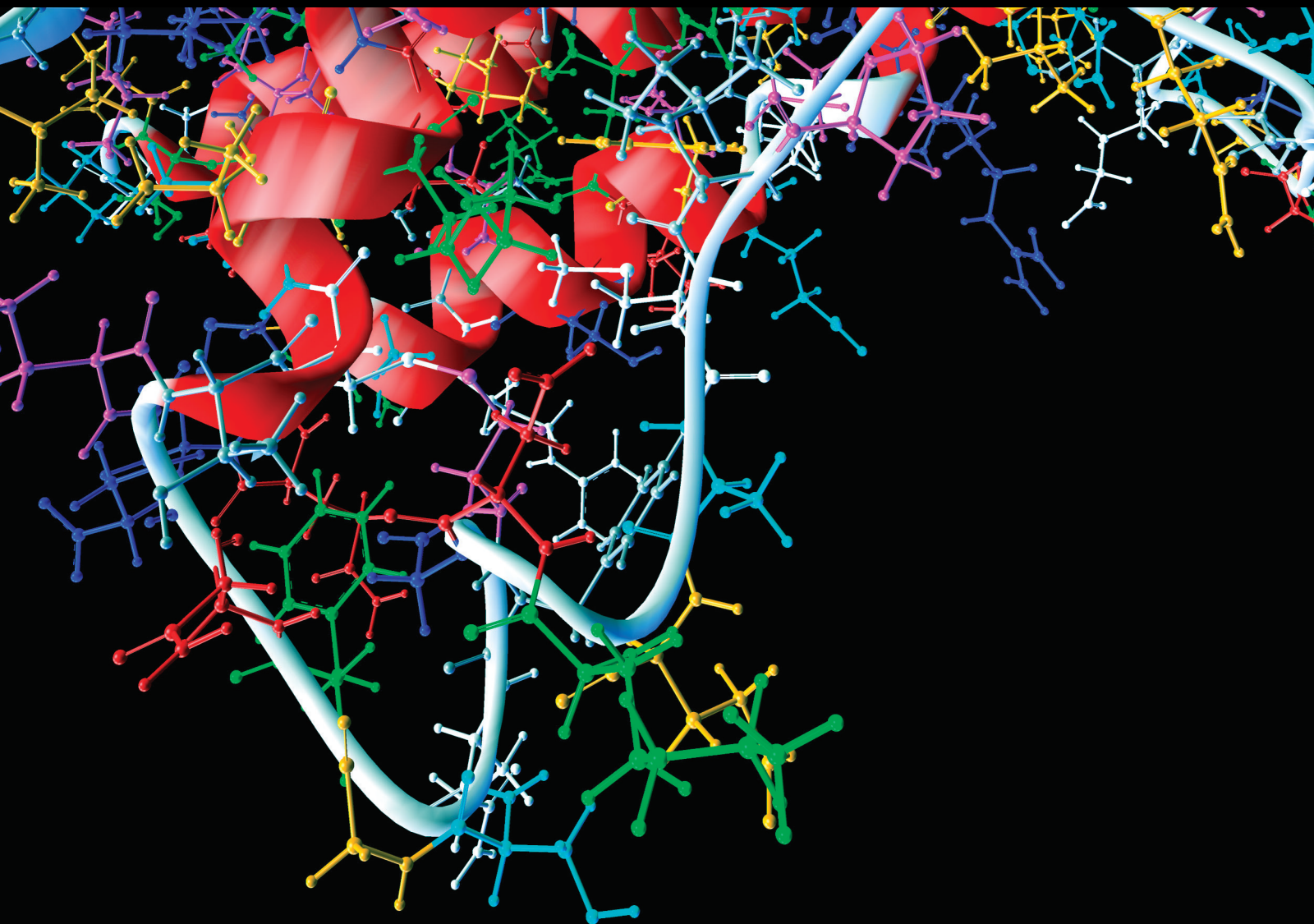


Data Analysis and Computational Methods in Public Health Surveillance Data

Lead Guest Editor: Miguel G. Torres

Guest Editors: David Bacerra-Alonso, José Luis Vázquez Noguera, and
Diego Pinto





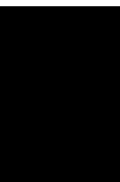
Data Analysis and Computational Methods in Public Health Surveillance Data

Computational and Mathematical Methods in Medicine

**Data Analysis and Computational
Methods in Public Health Surveillance
Data**

Lead Guest Editor: Miguel G. Torres




Guest Editors: David Bacerra-Alonso, José Luis
Vázquez Noguera, and Diego Pinto



Copyright © 2022 Hindawi Limited. All rights reserved.

This is a special issue published in “Computational and Mathematical Methods in Medicine.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editors

Ahmed Albahri, Iraq
Konstantin Blyuss , United Kingdom
Chuangyin Dang, Hong Kong
Farai Nyabadza , South Africa
Kathiravan Srinivasan , India

Academic Editors

Laith Abualigah , Jordan
Yaser Ahangari Nanehkaran , China
Mubashir Ahmad, Pakistan
Sultan Ahmad , Saudi Arabia
Akif Akgul , Turkey
Karthick Alagar, India
Shadab Alam, Saudi Arabia
Raul Alcaraz , Spain
Emil Alexov, USA
Enrique Baca-Garcia , Spain
Sweta Bhattacharya , India
Junguo Bian, USA
Elia Biganzoli , Italy
Antonio Boccaccio, Italy
Hans A. Braun , Germany
Zhicheng Cao, China
Guy Carrault, France
Sadaruddin Chachar , Pakistan
Prem Chapagain , USA
Huiling Chen , China
Mengxin Chen , China
Haruna Chiroma, Saudi Arabia
Watcharaporn Cholamjiak , Thailand
Maria N. D.S. Cordeiro , Portugal
Cristiana Corsi , Italy
Qi Dai , China
Nagarajan Deivanayagam Pillai, India
Didier Delignières , France
Thomas Desaive , Belgium
David Diller , USA
Qamar Din, Pakistan
Irina Doytchinova, Bulgaria
Sheng Du , China
D. Easwaramoorthy , India




Esmaeil Ebrahimie , Australia
Issam El Naqa , USA
Ilias Elmouki , Morocco
Angelo Facchiano , Italy
Luca Faes , Italy
Maria E. Fantacci , Italy
Giancarlo Ferrigno , Italy
Marc Thilo Figge , Germany
Giulia Fiscon , Italy
Bapan Ghosh , India
Igor I. Goryanin, Japan
Marko Gosak , Slovenia
Damien Hall, Australia
Abdulsattar Hamad, Iraq
Khalid Hattaf , Morocco
Tingjun Hou , China
Seiya Imoto , Japan
Martti Juhola , Finland
Rajesh Kaluri , India
Karthick Kanagarathinam, India
Rafik Karaman , Palestinian Authority
Chandan Karmakar , Australia
Kwang Gi Kim , Republic of Korea
Andrzej Kloczkowski, USA
Andrei Korobeinikov , China
Sakthidasan Sankaran Krishnan, India
Rajesh Kumar, India
Kuruva Lakshmana , India
Peng Li , USA
Chung-Min Liao , Taiwan
Pinyi Lu , USA
Reinoud Maex, United Kingdom
Valeri Makarov , Spain
Juan Pablo Martínez , Spain
Richard J. Maude, Thailand
Zahid Mehmood , Pakistan
John Mitchell , United Kingdom
Fazal Ijaz Muhammad , Republic of Korea
Vishal Nayak , USA
Tongguang Ni, China
Michele Nichelatti, Italy
Kazuhisa Nishizawa , Japan
Bing Niu , China

Hyuntae Park , Japan
Jovana Paunovic , Serbia
Manuel F. G. Penedo , Spain
Riccardo Pernice , Italy
Kemal Polat , Turkey
Alberto Policriti, Italy
Giuseppe Pontrelli , Italy
Jesús Poza , Spain
Maciej Przybyłek , Poland
Bhanwar Lal Puniya , USA
Mihai V. Putz , Romania
Suresh Rasappan, Oman
Jose Joaquin Rieta , Spain
Fathalla Rihan , United Arab Emirates
Sidheswar Routray, India
Sudipta Roy , India
Jan Rychtar , USA
Mario Sansone , Italy
Murat Sari , Turkey
Shahzad Sarwar, Saudi Arabia
Kamal Shah, Saudi Arabia
Bhisham Sharma , India
Simon A. Sherman, USA
Mingsong Shi, China
Mohammed Shuaib , Malaysia
Prabhishek Singh , India
Neelakandan Subramani, India
Junwei Sun, China
Yung-Shin Sun , Taiwan
Min Tang , China
Hongxun Tao, China
Alireza Tavakkoli , USA
João M. Tavares , Portugal
Jlenia Toppi , Italy
Anna Tsantili-Kakoulidou , Greece
Markos G. Tsipouras, North Macedonia
Po-Hsiang Tsui , Taiwan
Sathishkumar V E , Republic of Korea
Durai Raj Vincent P M , India
Gajendra Kumar Vishwakarma, India
Liangjiang Wang, USA
Ruisheng Wang , USA
Zhouchao Wei, China
Gabriel Wittum, Germany
Xiang Wu, China

KI Yanover , Israel
Xiaojun Yao , China
Kaan Yetilmezsoy, Turkey
Hiro Yoshida, USA
Yuhai Zhao , China



Contents

Diagnosing Breast Cancer Based on the Adaptive Neuro-Fuzzy Inference System

S. Chidambaram, S. Sankar Ganesh , Alagar Karthick, Prabhu Jayagopal , Bhuvaneswari Balachander, and S. Manoharan 



Research Article (11 pages), Article ID 9166873, Volume 2022 (2022)

Cross-Sectional Analysis of Impulse Indicator Saturation Method for Outlier Detection Estimated via Regularization Techniques with Application of COVID-19 Data

Sara Muhammadullah , Amena Urooj, Muhammad Hashim Mengal, Shahzad Ali Khan, and Fereshteh Khalaj 

Research Article (11 pages), Article ID 2588534, Volume 2022 (2022)

Complex Survival System Modeling for Risk Assessment of Infant Mortality Using a Parametric Approach

Hang Chen, Maryam Sadiq , and Zishen Song 

Research Article (8 pages), Article ID 7745628, Volume 2022 (2022)

The Partial Least Squares Spline Model for Public Health Surveillance Data

Maryam Sadiq , Dalia Kamal Fathi Alnagar , Alanazi Talal Abdulrahman, and Randa Alharbi

Research Article (7 pages), Article ID 8774742, Volume 2022 (2022)

Data Analysis and Computational Methods for Assessing Knowledge of Obesity Risk Factors among Saudi Citizens

Alanazi Talal Abdulrahman  and Dalia Kamal Alnagar 

Research Article (6 pages), Article ID 1371336, Volume 2021 (2021)

Research Article

Diagnosing Breast Cancer Based on the Adaptive Neuro-Fuzzy Inference System

S. Chidambaram,¹ S. Sankar Ganesh ,² Alagar Karthick,^{3,4} Prabhu Jayagopal ,⁵ Bhuvaneshwari Balachander,⁶ and S. Manoharan ⁷

¹Department of Information Technology, National Engineering College, Kovilpatti, 628503, Tamil Nadu, India

²Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, Arasur, Coimbatore, 641407, Tamil Nadu, India

³Renewable Energy Lab, Department of Electrical and Electronics Engineering, KPR Institute of Engineering and Technology, Arasur, Coimbatore, 641407, Tamil Nadu, India

⁴Departamento de Química Organica, Universidad de Cordoba, Edificio Marie Curie (C-3), Ctra Nnal IV-A, Km 396, E14014 Cordoba, Spain

⁵School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India

⁶Department of ECE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India

⁷Department of Computer Science, School of Informatics and Electrical Engineering, Institute of Technology, Ambo University, Ambo, Post Box No.: 19, Ethiopia

Correspondence should be addressed to S. Manoharan; manoharan.subramanian@ambou.edu.et

Received 22 December 2021; Revised 27 January 2022; Accepted 19 April 2022; Published 11 May 2022

Academic Editor: David Becerra-Alonso

Copyright © 2022 S. Chidambaram et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this work, a novel hybrid neuro-fuzzy classifier (HNFC) technique is proposed for producing more accuracy in input data classification. The inputs are fuzzified using a generalized membership function. The fuzzification matrix helps to create connectivity between input pattern and degree of membership to various classes in the dataset. According to that, the classification process is performed for the input data. This novel method is applied for ten number of benchmark datasets. During preprocessing, the missing data is replaced with the mean value. Then, the statistical correlation is applied for selecting the important features from the dataset. After applying a data transformation technique, the values normalized. Initially, fuzzy logic has been applied for the input dataset; then, the neural network is applied to measure the performance. The result of the proposed method is evaluated with supervised classification techniques such as radial basis function neural network (RBFNN) and adaptive neuro-fuzzy inference system (ANFIS). Classifier performance is evaluated by measures like accuracy and error rate. From the investigation, the proposed approach provided 86.2% of classification accuracy for the breast cancer dataset compared to other two approaches.

