sales forecast based on SCIM," *Applied Soft Computing*, vol. 109, Article ID 107520, 2021.

[58] J. Manickavasagam, S. Visalakshmi, and N. Apergis, "A novel hybrid approach to forecast crude oil futures using intraday data," *Technological Forecasting and Social Change*, vol. 158, Article ID 120126, 2020.

[59] A. M. Ebid, L. I. Nwobia, K. C. Onyelowe, and F. I. Aneke, "Predicting nanobinder-improved unsaturated soil consistency limits using genetic programming and artificial neural networks," *Applied Computational Intelligence and Soft Computing*, vol. 2021, Article ID 5992628, 13 pages, 2021.

[60] Provincial Electricity Authority, "Load research of PEA," 2020, https://peaoc.pea.co.th/loadprofile/index.php.

[61] National Statistical Office, "Statistical data," 2020, https://web.nso.go.th.

[62] Thai Meteorological Department, "Thailand meteorological department API: tmdapi," 2020, https://data.tmd.go.th/api/index1.php.

[63] P. Yang, G. Tang, and A. Nehorai, "A game-theoretic approach for optimal time-of-use electricity pricing," *IEEE Transactions on Power Systems*, vol. 28, pp. 884–892, 2012.