1. Introduction

Recently, data mining plays a major role in both industry and research organizations due to the accessibility of the huge volume of data and transforms these data into significant information and knowledge. Mainly classification [1] is the approach determining a classifier that compares and predict a target class with an unidentified class label. During

the training phase, it follows two phases; a classifier is developed, as well as its relevant class variables. During the test phase, a set of features are applied to approximate the level of the classifier.

Before the data classification process, many preprocessing procedures have been applied. The artificial neural network (ANN) can do intellectual responsibilities like the human brain. A popular trustworthy classification method

from the NN is the multilayer backpropagation network [2]. And a radial basis function [3] is a dominant neural approach that uses radial basis procedures. In that, neuron parameters are considered for producing better performance.

The artificial neural network (ANN) is a trendy data modeling approach which can carry out intelligent tasks in the same way as the human brain. ANN is well suitable for high-precision and high-learning ability purpose. One of the reliable approaches of data classification from the neural network area is the multilayer perceptron backpropagation network (MLPBPN) approach [4]. The output of this neural network technique is the linear combination of radial basis functions of inputs and neuron factors. RBFNNs helps for classification, function approximation, and prediction of time series applications.

The discrete-time linear dynamical systems [5] are used to make a spirit for the approximation. It includes time-varying systems by recurrent neural networks (RNNs). For the subclass of linear time-invariant (LTI) systems, learning a differential equation is the easiest feasible mathematical incarnation. In the experimental results, the dynamics of physical, biological, mechanical, or chemical procedures are recognized from practical input-output traces. An adaptive proportional-integral controller is used along with the proper gain variation according to the adaptive neuro-fuzzy inference system (ANFIS) to promise high performances of electric drive models with respect to the parametric differences.

In a fuzzy approach, the selected attributes are linked with a degree of membership to various groups. Both NN and fuzzy approaches are flexible to measure *I/O* correlations. Fuzzy systems consider figurative as well as quality-based data. The condition-oriented neuro-fuzzy approach is categorized as the linguistic fuzzy modeling that deals with the inference and fuzzy modeling technique like the Sugeno model which considers accuracy [6].

The modular neural network is an incorporation of smaller subcomplete neural network models [7]. Each model functions separately on a subportion of larger size pattern vectors. There are two ways of modularizing the neural network, i.e., modularizing learning and modularizing structure. The modular learning for pattern classification of hand-written Hindi alphabets is considered. Here, twenty-four individual subneural networks have been considered for first phase computing. Then, the collective outputs of the first phase are applied as input to the global neural network. Thus, the output of the second phase presents the desired classification of the given large training set. Neural networks of the first phase are trained locally for decomposed input patterns with gradient descent learning. Updated weights of the first phase are mapped to the global neural network. The global neural network is further trained for the collective output patterns of the first phase computing. Here, decomposition and replication concepts have been applied to perform the classification task [8].

A forecast time series model [9] is proposed which uses generalized regression neural networks. The objective is to take advantage of their inherent properties to produce fast

and accurate forecasts. The key modeling decisions are involved in forecasting with generalized regression neural networks. For every modeling decision, several strategies are proposed. Each strategy is analyzed in terms of forecast accuracy and computational time. Apart from the modeling decisions, any successful time series forecasting methodology has to be able to capture the seasonal and trend patterns found in a time series. There are three different forecasting models proposed such as the sigmoid function regression model [9], the feedforward neural network, and the recurrent neural network model. The models were trained, compared, and validated using gas consumption data.

A novel adaptive backstepping approach is used to manage the induction motor (IM) rotor resistance tracking issue. The robustness of the device can be forecast with the experimental results. The various parameters are determined such as rotor resistance, sensitivity and torque [10].

The genetic optimization algorithm [11] was applied to train the neural networks, and the Levenberg-Marquardt algorithm was applied to attain the parameters of the sigmoid model. From the results, it shows that both neural network models perform similarly and are superior to the sigmoid model. The models were prepared for use in conjunction with a weather forecasting service to generate day-ahead or within-day forecasts and are relevant to any geographical area.

The risk management framework is used to represent digitally the product of probability and consequence. In the conventional approach, it has been increasingly discussed to include strength of evidence combined with the traditional consequence and probability. It also focuses on addressing these challenges and makes the risk expression fully digital analysis and visualization. In the proposed approach to address the challenges by forming a fuzzy logic index based on fuzzy logic theory, this enables a transfer from a linguistic variable to a digital one. Then, it can be applied into a node size index to express its practical application. It enables an improved risk visualization, risk management, and risk communication for system analysis, towards risk digitalization.

The rule-based neuro-fuzzy approach [12] is split into two categories: the linguistic fuzzy modeling which can focus on interpretability, primarily the Mamdani model, and the fuzzy modeling that concentrated on accuracy, primarily the Sugeno model or Takagi-Sugeno-Kang (TSK) model. This rule-based approach normally applies the concept of the adaptive neural network. An adaptive network [13] is a network of nodes and directed links that is functionally equivalent to a fuzzy inference system.

In that, IF-THEN conditions are generated [14]. Individual nodes are attached with some significant parameters. The Sugeno model fuzzy rule is represented as

$$\text{IF } a \text{ is } M \text{ and } b \text{ is } N, \text{ THEN } Z = f(a, b), \quad (1)$$

where M and N are the fuzzy sets in the rule and “ Z ” is an output function.

This research work is arranged in this paper as follows: Section 2 describes the interrelated concepts carried out in

this research domain. Section 3 explains artificial neural network classification approach functionalities Section 4 explains the architecture and learning method of the RBFNN classifier. Section 5 explains the step by step procedure for the proposed neuro-fuzzy approach. Section 6 discusses the performance analysis and results, and finally, Section 7 concludes the paper.

2. Related Work

The fuzzy neural networks (FNN) are proposed, which have the main objective of practicing numerical relationships and practicing numerical and perception oriented information. It can reduce error rate and find the connection weights as well as bias values. A particle swarm optimization [2] is matched with the backpropagation approach for training the dataset. It produces maximum accuracy in prediction.

A novel hybrid forecasting approach [15] is based on the firefly algorithm. In that, an algorithm optimizer is combined with the adaptive neuro-fuzzy inference system for assessing the fragmentation. The proposed hybrid models were evaluated based on the statistical criteria such as coefficient of calculation and Nash and Sutcliffe. The adaptive neuro-fuzzy inference system (ANFIS) [16] is proposed to determine axial velocity and flow depth in a 90° sharp bend. The velocity and flow depth data for five discharge rates are applied for training and testing the models. In the ANFIS training phase, the two algorithms are backpropagation and a hybrid of backpropagation and least squares. In the proposed model design, the grid partitioning and subclustering methods are applied for generating the fuzzy inference system.

The fuzzy set theory [17] is used to describe an essential involvement to fuzzy concepts in data mining techniques. It manages interpretable and subjective information. A sliding window approach is used to produce time series subsequences and then analyze the fuzzy item sets. It handles temporal data to determine association rules.

The Adaptive Genetic Fuzzy System (AGFS) [18] is used for optimizing rules in the healthcare data classification. The main objective is to produce optimized rules from data. Fuzzy set theory [19] in machine learning deals with techniques for applying automated induction approaches and pattern extraction from experiential data.

A novel fuzzy partition learning approach [20] is used for applying artificial immune system methods for improving classification accuracy. An efficient CRM-data mining framework [21] is used to establish tight customer relationships and deal with the association between organizations and customers in order to take a decision. With the development of the database, the volume of data in the database increases quickly and sensitive data is protected by applying some security mechanism.

A genetic algorithm (GA) [22] is engaged to determine the optimal selection of adaptive neuro-fuzzy inference system (ANFIS) membership functions and the evolutionary design of a generalized group method of data handling (GMDH) structure for prediction of the side weir discharge coefficient. The Singular Value Decomposition (SVD) method is applied to measure the linear parameters of the

ANFIS classifier and linear coefficient vectors in GMDH. The uncertainty investigation is also performed to measure the quantitative performance of all types of models.

The multilayer perceptron network [23] is applied with three types of training algorithms which include variable learning rate (MLP-GDX), resilient backpropagation (MLP-RP), and Levenberg-Marquardt (MLPLM) [23, 24]. These approaches were studied based on the ability to approximate the sediment transport in a clean pipe. Model ANN that employs volumetric sediment concentration (CV), median relative size of particles, ratio of median diameter particle size to hydraulic radius, and overall sediment friction factor as input parameters is more accurate than the other existing models.

The subfeature selection of the attributes [24] uses fuzzy methodologies to preserve privacy of the users in the distributed environment. An effective knowledge extraction approach is proposed which can get knowledge in terms of rules. At first, train the model and prune the decision tree to take out optimized rules. A correlation-oriented feature selection is introduced with a linear search approach for cardiac arrhythmia disease classification.

An adaptive neuro-fuzzy-embedded subtractive clustering (ANFIS-SC) [25] approach is applied for evaluating the abutment scour hole depth under clear water condition with uniform bed sediments. The accuracy of the ANFIS-SC approach is compared with that of two other ANFIS approaches embedded with fuzzy C-mean clustering [26] and grid partitioning. The decisive factors on the abutment scour hole depth include the ratio of the average diameter of particle size to abutment transverse length, excess Froude number of the abutment, shape factor, and the ratio of approach stream depth to abutment transverse length.

A genetic algorithm [27] is used for training neural networks, and analysis is made to compute the convergence error rate in a neural network. A hybrid fuzzy min-max neural network [28] is proposed, which is suitable for outlier detection. A hybrid algorithm with respect to a genetic algorithm and particle swarm optimization technique can also be applied to model a fuzzy neural network. A fuzzy wavelet neural network (FWNN) technique is another approach for obtaining better accuracy in classification.

The multilayer perceptron neural network [29] applies the artificial neural network to pick up the essential characteristics of the input layer of the network. A fuzzy radial basis polynomial network design approach [30, 31] is suitable for granular information classification. An automated healthcare classification technique [32] is introduced for wavelet transformation (WT). It is helpful in the decision support system for medical practitioners.

Neural network-based sentiment classification approaches [33, 34] such as BPNN and probabilistic NN approaches using different stages of word granularity are compared as attributes.

3. ANFIS Architecture

Consider the fuzzy inference method has two input values such as “ x ” and “ y ,” and “ s ” is the output. The Sugeno fuzzy method has two if-then rule constraints shown in Figure 1:

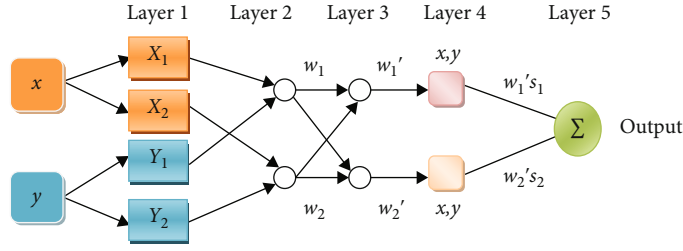


FIGURE 1: ANFIS architecture layer-wise representation.

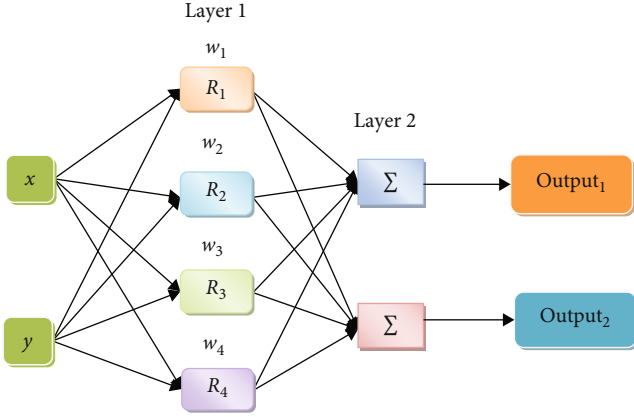


FIGURE 2: Input, hidden, and output layer of the RBFNN model.

- (i) If x has value X_1 and y has value Y_1 , then $s_1 = p_1x + q_1y + r_1$
- (ii) If x has value X_2 and y has value Y_2 , then $s_2 = p_2x + q_2y + r_2$

Layer 1: all the nodes are represented in

$$\begin{aligned} O_{1,i} &= \mu X_i(x), \quad \text{for } i = 1, 2 \dots, \\ O_{1,i} &= \mu Y_{i-2}(x), \quad \text{for } i = 3, 4 \dots, \end{aligned} \quad (2)$$

where “ x ” or “ y ” is the input to the node “ i ” and X_i (or Y_{i-2}) is an associated node.

The generalized bell function is represented by

$$\mu X_i(x) = \frac{1}{1 + |(x - n_i)/l_i|^{2m}}, \quad (3)$$

where $\{l_i, m_i, n_i\}$ is the argument set. When parameters are modified, the bell-shaped function varies consequently. These parameters are called as premise parameters.

Layer 2: in that, all the nodes are fixed. Its output is determined by finding a product of inputs. It is represented by

$$O_{2,i} = w_i = \mu X_i(x) \mu Y_i(x), \quad \text{where } i = 1, 2. \quad (4)$$

Layer 3: the node evaluates the proportion of the i^{th} rule’s weighted value to the summation of the value of all

weighted rules. It is denoted by

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2. \quad (5)$$

Layer 4: in that, all nodes are adaptive in nature. It is denoted by

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i (l_i x + m_i y + n_i), \quad (6)$$

where \bar{w}_i represents a normalized weighted value of the output layer and $\{l_i, m_i, n_i\}$ is the consequential attribute set.

Layer 5: one node can measure the resultant value by the sum of all inputs. It is denoted by

$$\text{Output} = O_{5,1} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}. \quad (7)$$

This network is the same as the Sugeno fuzzy model with respect to functionality. But structure-wise, it is different.

4. Radial Basis Function Networks

4.1. *Architecture and Learning Methods.* The activation stage in the hidden layer is denoted by

$$w_i = R_i(v) = R_i\left(\frac{\|x - u_i\|}{\sigma_i}\right), \quad i = 1, 2, \dots, M, \quad (8)$$

where “ v ” represents the input vector, u_i denotes the vector with the similar measurement like v , M denotes the count, and $R_i(\cdot)$ is the i^{th} radial basis function. Weighted value have not been assigned among the input and the hidden layer shown in Figure 2.

Normally, $R_i(\cdot)$ represents the Gaussian function in

$$R_i(v) = \exp\left(-\frac{\|v - u_i\|^2}{2\sigma_i^2}\right). \quad (9)$$

The activation stage w_i measured by the i^{th} hidden layer is greatest. In RBFN, the overall output is calculated as the weighted sum of the outputs related to the attributes. It is represented by

$$d(v) = \sum_{i=1}^H c_i w_i = H \sum_{i=1}^H c_i R_i(v), \quad (10)$$

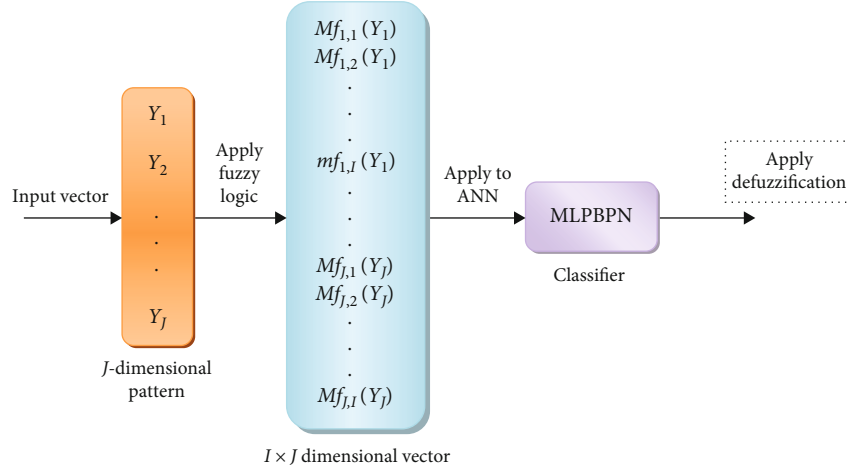


FIGURE 3: Proposed neuro-fuzzy classification approach.

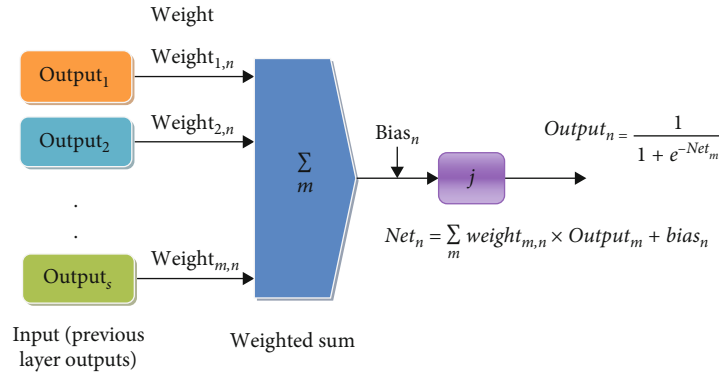


FIGURE 4: MLPBPN architecture layer-wise procedure.

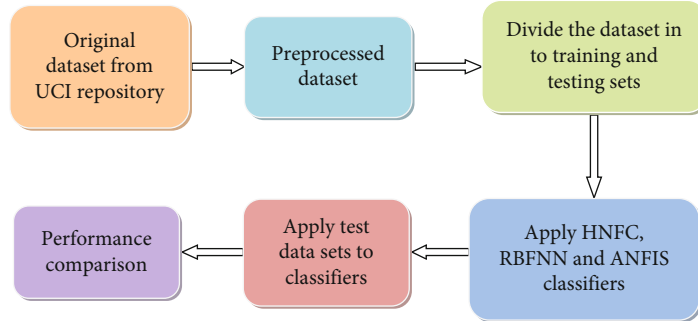


FIGURE 5: Detailed steps for the development of the proposed system.

where c_i is represented as the connection weight between the field and the output. It is denoted in

$$d(v) = \frac{\sum_{i=1}^H c_i w_i}{\sum_{i=1}^H w_i} = \frac{\sum_{i=1}^H c_i R_i(v)}{\sum_{i=1}^H R_i(v)}. \quad (11)$$

Both FIS and RBFN have a procedure whereby it can generate a center-weighted radical-shaped function. In the

following constraints, an RBFN and a FIS have equal functionality:

- (i) Both RBFN and FIS utilize the identical aggregation approach such as weighted sum and weighted average
- (ii) The receptive field unit's count in the RBFN is equivalent to the if-then rule condition in the fuzzy approach

TABLE 1: List of features of the breast cancer dataset.

S. no	List of attributes	Type of data
1	Age	Numeric
2	Mefalsepause	Numeric
3	Tumor size	Numeric
4	Inv-falsedes	Numeric
5	Falsede-caps	Numeric
6	Deg-malig	Numeric
7	Breast quad	Numeric
8	Irradiat	Numeric
9	Class	Categorical

(iii) Subsequent radial basis function and fuzzy rule representation have similar response to the input

5. Proposed Method

The proposed approach can perform the selected features from a set of input prototypes, fuzzifies the equivalent prototype measures, and applies a membership function of each prototype in classes. Consider the input patterns (N), set of classes (M), and attributes (k). The proposed classification approach is shown in Figure 3.

The proposed technique contains three steps:

Step 1. In this fuzzification stage, a matrix order $J \times I$ is produced that contains the membership degree of J patterns. Every data element in this matrix is denoted as $mf_{m,n}(y_m)$, where y_m represents the m^{th} input pattern vector value, where $m = 1, 2, \dots, J$ and $n = 1, 2, \dots, I$. The membership function is represented as

$Mf_{m,n}(y_m)$ = membership pattern from class m to n ,

where the m^{th} pattern $y_m = x_{m1}, x_{m2}, \dots, x_{mk}$.

The input pattern vector “ y ” is represented by

$$y = [y_1, y_2, \dots, y_J]^T. \quad (12)$$

A generalized bell-shaped membership function is used which is based on three parameters such as p , q , and r as given by

$$mf(y : p, q, r) = \frac{1}{1 + |(y - r)/p|^{2q}}. \quad (13)$$

The resultant membership of the pattern vector matrix y is denoted by

$$MF(y) = \begin{bmatrix} mf_{1,1}(y_1) & mf_{1,2}(y_1) & mf_{1,3}(y_1) & \dots & mf_{1,I}(y_1) \\ mf_{2,1}(y_2) & mf_{2,2}(y_2) & mf_{2,3}(y_2) & \dots & mf_{2,I}(y_2) \\ mf_{3,1}(y_3) & mf_{3,2}(y_3) & mf_{3,3}(y_3) & \dots & mf_{3,I}(y_3) \\ \dots & \dots & \dots & \dots & \dots \\ mf_{J,1}(y_J) & mf_{J,2}(y_J) & mf_{J,3}(y_J) & \dots & mf_{J,I}(y_J) \end{bmatrix}, \quad (14)$$

where $mf_{m,n}(y_m)$ is the member of m^{th} pattern of input values “ y ” where $m = 1, 2, \dots, J$.

Step 2. In this step, MLPBPN is constructed. It converts the matrix values into an $M \times N$ vector by transposing it. This converted vector value is applied as input to the classifier.

A distinctive MLPBPN approach has a one input and output layer and a minimum one hidden layer. It demonstrates two types of procedures: feedforward and backpropagation. The nodes are associated in a feedforward approach. The input nodes are linked to the hidden nodes, and the hidden elements are entirely related to the output layer elements. The input and hidden nodes are linked with the weighted value. All weighted values of nodes are preferred arbitrarily shown in Figure 4.

In the backpropagation method, the happening of errors and the learning process such as revising the weighted value and biases are transmitted in the reverse route beginning from the output level to the internal values. This procedure is replicated many times. The main objective is to reduce the root-mean-square error among the forecast and actual values up to completion of the preparation process or the final condition attained [35–39].

The predicted output of element “ n ” is represented by

$$\text{Output}_n = \frac{1}{1 + e^{-\text{Net}_n}}, \quad (15)$$

where Net_n is the total input of element “ n ” in this model. The total input value is represented as a sum of the connection strengths and the result from the previous stage. It is represented in

$$\text{Net}_n = \sum_m \text{weight}_{m,n} \times \text{Output}_m + \text{bias}_n, \quad (16)$$

where $\text{weight}_{m,n}$ is the connection strength of the connection from element “ m ” in the preceding stage to unit “ n .” Output_m is the output of element “ m ” from the previous stage, and bias_n is the bias of the element.

The total of squared error values from the predictable result is measured by

$$\text{Error} = \frac{1}{2} \sum_n (\text{Target}_n - \text{Output}_n)^2. \quad (17)$$

The weighted value of the backpropagation network model is changed to decrease this error. It is denoted in

$$\Delta \text{Weight} \propto - \frac{\partial \text{Error}}{\partial \text{Weight}}. \quad (18)$$

The final output stage “ n ” with a weight value, $\text{weight}_{m,n}$,

TABLE 2: Detailed performance comparison for three classifiers.

Datasets	Classifiers	Acc (%)	TP rate/recall (%)	FP rate (%)	Precision (%)	F-measure (%)	TT (sec)
Breast cancer	HNFC	86.2	85.2	14.8	85.5	85.2	1521.2
	RBFNN	83.3	82.3	17.7	82.8	82.3	1821.5
	ANFIS	82.2	80.5	19.5	82.1	80.5	1712.2
Diabetes	HNFC	85.4	83.5	16.5	82.6	83.5	1514.6
	RBFNN	80.4	78.6	21.4	79.6	78.6	1815.2
	ANFIS	79.2	77.5	22.5	78.5	77.5	1945.2
E. coli	HNFC	75.9	74.2	25.3	76.5	74.2	1612.2
	RBFNN	73.6	72.8	27.2	71.6	72.8	1812.2
	ANFIS	72.8	71.5	28.5	72.1	71.5	1921.2
Liver disorder	HNFC	78.8	77.8	22.2	77.8	77.8	1621.2
	RBFNN	76.8	74.5	25.5	74.5	74.5	1752.6
	ANFIS	70.2	71.5	28.5	71.5	71.5	1721.6
Primary tumor	HNFC	80.4	80.2	19.8	78.5	80.2	1825.5
	RBFNN	78.6	77.3	22.7	75.6	77.3	2112.5
	ANFIS	77.6	75.8	24.2	72.8	75.8	2512.6
Mushroom	HNFC	92.5	91.1	8.9	90.5	91.1	1321.2
	RBFNN	90.6	88.5	11.5	89.5	88.5	1521.9
	ANFIS	88.9	87.8	12.2	87.2	87.8	1569.3
Ionosphere	HNFC	95.5	94.1	5.9	93.5	94.1	1125.6
	RBFNN	93.6	92.5	7.5	91.3	92.5	1253.2
	ANFIS	92.2	90.1	9.9	89.6	90.1	1245.6
Credit-g	HNFC	96.8	94.6	5.4	89.5	94.6	1325.2
	RBFNN	93.8	92.2	7.8	90.6	92.2	1452.2
	ANFIS	91.8	90.8	9.2	89.9	90.8	1441.3
Anneal-org	HNFC	95.9	94.8	5.2	93.5	94.8	1221.2
	RBFNN	93.8	92.5	7.5	91.8	92.5	1362.3
	ANFIS	91.5	90.6	9.4	90.6	90.6	1401.2
Iris	HNFC	96.8	95.6	4.4	95.2	95.6	1323.1
	RBFNN	94.2	94.1	5.9	93.6	94.1	1391.2
	ANFIS	93.5	92.5	7.5	90.8	92.5	1423.2

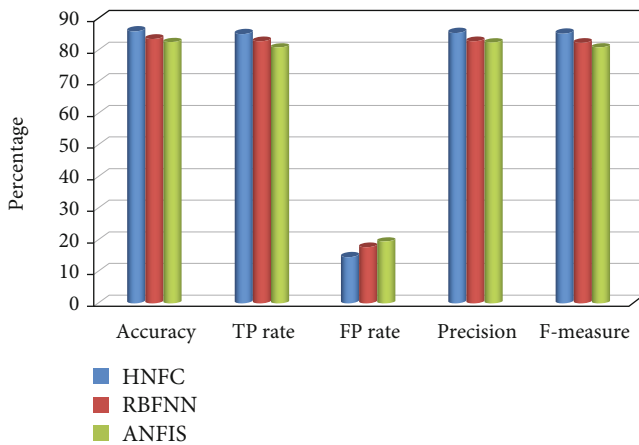


FIGURE 6: Performance measures for the breast cancer dataset.

is determined by

$$\Delta \text{Weight}_{m,n} \propto -\frac{\partial \text{Error}}{\partial \text{Weight}_{m,n}},$$

$$\Delta \text{Weight}_{m,n} = -\eta \frac{\partial \text{Error}}{\partial \text{Output}_n} \times \frac{\partial \text{Output}_n}{\partial \text{Net}_n} \times \frac{\partial \text{Net}_n}{\partial \text{Weight}_{m,n}}, \quad (19)$$

where η denotes the learning rate. Here, the weight updating formula is represented as

$$\text{Weight}_{m,n} = \text{Weight}_{m,n} + \Delta \text{Weight}_{m,n}. \quad (20)$$

Similarly, bias updation is performed by

$$\text{bias}_n = \text{bias}_n + \Delta \text{bias}_n. \quad (21)$$

In the MLPBPN approach, only one hidden layer is used. The neural network approach uses gradient descent with impetus as supervised conditions. Both hidden and last layers follow the tan sigmoidal transfer function.

The input layer nodes are equivalent to the amount of input features in the datasets. In the same way, the count of resultant nodes is equal to the quantity of class labels. The elements in the hidden layer are denoted as L in

$$L = (\text{Input featurecount} + \text{Total classcount}) * \frac{2}{3}. \quad (22)$$

Step 3. In this defuzzification stage, the proposed classifier classifies and defuzzifies the activation result. The input prototype is selected to the class “ n ” with the highest membership label.

5.1. Detailed Procedure

- (Step 1) Apply data cleaning in which preprocessing of data is performed by eliminating or decreasing noise. The attribute missing values are replaced by its mean value.
- (Step 2) Apply data selection in which statistical correlation analysis is applied to remove duplicate features, and then, only the relevant features can be collected.
- (Step 3) Apply transformation of data in which normalization is applied to the dataset. The neural network-based technique involves transformation of values ranging from -1.0 to $+1.0$.
- (Step 4) The data is separated into two subsets, training and test datasets, after preprocessing.
- (Step 5) In the training stage, the data is applied to the proposed system for creating a prototype. It also implemented for both RBFNN and ANFIS approaches for developing other classifiers.
- (Step 6) In the testing stage, three classifiers such as NFS, RBFNN, and ANFIS are applied for calculating its performance.
- (Step 7) The performance measures of these models are compared.

The detailed procedure is shown in Figure 5.

6. Results and Analysis

In our experiment, three classification approaches such as HNFC, RBFNN, and ANFIS are applied on benchmark datasets, namely, primary tumor, breast cancer, *E. coli*, mushroom, diabetes, ionosphere, liver disorder, Credit-g, Anneal-org, and iris. From the machine learning repository

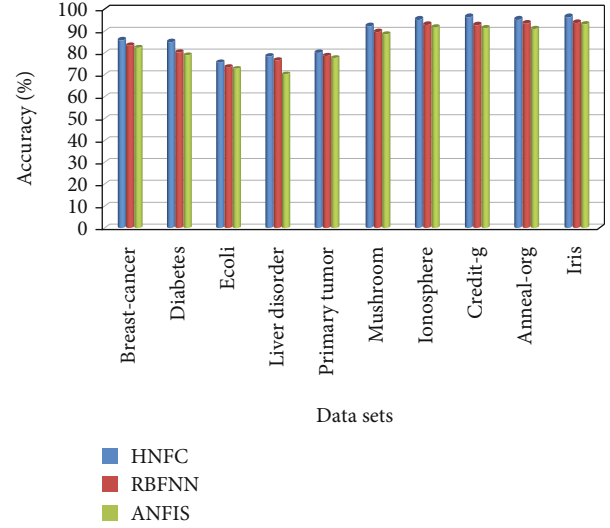


FIGURE 7: Comparison of classification accuracy for various datasets.

using the R tool, there are 50,000 records that are created for each dataset and compare its performance. In the breast cancer dataset, it has ten numbers of features such as age, mefalsepause, tumor size, inv-falsedes, falsede-caps, deg-malig, breast, breast-quad, and irradiat. All the features are multivariate categorical type of attributes.

A performance comparison has been done by considering various metrics such as accuracy, TP-rate, FP-rate, precision, F -measure, and root mean square error (RMSE). From the experimental outcomes given in Table 1, for the above specified datasets, the proposed HNFC method has produced better classification accuracy compared to other two approaches such as RBFNN and ANFIS.

6.1. Performance Measures. The performances of the classifiers are evaluated as per the following metrics:

6.1.1. Confusion Matrix. The confusion matrix is an illustration which gives the detailed visualization of the classification performance. Each column represents the records in a predicted variable. The row denotes the records in an actual variable.

- (i) True positive is a count of correct and positively classified objects
- (ii) False positive is a count of incorrectly classified instances which are positive
- (iii) False negative is a count of incorrectly classified instances which are negative
- (iv) True negative is a count of correctly classified objects that are negative

Accuracy of the correctly classified instance is determined by

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}. \quad (23)$$

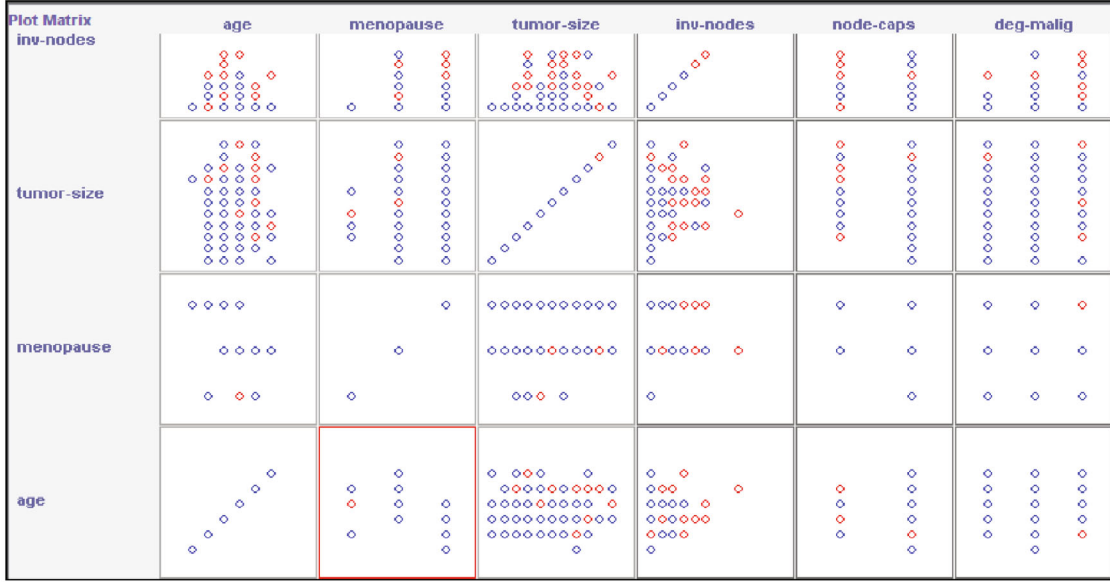


FIGURE 8: Distribution and recurrent relationship among features in the dataset.

The relation of the forecast positive objects which are accurate is determined by

$$\text{Precision} = \frac{tp}{tp + fp}. \quad (24)$$

The relation of negative objects which are incorrectly classified as positive is represented by

$$\text{FP-rate} = \frac{fp}{fp + tn}. \quad (25)$$

The relation of positive objects which are suitably classified is calculated by

$$\text{recall} = \text{tp-rate} = \frac{tp}{tp + tn}. \quad (26)$$

In some situations, maximum recall value may be important. To increase the performance measures, both precision recall values are represented by

$$F\text{-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (27)$$

From Table 2, while the breast cancer dataset is taken as input and applied to existing classifiers such as RBFNN and ANFIS, it produces the accuracy of 83.3% and 82.2%, respectively. But at the same time, for the proposed hybrid neuro-fuzzy classification, it gives 86.2%. It is comparatively higher than that of the other two approaches. Root mean square error is also low (0.323) for the proposed system. Time complexity is high compared to that of other decision tree classification approaches. But this drawback can be overcome by means of producing maximum accuracy in classifications. The existing approaches with the proposed algorithms have

been applied for other datasets such as diabetes, liver disorder, E. coli, primary tumor, mushroom, ionosphere, Credit-g, Anneal-org, and iris. The performance comparison has been shown in Figures 6 and 7.

Figure 8 denotes the various features of the breast cancer dataset and relationship among various data items in the dataset. Red color circle denotes the recurrence events, and blue color denotes no recurrent events in the dataset.

From the above experimental graphs, for the breast cancer dataset, a number of instances are distributed and the class labels are indicated in red and blue color circle format. With respect to various attribute values, the distribution ranges and values will be varied. The chart representation indicates the accuracy for the various existing algorithms such as the radial basis function neural network and adaptive network-based fuzzy inference system; the proposed hybrid neuro-fuzzy classifier provides better classification accuracy for the various input data such as breast cancer, diabetes, E. coli, liver disorder, primary tumor, mushroom, ionosphere, Credit-g, Anneal-org, and iris.

7. Conclusion

In this paper, we compared the proposed HNFC approach with RBFNN and ANFIS classification. The classifiers were experimented with ten UCI repository datasets. From the experimental outcome, the proposed classifier produces 86.2% better performance in classification of datasets compared to the existing algorithms. Similarly, this classifier provides better performance in classifying the input data. And it also provides a valuable contribution to the performance improvement of conventional classification approaches in the data mining research field. Still there is a research opening to apply other classifiers to predict the disease based the medical records.

Data Availability

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

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

Alagar Karthick gratefully acknowledges the group FQM-383 from Universidad de Cordoba, Spain, for the provision of an honorary visiting research position in the group.

References

- [1] R. A. Alieva, B. G. Guirimova, B. Fazlollahib, and R. R. Aliev, "Evolutionary algorithm-based learning of fuzzy neural networks. Part 2: recurrent fuzzy neural networks," *Fuzzy Sets and Systems*, vol. 160, no. 17, pp. 2553–2566, 2009.
- [2] M. Subramanian, M. S. Kumar, V. E. Sathishkumar et al., "Diagnosis of retinal diseases based on Bayesian optimization deep learning network using optical coherence tomography images," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 8014979, 15 pages, 2022.
- [3] S. Saroja, R. Madavan, S. Haseena et al., "Human centered decision-making for COVID-19 testing center location selection: Tamil Nadu—a case study," *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 2048294, 13 pages, 2022.
- [4] A. Gholami, H. Bonakdari, I. Ebtehaj, and A. A. Akhtari, "Design of an adaptive neuro-fuzzy computing technique for predicting flow variables in a 90° sharp bend," *Journal of Hydroinformatics*, vol. 19, no. 4, pp. 572–585, 2017.
- [5] R. Nanmaran, S. Srimathi, G. Yamuna et al., "Investigating the role of image fusion in brain tumor classification models based on machine learning algorithm for personalized medicine," *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 7137524, 13 pages, 2022.
- [6] S. V. Kogilavani, J. Prabhu, R. Sandhiya et al., "COVID-19 detection based on lung CT scan using deep learning techniques," *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 7672196, 13 pages, 2022.
- [7] D. S. Broomhead and L. David, "Radial basis functions, multi-variable functional interpolation and adaptive networks," Tech. Rep. 4148, Royal Signals and Radar Establishment, Technical report, 2010.
- [8] S. Kaliappan, R. Saravanakumar, A. Karthick et al., "Hourly and day ahead power prediction of building integrated semi-transparent photovoltaic system," *International Journal of Photoenergy*, vol. 2021, Article ID 7894849, 8 pages, 2021.
- [9] C. H. Chen, T. P. Hong, and V. S. Tseng, "Fuzzy data mining for time-series data," *Applied Soft Computing*, vol. 12, no. 1, pp. 536–542, 2012.
- [10] U. Subramaniam, M. M. Subashini, D. Almakhles, A. Karthick, and S. Manoharan, "An expert system for COVID-19 infection tracking in lungs using image processing and deep learning techniques," *BioMed Research International*, vol. 2021, Article ID 1896762, 17 pages, 2021.
- [11] K. Alagar and S. Thirumal, "Standalone PV-wind-DG-battery hybrid energy system for zero energy buildings in smart city Coimbatore, India," in *Advanced Controllers for Smart Cities*, pp. 55–63, Springer, Cham, 2021.
- [12] B. Dennis and S. Muthukrishnan, "AGFS: adaptive genetic fuzzy system for medical data classification," *Applied Soft Computing*, vol. 25, pp. 242–252, 2014.
- [13] E. Hullermeier, "Fuzzy sets in machine learning and data mining," *Applied Soft Computing*, vol. 11, no. 2, pp. 1493–1505, 2011.
- [14] I. Ebtehaj and H. Bonakdari, "Bed load sediment transport estimation in a clean pipe using multilayer perceptron with different training algorithms," *Environmental Engineering*, vol. 20, no. 2, pp. 581–589, 2016.
- [15] E. Me and O. Unold, "Mining fuzzy rules using an artificial immune system with fuzzy partition learning," *Applied Soft Computing*, vol. 11, no. 2, pp. 1965–1974, 2011.
- [16] F. Moradi, H. Bonakdari, O. Kisi, I. Ebtehaj, J. Shiri, and B. Gharabaghi, "Abutment scour depth modeling using neuro-fuzzy-embedded techniques," *Marine Georesources & Geotechnology*, vol. 32, 2019.
- [17] F. Martínez, F. Charte, M. P. Frías, and A. M. Martínez-Rodríguez, "Strategies for time series forecasting with generalized regression neural networks," *Neurocomputing*, 2021.
- [18] S. F. F. Mojtahedi, I. Ebtehaj, M. Hasanipanah, H. Bonakdari, and H. B. Amnieh, "Proposing a novel hybrid intelligent model for the simulation of particle size distribution resulting from blasting," *Engineering with Computers*, vol. 35, no. 1, pp. 47–56, 2019.
- [19] F. B. Ta and S. E. Mb, "An efficient CRM-data mining framework for the prediction of customer behaviour," *Procedia Computer Science*, vol. 46, pp. 725–731, 2015.
- [20] F. Stahlberg, "Neural machine translation: a review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.
- [21] G. Tewary, "Effective data mining for proper mining classification using neural networks," *International Journal Of Data Mining & Knowledge Management Process (IJDKP)*, vol. 5, no. 2, 2020.
- [22] H. K. Bhuyana and N. K. Kamila, "Privacy preserving subfeature selection in distributed data mining," *Applied Soft Computing*, vol. 36, pp. 552–569, 2015.
- [23] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan and Kaufmann, 2nd edition, 2005.
- [24] I. Khan and A. Kulkarni, "Knowledge extraction from survey data using neural networks," *Procedia Computer Science*, vol. 20, pp. 433–438, 2013.
- [25] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall, USA, 2012.
- [26] L. Lu, F. Goerlandt, O. A. Valdez Banda, and P. Kujala, "Developing fuzzy logic strength of evidence index and application in Bayesian networks for system risk management," *Expert Systems with Applications*, vol. 192, article 116374, 2022.
- [27] M. P. Singh, "Two phase learning technique in modular neural network for pattern classification of handwritten Hindi alphabets," *Machine Learning with Applications*, vol. 6, pp. 100174–108270, 2021.
- [28] M. Mitra and R. K. Samanta, "Cardiac arrhythmia classification using neural networks with selected features," *Procedia Technology*, vol. 10, pp. 76–84, 2013.

- [29] M. H. Mohamed, "Rules extraction from constructively trained neural networks based on genetic algorithms," *Neurocomputing*, vol. 74, no. 17, pp. 3180–3192, 2011.
- [30] N. Upasania and H. Om, "Evolving fuzzy min-max neural network for outlier detection," *Procedia Computer Science*, vol. 45, pp. 753–761, 2015.
- [31] O. Khayat, M. M. Ebadzadeh, H. R. Shahdoosti, R. Rajaei, and I. Khajehnasiri, "A novel hybrid algorithm for creating self-organizing fuzzy neural networks," *Neurocomputing*, vol. 73, no. 1-3, pp. 517–524, 2009.
- [32] P. Zhanga and H. Wang, "Fuzzy wavelet neural networks for city electric energy consumption forecasting," *Energy Procedia*, vol. 17, pp. 1332–1338, 2012.
- [33] K. Rajeswari, V. Vaithyanathan, and T. R. Neelakantan, "Feature selection in ischemic heart disease identification using feed forward neural networks," *Procedia Engineering*, vol. 41, 2012.
- [34] J. Ravnik, A. Jovanovac, N. Trupej, and M. Vistica, "A sigmoid regression and artificial neural network models for day-ahead natural gas usage forecasting," *Cleaner and Responsible Consumption*, vol. 3, article 100040, 2021.
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [36] S.-B. Roha, S.-K. Ohb, and W. Pedrycz, "Design of fuzzy radial basis function-based polynomial neural networks," *Fuzzy Sets and Systems*, vol. 185, pp. 15–37, 2011.
- [37] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Medical data classification using interval type-2 fuzzy logic system and wavelets," *Applied Soft Computing*, vol. 30, pp. 812–822, 2015.
- [38] G. Vinodhini and R. M. Chandrasekaran, "A comparative performance evaluation of neural network based approach for sentiment classification of online reviews," *Journal of King Saud University-Computer and Information Sciences*, vol. 28, 2014.
- [39] Y. F. Wanga, D. H. Wangb, and T. Y. Chai, "Active control of friction self-excited vibration using neuro-fuzzy and data mining techniques," *Expert Systems with Applications*, vol. 40, no. 4, pp. 975–983, 2013.

Research Article

Cross-Sectional Analysis of Impulse Indicator Saturation Method for Outlier Detection Estimated via Regularization Techniques with Application of COVID-19 Data

Sara Muhammadullah ¹, Amena Urooj,¹ Muhammad Hashim Mengal,² Shahzad Ali Khan,³ and Fereshteh Khalaj ⁴

¹Department of Economics and Econometrics, Pakistan Institute of Development Economics, Islamabad, Pakistan

²World Health Organization, Pakistan

³Vice-Chancellor of Health Services Academy Islamabad, Pakistan

⁴Department of Mathematics and Statistics, Parand and Robat Karim Branch, Islamic Azad University, Tehran, Iran

Correspondence should be addressed to Fereshteh Khalaj; fekhaj@gmail.com

Received 21 December 2021; Revised 23 February 2022; Accepted 22 March 2022; Published 6 May 2022

Academic Editor: Diego Pinto

Copyright © 2022 Sara Muhammadullah et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Impulse indicator saturation is a popular method for outlier detection in time series modeling, which outperforms the least trimmed squares (LTS), M-estimator, and MM-estimator. However, using the IIS method for outlier detection in cross-sectional analysis has remained unexplored. In this paper, we probe the feasibility of the IIS method for cross-sectional data. Meanwhile, we are interested in forecasting performance and covariate selection in the presence of outliers. IIS method uses Autometrics techniques to estimate the covariates and outlier as the number of covariates $P > n$ observations. Besides Autometrics, regularization techniques are a well-known method for covariate selection and forecasting in high-dimensional analysis. However, the efficiency of regularization techniques for the IIS method has remained unexplored. For this purpose, we explore the efficiency of regularization techniques for out-of-sample forecast in the presence of outliers with 6 and 4 standard deviations (SD) and orthogonal covariates. The simulation results indicate that SCAD and MCP outperform in forecasting and covariate selection with 4 SD (20% and 5% outliers) compared to Autometrics. However, LASSO and AdaLASSO select more covariates than SCAD and MCP and possess higher RMSE. Overall, regularization techniques possess the least RMSE than Autometrics, as Autometrics possesses the least average gauge at the cost of the least average potency. We use COVID-19 cross-sectional data collected from 1 July 2021 to 30 September 2021 for real data analysis. The SCAD and MCP select CRP level, gender, and other comorbidities as an important predictor of hospital stay with the least out-of-sample RMSE of 7.45 and 7.50, respectively.

1. Introduction

The ordinary least squares (OLS) approach has been a widely chosen technique among the numerous available methods in regression analysis because it is computationally straightforward and possesses the best linear unbiased estimate. However, it possesses strong assumptions on the distribution of error (ε) termed as $\varepsilon \sim N(0, \sigma)$, which is usually violated while dealing with real data analysis. The leading cause of distortion is outliers, which violates the nor-

malinity assumption of residuals. Outlying data in the dependent and regressor variables pose a risk to least squares regression since they might negatively impact the estimate if they go unreported. Even cross-sectional data with high quality contain outliers; however, it is rare in time series economic data (because of the differencing variables) [1].

Robust regression techniques are used significantly in the literature of outlier presences. Langford and Lewis [2] well defined an outlier as data points that look inconsistent with the rest of the data. Such influential points are

frequently concealed from the user since they do not always appear in the standard least-squares residual graph [3]. Zaman et al. [1] indicate that the OLS residuals are ineffective in finding outliers in small and big sample sizes, whereas Rousseeuw and Leroy [4] demonstrate several real data sets in which the OLS residuals miss to detect any outliers despite significant outliers. However, new statistical procedures have been proposed that are less susceptible to outliers; Rousseeuw [5] introduced the primary feasible robust regression estimators (least median squares (LMS), least trimmed squares (LTS), and variations) that perform correctly even when a high number of outliers are present. Huber M estimation, MM estimation, least absolute value method (LAV), and S estimation are examples of robust approaches [6–8]. A conspicuous technique is established on Huber’s M-estimators, which offer robustness in location parameters. Regrettably, generalizations to regression models miss the mark to accomplish robustness. As Rousseeuw [5] illustrates, regression M-estimators likewise have a 0% breakdown value. The generalization of MM-estimators likewise fails to attain large breakdown values. A direct method to robust regression is to use LTS analysis in huge residuals. The LTS analysis discards outlying observations and then can run a standard OLS regression, proposed in Rousseeuw [5]. However, removing too many data points in the case of too many outlier observations turns the risk of the final regression model not reflecting the association that the econometrician wants to assess [1].

On the contrary, Doornik [9] and Johansen and Nielsen [10] illustrate the impulse indicator saturation (IIS) as a robust estimator. Similarly, Johansen and Nielsen [10] describe and demonstrate that a split-sample estimator for the indicator-saturated regression model is a one-step M-estimator that is iterated twice. Doornik [9] illustrates that robustified least squares and indicator saturation are more efficient than least trimmed squares. When the covariates are static and only outliers occur in the dependent variable’s data, M estimation works effectively. The impulse indicator saturation method was initially designed to detect unidentified numbers of outliers with indefinite magnitudes at uncertain points in the sample, together with the start and end of observations [11]. However, the step indicator saturation (SIS) method is a modified version of IIS techniques for multiple break detection. Indicator saturation (IS) is used as a border term that detects outlier (via IIS) and multiple break shifts (via SIS) and simultaneously estimates the underlying modeling [9–13].

As the IS method possesses the number of candidates regressor more than the number of data points, the OLS estimates fail to estimate the thriving model. However, Autometrics handles such phenomena efficiently regardless of candidate regressors exceeding the number of observations; due to this reason, IS method is feasible to estimate via Autometrics. Autometrics uses extending and contracting multiple-path search algorithms with user-specified significance levels through the model selection process. However, the choice of the significance level is the trade-off between the irrelevant and relevant dummy indicators or regressors, with tight significance level (0.001) significant variable omit-

ted in the final model whereas, with 0.05 significance level, the model consists of irrelevant regressors [13–15].

Other than Autometrics, regularization techniques are emerging techniques when the number of covariates excel the number of data points (observations); some of these popular techniques are Least Absolute Subset Selection Operator (LASSO), Adaptive LASSO, Smoothly Clipped Absolute Deviations (SCAD), and Minimax Concave Penalty (MCP) [16–19]. However, every few studies compare the computational efficiencies of Autometrics with regularization techniques [20] [21–23] for covariate selection and forecasting under the normality assumption. They do not consider outliers with the IIS setup. As it is challenging to choose the level of significance for thriving models in Autometrics, the regularization techniques can be used as an alternative model selection method in this case. Up to date, the prevailing studies do not compare the computational efficiency of regularization techniques with Autometrics in cross-sectional analysis with outlier in IIS setup. This study is aimed at analyzing the computational efficiency of regularization techniques with IIS setup in cross-sectional phenomena. The computational proficiency of these methods is evaluated with potency, gauge, and out-of-sample Root Mean Square Error (RMSE) in the simulation experiment. For the simulation experiment, the Data Generating Process (DGP), we opt with the orthogonal regressors and possess three scenarios 5%, 10%, and 20% outlying observations with 4 and 6 standard deviation (SD). Meanwhile, in DGP, we intake orthogonal cases for this purpose we use some well-known orthogonal techniques of regularization like LASSO, Adaptive LASSO, Smoothly Clipped Absolute Deviation (SCAD), and Minimax Concave Penalty (MCP) [16–19].

Outlier detection is a rapidly developing procedure in the healthcare and medical data industries, and it is a significant source of concern. Hauskrecht et al. [24] study data-driven outlier-based surveillance and forewarning system that uses data from former patient cases. Wilson et al. [25] used the outlier identification method for hypoglycemia safety in patients, calculating a flair outlier value within a year, comparator group, and A1c threshold while considering at hazard population proportions. Jyothi et al. [26] used outlier detection in healthcare data, a key source of concern for health insurers. The development of a Supervised Outlier Detection Approach in Healthcare Claims (SODAC) and carried out in two parts. Noma et al. [27] offer optimal effect measures for network meta-analysis models with mislaid outcomes and appropriate degree of freedom adjustments. The real data application of the IIS method in healthcare and medicine with outliers for cross-sectional analysis does not exist in the current literature [24–30]. To probe the efficacy of the IIS method estimated via regularization techniques for real data techniques, we use COVID-19 cross-sectional surveillance data, which has been collected from July 2021 to 30 September 2021 in Isolation Hospital and Infectious Treatment Center (IHITC) Islamabad. We aim to analyze the factors associated with prolonging the length of hospital stay of COVID-19 patients in the capital territory of Islamabad.

2. Outlier Detection and Model Selection Techniques

2.1. Impulse Indicator Saturation. Impulse indicator saturation is a popular method of outlier detection as it already dominates the existing outlier selection techniques like least trimmed squares (LTS), M-estimator, and MM-estimator [9, 10]. Usually, in multivariate regression, we assume that error is normally distributed, which is usually violated in real data analysis. In the equation below, we assume that error is not normally distributed, and α is the intercept of the model, y is the continuous dependent variable, and x_{ji} is the orthogonal regressors, where $j = 1, 2, 3, \dots, k$ number of orthogonal regressors and $i = 1, 2, 3, \dots, n$ observations.

$$y_i = \alpha + \sum_{j=1}^k \beta_j x_{ji} + \varepsilon_i. \quad (1)$$

As in equation (1), the error is not normally distributed due to the presence of an outlier; in this case, the IIS method introduces an impulse dummy indicator to each of the data points, and the above equation would be

$$y_i = \alpha + \sum_{j=1}^k \beta_j x_{ji} + \sum_{i=1}^n \gamma_i I + \varepsilon_i, \quad (2)$$

where

$$I = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (3)$$

Here, I is an identity matrix of each corresponding observation in the above equation. $I'_1 = (1, 0, 0, \dots, 0)$, $I'_2 = (0, 1, 0, 0, \dots, 0)$, and $I'_i = (0, 0, 0, \dots, 1)$. The OLS estimate is not feasible to estimate the above Generalized Unrestricted Model (GUM). Estimating the above equation is possible because Autometrics (created on general-to-specific modeling) is used to detect the outlier and estimate the model instantaneously. In the general-to-specific methodology, each observation would have one dummy variable, and additional exogenous variables can be considered that possibly distress the dependent variable [10, 12].

2.2. Model Selection Methods. There are two main fields of model selection methods when covariates are higher than the number of data points: the regularization technique and the classical (general-to-specific, Autometrics) approach. The classical method (Autometrics) is initiated by a saturated model and uses the multipath search process to eliminate insignificant covariates. The model selection is primarily dependent on the preset significance threshold. On the other hand, the regularization approach applies spar-

sity to the p -dimensional vector of parameters, resulting in numerous parameters of covariates equal to zero. This approach resolves the issues that arise in high dimensionality. We go through each of these methods further; however, we only looked at orthogonal regularization approaches.

2.2.1. Autometrics. The general-to-specific model procedure, presented by Hoover et al. [31], combines several components of Krolzig and Hendry [32]. PcGets is a second-generation extension of general-to-specific method; it extends and clarifies Hoover and Perez's methodology [32, 33]. Modifying the existing techniques, Doornik [9] introduced Autometrics which is based on the same concept of general to specialized (gets) modeling. Autometrics is a third-generation algorithm based on the same concept of PcGets.

Autometrics employs a tree path search that includes multistep simplifications along several pathways. The GUM contains all covariates at first and estimates them using the OLS technique, removing statistically insignificant covariates; the compact model's reliability is tested at each individual stage to guarantee consistency with the test diagnostics. Autometrics employs a tree-path exploration strategy that involves multiple multistep simplifications. The ultimate models are constructed that used a tree-path approach and assessed using screening procedures; the parameters are automatically eliminated if the parameter estimates are statistically irrelevant. Autometrics retests their union once a high number of terminal models are discovered. A novel GUM is formed once the "surviving" terminal models are merged, permitting another tree-path search repetition. The whole search procedure is completed by reexamining the terminal models and their consolidations. If a large number of models pass all of the tests, the final decision is made on specified information criteria.

The test diagnostics are being used to ensure the simple models, whereas inclusive tests are used to resolve several terminal models. Epprecht et al. [20] argue that Autometrics is a kind of black box technique. While developing modeling techniques, the user can select among 1-cut and tight significance level and nominal significance level. The multipath technique in Autometrics identifies multiple breaks/outliers more effectively and has reduced estimator variance [34]. The multipath technique eliminates path reliance by employing a tree structure and alike stepwise sequential backward, an integral function of the gets package in R software [15].

2.2.2. Regularization Techniques. Other than Autometrics, regularization approaches manage saturated models with irrelevant variables even if the amount of regressors excel the quantity of data points (observations), shrinking the irrelevant parameters to zero with a nearly biased estimate. The Least Absolute Shrinkage and Selection Operator (LASSO) was introduced by Tibshirani [17]. It is a standard estimation method in a linear regression framework due to its decreased computing cost. The LASSO does not hold an oracle property; Zou [19] proposed Adaptive LASSO. The

regularization penalty is defined in

$$\hat{\gamma}_j = \operatorname{argmin}_{\gamma} \sum_{i=1}^n \left(y_i - \alpha - \sum_{j=1}^k \beta_j x_{ji} - \gamma_i I \right) + \lambda p(|\gamma_j|). \quad (4)$$

In the above equation, y is a continuous dependent variable, x is an orthogonal covariate, and I is the impulse dummy for outliers. The following regularization techniques contemplate different choices for the penalty function, which is summarized in Table 1.

$$(\text{Lasso}) p_{\lambda_j}(|\gamma_j|) = \lambda_j |\gamma_j|. \quad (5)$$

The “L1 penalty” for the LASSO estimator is the subsequent term in the preceding equation, and it primes to a sparse solution with a very precise set of parameters exactly equivalent to zero through a particular level of bias. The choice of λ determines the quantity of reduction, and it varies from $0 < \lambda < \infty$.

Zou [19] revealed that the LASSO method violated the oracle property and proposed the Adaptive LASSO as a modest and effective alternative. On the other hand, the coefficients in LASSO are altogether penalized similarly in the “L1 penalty.” Nevertheless, in the AdaLASSO method, individual parameter is assigned its own weight. Zou [19] demonstrated that if the weights are data-dependent and correctly set, the AdaLASSO may have the best outcomes and exhibit the oracle property.

$$(\text{Adaptive Lasso}) p_{\lambda_j}(|\gamma_j|) = \lambda_j w_j |\gamma_j|, \quad \text{where } w_j = |\hat{\gamma}_j^*|^{-\tau}. \quad (6)$$

$\hat{w}_j = 1/|\hat{\gamma}_j^*|^\tau$, $\tau > 0$, and $\hat{\gamma}_j^*$ is a preliminary parameter estimate. The weights of irrelevant parameters approach infinity as the sample increases, whereas relevant parameters approach a finite constant. Zou [19] suggested using the OLS technique to estimate $\hat{\gamma}_j^*$. On the other hand, the OLS approach does not work as soon as the amount of candidate regressors excel the quantity of data points (observations). A ridge estimate might be used as a preliminary estimator in this scenario.

Fan and Li [16] introduced a new approach that satisfied the condition of unbiased, sparsity, and continuity known as Smoothly Clipped Absolute Deviation (SCAD).

$$\text{SCAD} = \lambda \left\{ \begin{array}{l} |\gamma| \text{ if } |\gamma| \leq \lambda, \\ -\frac{(\gamma^2 - 2a\lambda|\gamma| + \lambda^2)}{2(a+1)\lambda} \text{ if } \lambda < |\gamma| \leq a\lambda \text{ and} \\ \frac{1}{2}(a+1)\lambda \text{ if } |\gamma| \geq a\lambda \end{array} \right\}. \quad (7)$$

Distinct to LASSO, SCAD uses two tuning parameters α and λ ; $P(\gamma | \lambda, \alpha)$ of SCAD method is known as folded con-

TABLE 1: Regularization penalties.

Method	Penalty function
LASSO	$P(\cdot) = \sum_{j=1}^k p_{\lambda_j}(\gamma_j)$
AdaLASSO	$P(\cdot) = \lambda \sum_{j=1}^k \hat{w}_j \gamma_j $
SCAD	$P(\cdot) = \sum_{j=1}^k p_j(\gamma_j ; \lambda; \alpha)$
MCP	$P(\cdot) = \sum_{j=1}^k p_j(\gamma_j ; \lambda; \alpha)$

$p_{\lambda_j}(\cdot)$ is a function denoted as penalty function, and λ_j is the function parameter.

cave penalty that depends on λ in a nonmultiplicative way; hence, $\lambda P(\alpha) = P(\alpha | \lambda)$. In addition, the tuning parameter (λ) affects the penalty’s concavity. The objective function’s intensification is determined by λ and α , λ being chosen via cross-validation and α is fixed equal to 3.7 [16].

Zhang [18] proposed the Minimax Concave Penalty (MCP), a nonconvex regularization approach that uses spares zone up to a specified choice of threshold to produce an unbiased estimate.

$$\text{MCP} = \lambda \left\{ \begin{array}{l} \left(\lambda - \frac{|\gamma|}{\alpha} \right) \operatorname{sign}(\gamma) \text{ if } |\gamma| \leq \alpha\lambda \\ 0 \text{ if } |\gamma| > \alpha\lambda \end{array} \right\}. \quad (8)$$

MCP employs the $p_j(|\gamma_j|; \lambda; \alpha)$ regularization pathway, which is constructed on a family of nonconvex penalty functions through two tuning parameters λ and α , whereas α is constant and λ is chosen by cross-validation. The λ tuning parameter regulates the degree of penalty shrinking and concavity. Because the maximum concavity is minimized, MCP minimizes the convexity of the spares to a greater extent [18]. SCAD and MCP estimates fall to the folded concave penalty family since the $P(\cdot)$ penalty function is neither convex nor concave.

2.2.3. Selection Criteria for Tuning Parameter. The selection of tuning parameter is critical since it determines the complication of the chosen model. The selection of the suitable tuning parameters results in a compact model with accurate forecast performance. In order to achieve prediction optimality, the tuning parameter is commonly selected by a cross-validation technique. The aim is to retrieve the primary collection of sparse covariates. Covariate selection typically needs a more substantial penalty parameter than optimum prediction [35]. The information criteria like Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) are used as another approach for penalizing the likelihood through the degrees of freedom of the fitted model. Degrees of freedom are frequently used to measure the complication of a model fit, and we can use them to

TABLE 2: Simulated results with different percentages of outliers with 6 SD.

	20% outliers	
	Gauge	Potency
SCAD	0.222	0.367
MCP	0.222	0.367
LASSO	0.611	0.767
AdaLASSO	0.333	0.433
Auto(0.05)	0.011	0.100
Auto(0.01)	0.011	0.100
	10% outliers	
SCAD	0.100	0.500
MCP	0.140	0.550
LASSO	0.650	0.850
AdaLASSO	0.220	0.600
Auto(0.05)	0.010	0.200
Auto(0.01)	0.000	0.200
	5% outliers	
SCAD	0.048	0.600
MCP	0.048	0.600
LASSO	0.591	0.933
AdaLASSO	0.124	0.667
Auto(0.05)	0.000	0.534
Auto(0.01)	0.000	0.534

TABLE 3: Simulated results with different percentages of outliers with 4 SD.

	20% outliers	
	Gauge	Potency
SCAD	0.222	1.000
MCP	0.144	1.000
LASSO	0.611	0.967
AdaLASSO	0.189	0.933
Auto(0.05)	0.000	0.367
Auto(0.01)	0.011	0.367
	10% outliers	
SCAD	0.230	0.600
MCP	0.150	0.550
LASSO	0.650	0.850
AdaLASSO	0.360	0.700
Auto(0.05)	0.000	0.500
Auto(0.01)	0.000	0.500
	5% outliers	
SCAD	0.114	0.667
MCP	0.095	0.667
LASSO	0.657	0.867
AdaLASSO	0.352	0.667
Auto(0.05)	0.000	0.667
Auto(0.01)	0.000	0.667

decide how much regularization to utilize. Meanwhile, in terms of covariate selection and out-of-sample forecast, WLAdaLASSO with a BIC-based tuning parameter possesses optimal results [23, 36].

$$\text{BIC} = n \log(\hat{\sigma}^2) + \log(n) + df(\hat{y}), \quad (9)$$

whereas $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $df(\hat{y})$ signifies the degrees of freedom of the fitted model. The BIC-based tuning parameter, on the other hand, is superior to cross-validation for covariate selection, although there is no theoretical justification [35]. Henceforth, the BIC-based tuning parameter is used for outlier and covariate selection in simulation and real data analysis.

2.3. Theoretical Assessment. The study is aimed at evaluating the out-of-sample forecasting performance of regularization methods in the presence of an outlier in the IIS setup. However, other than out-of-sample RMSE, we also emphasize the average gauge and potency in the simulation study. Gauge is defined as the empirical null retention frequency of how insignificant variables/outliers are reserved, whereas potency is identified as correct covariate/outlier identifications. The assessment of regularization methods and Automatics was evaluated via an accurate zero identification taken as potency and improper zero identification denoted as gauge [37]. If the considered techniques appropriately classify the model, the evaluations of the subsequent parameters should be expected:

- (1) The gauge is getting close to the significance level (0.05) or the tight significance level (0.01 or 0.001)

$$E\left(\frac{\widehat{k}_{\text{irr}}}{k_{\text{irr}}}\right) \longrightarrow \alpha. \quad (10)$$

- (2) When estimating techniques are used to estimate the exact model efficiently, potency approaches 1

$$E\left(\frac{\widehat{k}_{\text{rel}}}{k_{\text{rel}}}\right) \longrightarrow 1. \quad (11)$$

For out-of-sample RMSE, we randomly trained our model on 90% of observations, and 10% of observations were discarded to test the model's accuracy in terms of RMSE [23]. The RMSE of regularization techniques, even in an outlier, is expected to be smaller than Automatics. However, LASSO will retain more regressor variables than SCAD, MCP, and Automatics.

3. Data Generating Process and Simulation Experiment Result

The Data Generating Process (DG) in this section has opted from [9] where the models consist of irrelevant regressors

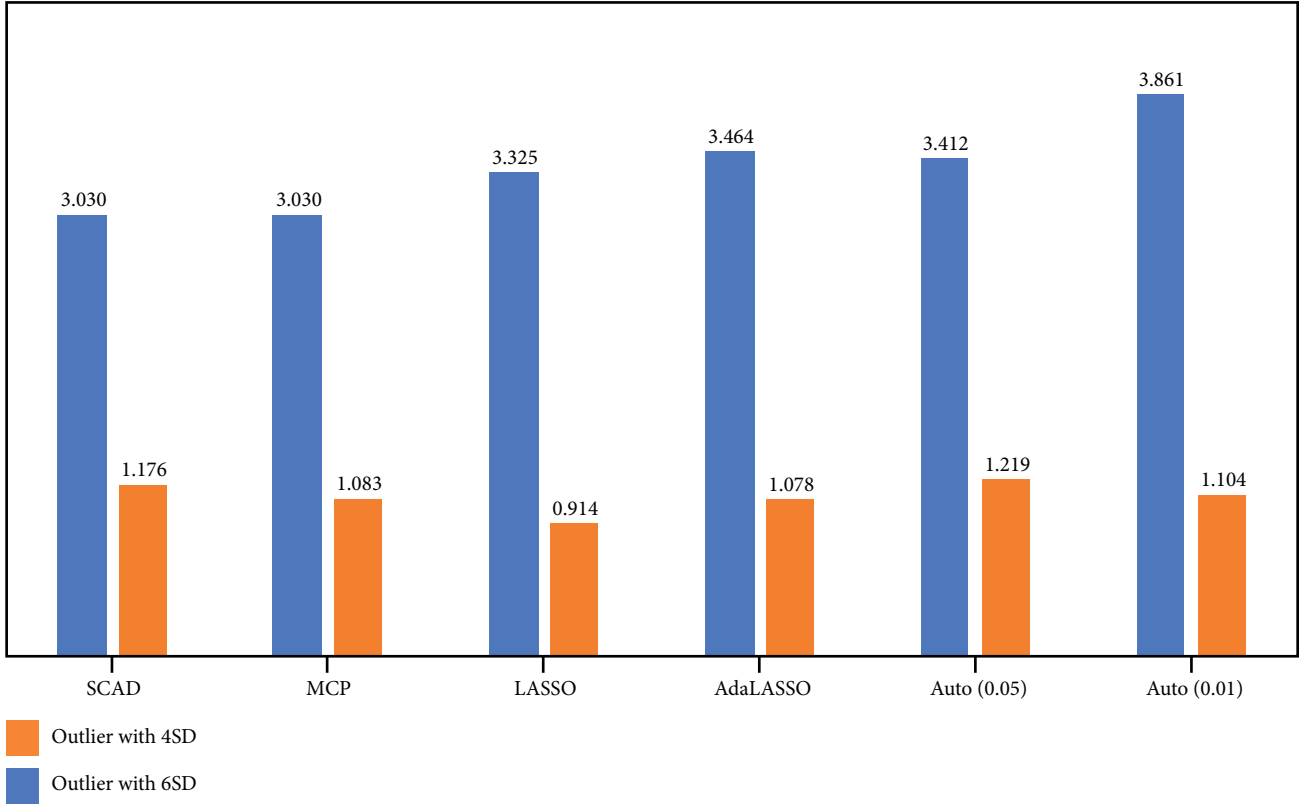


FIGURE 1: Average RMSE with less than 5% outliers.

and outliers. We assumed well scatter outlier among DGP with 5%, 10%, and 20% observations, which is different from Doornik [9], as it has been illustrated 20% outlier at the end of observations with magnitude coefficients equal to 6 in the static DGP, where the DGP can be defined as

$$y_i = 0.1 + \sum_{j=1}^k \beta_j x_{ji} + 6(\tau) + \varepsilon_i, \quad (12)$$

where $\beta_1 = \dots = \beta_{k^*} = 1$ whereas k^* is equal to 10 and the rest of the other beta coefficients equal zero, and $k = 20$ with $i = 1, 2, \dots, 100$ observations. The regressors $x_{ji} \sim \text{IID}(0, 1)$ and $\varepsilon \sim \text{IID}(0, 1)$, whereas the outlying observations (τ) are equal to 5%, 10%, and 20% with 6 SD and 4 SD of error term. To estimate the above DGP, we use the Generalized Unrestricted Model (GUM) and introduce an impulse dummy indicator for each observation in the model. The experiment is repeated 1000 times.

3.1. Simulation Experiment Result. The comparison is assessed under scenarios with 5%, 10%, and 20% scattered outliers with 6 SD and 4 SD. The glmnet package for R software is used to estimate LASSO and AdaLASSO. For MCP and SCAD estimation, we use the ncvreg package of R; the ncvreg package uses a coordinate descent algorithm, while for Autometrics we use the gets package of R. To achieve our study objective, we use a static DGP with orthogonal covariates and dummy indicator saturation opts from Doornik [9]. It provides a convenient base for comparing

regularization techniques with Autometrics in the presences of outliers. Outcomes of the simulated scenarios are obtainable in Table 2. Table 2 illustrates the average gauge and potency Autometrics and regularization techniques; however, the RMSE error of the out-of-sample forecast has been presented below. We use Auto as an acronym of Autometrics in the tables and figures, and the computational efficiency of Autometrics is assessed with 0.05 and 0.01 significance levels.

Table 2 demonstrates the results of regularization techniques with Autometrics for covariate selection and outlier detection in potency and gauge. The result indicates that with a 20% and 6 SD outlier, Autometrics performs worse in average potency among all existing techniques. On the contrary, LASSO possesses the highest gauge and potency among regularization techniques. Meanwhile, SCAD and MCP accomplish similar performance in both average gauge and potency. The simulation result specifies that as the outlier percentage decreases to 10%, the performance of considered methods increases in average potency. However, the performance of SCAD and MCP improved with both gauge and potency. With 5% outlying observation, the considered techniques improved further. The SCAD and MCP estimate retains 60% average potency with an average gauge equal 5%.

In Table 3, the result indicates that with 20% and 4 SD outliers, Autometrics performs worse among all existing techniques in average potency; however, the average potency of SCAD and MCP drastically increased compared to outliers with 6 SD. Meanwhile, significant improvement in the

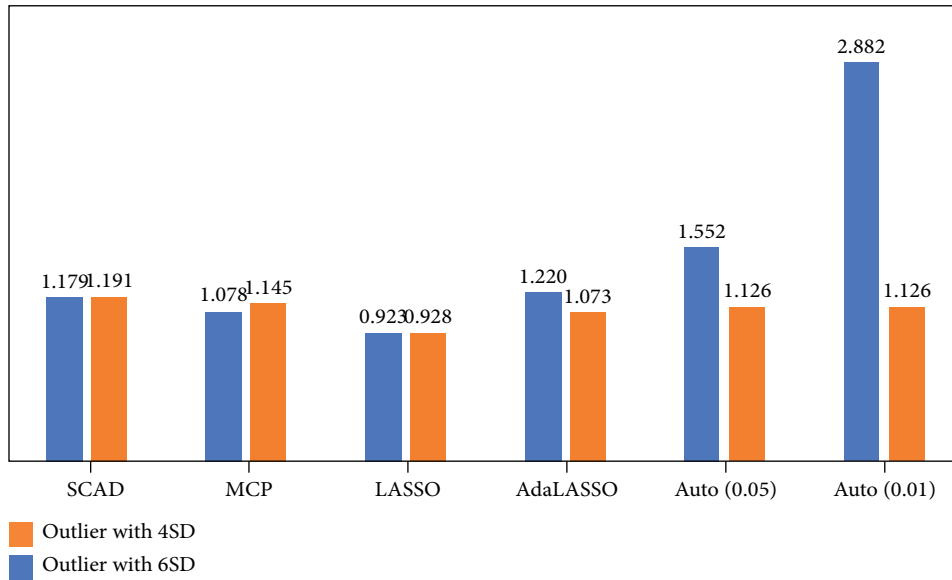


FIGURE 2: Average RMSE with 10% outliers.

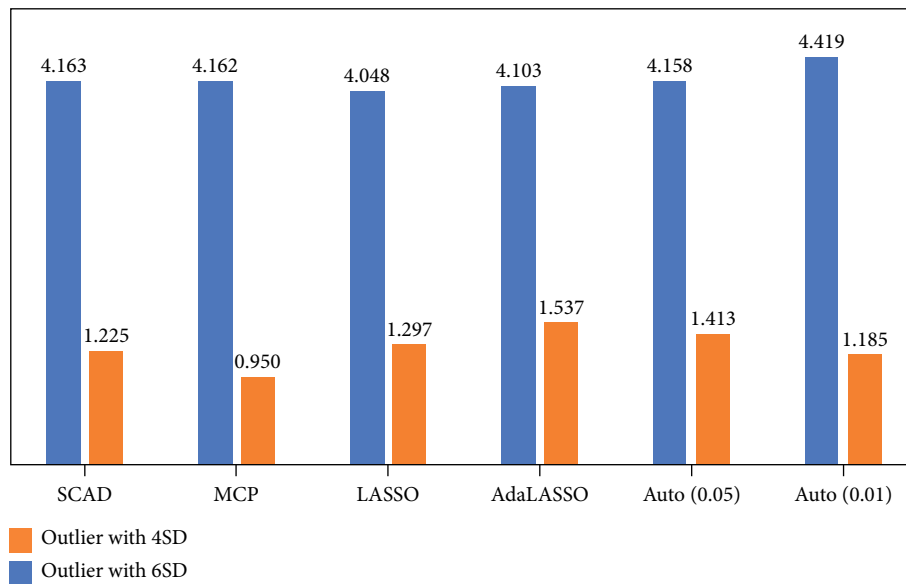


FIGURE 3: Average RMSE with 20% outliers.

average potency of the regularization technique with 4 SD outlier has been observed over 6 SD, whereas the performance of the average gauge remains the same in both seniors. On the contrary, LASSO possesses the highest gauge and potency among regularization techniques, similar to outliers with 6 SD. Compared to LASSO and SCAD, MCP performs significantly in gauge equal to 0.095 and 0.114 of SCAD with a 5% outlier. The simulation result shows that as the outlier percentage decreases to 10%, the performance of considered regularization methods decreases in average potency, whereas the average gauge remains similar to 20% outliers.

Overall, the simulation result indicates that outliers with 4 SD and 5% outlying observation regularization techniques perform better than 6 SD outliers in terms of average

potency, whereas the average gauge of regularization techniques with 6 SD is lower than 4 SD outliers. The Auto-metrics possesses the least average gauge in all scenarios (5%, 10%, and 20%, 6 SD and 4 SD) at the rate of the smallest average potency among all considered techniques. In contrast, LASSO possesses the highest potency and gauge of all other methods.

Figures 1–3 represent the out-of-sample forecasting performance of the considered methods. The graphs illustrate that the average RMSE error of LASSO with 20% and 10% outlier observations is the least among all considered techniques. The result aligns with existing literature as LASSO possesses the least forecasting error and selects a more irrelevant regressor (which can be observed from Table 1) [38]. However, with less than 5% outlier observations, the SCAD

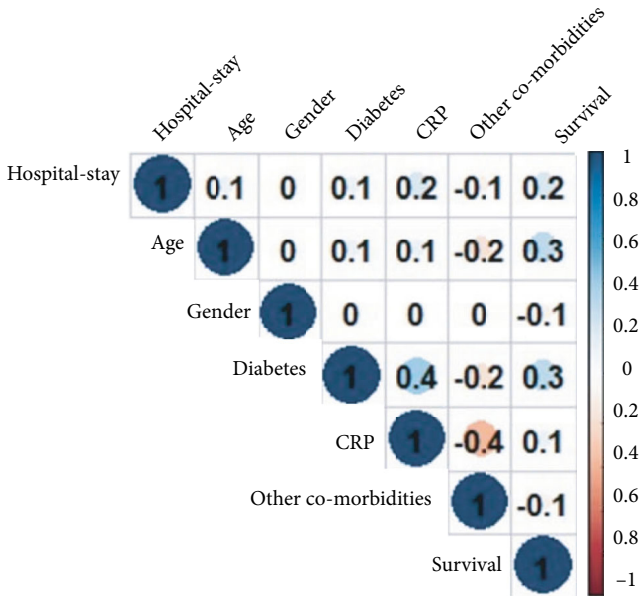


FIGURE 4: Correlation graph.

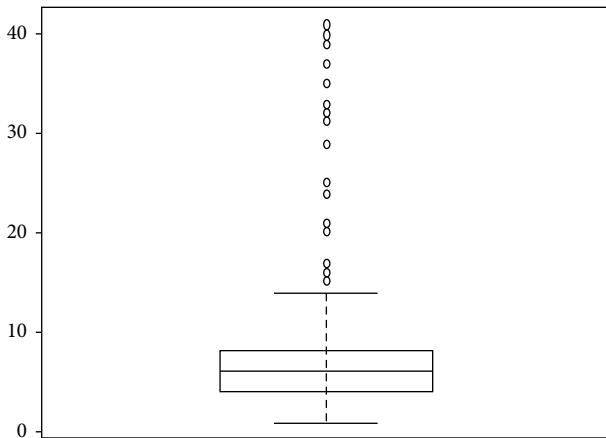


FIGURE 5: Box plot of hospital stay.

and MCP possess the least RMSE 3.03 than all other techniques, even less than Autometrics. We observed that Autometrics with 5% outliers possesses the least gauge but retain higher RMSE than SCAD and MCP. Autometrics with 0.05 level of significance possesses the least RMSE than 0.01 level of significance, the fact that Autometrics with 0.01 level of significance omits relevant regressors which increases the average RMSE.

There is a significant improvement in average RMSE with 4 SD with 5% and 20% outliers compared to 6 SD with 5% and 20% outliers. This difference can be justified as with 5% and 4 SD outliers, the average potency is higher (means that method correctly identified the correct variables/dummy indicator) compared to 6 SD, which ultimately impact the out-of-sample RMSE, and the same pattern can be observed with 20% outliers and 6 SD the average potency is least due to this reason the out-of-sample RMSE increases. However, the average potency of 20% outliers with 4 SD is close to 1 for regularization techniques; due to this, the

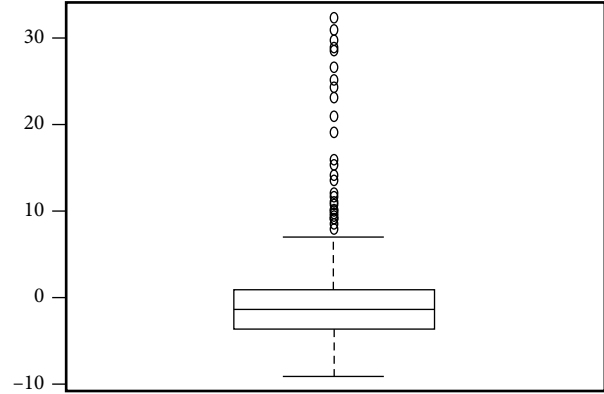


FIGURE 6: Residual box plot of linear regression.

TABLE 4: Real data analysis with covariate selection and number of selected outliers.

SCAD number of selected outliers (28)				
Variable	Gender	CRP level	Other comorbidities	
Coefficient	0.24463	0.00083	0.20533	
MCP number of selected outliers (31)				
Variable	Gender	CRP level	Other comorbidities	
Coefficient	0.22493	0.0004	0.2585	
LASSO number of selected outliers (204)				
Variable	Age	Gender	CRP level	Other comorbidities
Coefficient	0.00225	0.55747	0.00282	1.3966
Auto(0.05) number of selected outliers (14)				
Variable	CRP level	Other comorbidities		
Coefficient	0.00766	0.9653		

out-of-sample RMSE of regularization techniques is the least compared to 6 SD, as shown in Figure 3. On the contrary, as 10% and 4 SD and 10% and 6 SD outliers, the performance of considered methods is aligned in average potency, and consequently, the average RMSE are almost similar observed in Figure 2.

4. Real Data Analysis

Coronavirus disease 2019 (COVID-19) is a global outbreak triggered by coronavirus 2, which origins severe acute respiratory illness (SARS-CoV-2). The World Health Organization declared COVID-19 a pandemic in March 2020. Meanwhile, the confirmed number of cases around the globe has been reported as 504,079,039, with 6,204,155 fatalities as of April 20, 2022 (<https://covid19.who.int>). However, Pakistan is not among the nations with the uppermost number of COVID-19 cases and fatalities. The initial case of COVID-19 was identified in Pakistan on February 25, 2020. Up to April 20, 2022, 1,527,411 COVID cases had been reported, with 30,364 fatalities (<https://covid19.who.int/region/emro/country/pk>).

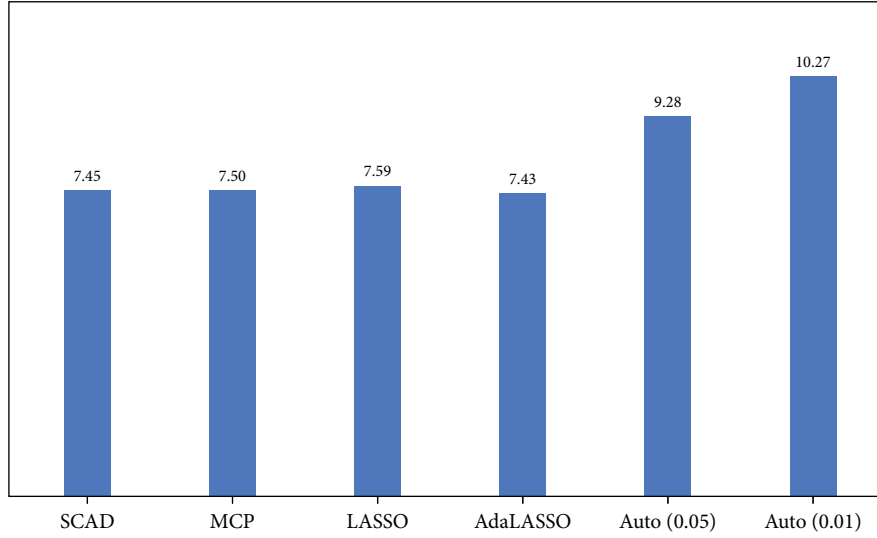


FIGURE 7: Out-of-sample RMSE of real data analysis.

Coronavirus pneumonia (COVID-19) is a worldwide health emergency because of its quick transmission and high death rate [39]. The clinical and physiological characteristics of SARS-CoV-2, as well as diagnostic approaches, have been studied all over the world [40]. During this pandemic, scientists and physicians face a global challenge in patient care and suitable treatment techniques, including creating an effective vaccine. Different diagnostic indicators have played a significant role in diagnosing and controlling the status of SARS-CoV-2 patients [41]. C-reactive protein (CRP) levels can be used as a biomarker to help diagnose pneumonia early, and individuals with severe lung infections have increased CRP levels [42]. Patients with COVID-19 have higher serum C-reactive protein (CRP) levels, which are used to help classify, diagnose, and make a prognosis of the disease [43]. This analysis is aimed at investigating the relationship between the length of hospital stay and CRP level, gender, age, diabetes, patient discharge status, and other comorbidities with permission of hospital authorities and consent of patient’s privacy. The data was gathered from Isolation Hospital and Infectious Treatment Center (IHITC) in Islamabad from July 2021 to 30 September 2021. A total of 275 patients agreed to join in the study between July and September. All the patients admitted they belonged to Rawalpindi and Islamabad regions. We extracted information for each individual, including age, gender, diabetic status, comorbidities, length of hospital stay, CRP level, and patient discharge status. Figure 4 illustrates the correlation graph of considered variables; this indicates the positive correlation between the hospital stay and CRP level with correlation equals 0.2 and negative correlation with other comorbidities with -0.1. However, patients’ survival and age are positively associated with hospital stay with a correlation equal 0.2 and 0.1, respectively. Figure 5 illustrates the box plot of the hospital stay. It indicates that the minimum length of hospital stay equals 1 and maximum 41, as the hospital stay is the dependent variable and contains an outlier, as shown in Figure 5. Furthermore, the residual plot of linear

regression presented in Figure 6 confirms outliers in model residuals. For the out-of-sample forecast, we randomly train the model on 90% of observations (233) and validate 10% of observations (26) [23, 44, 45].

After the confirmation of outlier in the data set, the estimated model with the IIS method is defined

$$\begin{aligned}
 \text{Hospital stay} = & \beta_0 + \beta_1 \text{gender} + \beta_2 \text{age} + \beta_3 \text{diabetes} + \beta_4 \text{CRP} \\
 & + \beta_5 \text{survival} + \beta_6 \text{other comorbidities} \\
 & + \sum_{i=1}^{233} \gamma_i I_i + \varepsilon_i.
 \end{aligned}
 \tag{13}$$

Table 4 indicates that SCAD and MCP perform similarly in covariate selection, as gender, CRP level, and other comorbidities are significant variables which increase the length of hospital stay. However, SCAD selected 28 outliers, and MCP selected 31 slightly higher than SCAD.

The real data analysis confirms that the LASSO estimates more covariates and outliers than other regularization techniques, aligned with our simulation findings. LASSO selects four more than covariates selected via SCAD and MCP. Autometrics with a 5% significance chooses two covariates and 14 outliers. AdaLASSO and Autometrics with a 1% significance do not select any covariate, only retain outliers. Overall, real data analysis indicates that gender, CRP level, and other comorbidities are significant covariates. These indicator dummies can be interpreted as an observed heterogeneity of individuals, which prolonged hospital stay length. We report the RMSE of regularization techniques in Figure 7.

The above figure indicates that SCAD and MCP outperform out-of-sample RMSE compared to all other considered techniques. As expected, the LASSO selected more indicator dummies and retained higher RMSE than SCAD and MCP. With 0.01 (level of significance), Autometrics holds the

highest RMSE compared to all other techniques because it dropped relevant covariate simulation finding aligned with existing studies of [20, 23]. Autometrics with tight significance levels omits relevant variables due to this RMSE increase (as observed from the simulation graph and table). In contrast, with a nominal significance level (0.05), Autometrics possesses higher RMSE than regularization techniques.

5. Conclusion

In cross-sectional data analysis, outlier occurred most frequently than the time series analysis, although outlier detection is a quick operation in healthcare and medical data, which is a significant cause of concern. Overall analysis indicates that regularization techniques perform more significantly than Autometrics in out-of-sample forecasting and covariate selection in simulation and real data analysis. However, the IIS method estimated via SCAD and MCP compromises promising covariate selection and forecasting results among regularization techniques. Regularization techniques with 20% and 4 SD outliers possess a higher average gauge than 20% and 6 SD. Conversely, 5% and 4 SD outlier's regularization technique possesses a higher average gauge than 5% and 4 SD outliers. Overall, with 4 SD outliers, the out-of-sample RMSE is optimal than 6 SD.

On the contrary, the LASSO estimates more outliers and covariates in simulation experiments and real data analysis than other regularization techniques. The real data analysis confirms the simulation findings, as the SCAD and MCP possess a minimum out-of-sample RMSE than Autometrics and LASSO. The real data analysis indicates that SCAD and MCP select three covariates, gender, CRP level, and other comorbidities, and possess the least RMSE. The real result is aligned with simulation findings as SCAD and MCP retain the highest potency and least RMSE compared to Autometrics. In contrast, LASSO possesses the highest gauge in simulation study compared to all considered techniques; the finding is aligned with real analysis as it retained the highest outliers. The concept of the IS method for outlier detection in the cross-sectional analysis would help to preserve unobserved heterogeneity in cross-sectional analysis, which simultaneously declines the RMSE of the estimated model. Our study proves that the IIS method for outlier detection and covariate selection estimated via SCAD and MCP gives more precise results than Autometrics in orthogonal covariates and outlier presences.

Data Availability

Data can be provided on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] A. Zaman, P. J. Rousseeuw, and M. Orhan, "Econometric applications of high-breakdown robust regression techniques," *Economics Letters*, vol. 71, no. 1, pp. 1–8, 2001.
- [2] I. H. Langford and T. Lewis, "Outliers in multilevel data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 161, no. 2, pp. 121–160, 1998.
- [3] Ö. G. Alma, "Comparison of robust regression methods in linear regression," *International Journal of Contemporary Mathematical Sciences*, vol. 6, pp. 409–421, 2011.
- [4] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, 2005.
- [5] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.
- [6] R. Berk, *A Primer on Robust Regression*, 1990, <https://escholarship.org/content/qt2cs5m2sh/qt2cs5m2sh.pdf>.
- [7] D. Birkes and Y. Dodge, *Alternative Methods of Regression*, John Wiley & Sons, 2011.
- [8] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Academic press, 2012.
- [9] J. A. Doornik, *Econometric Model Selection with More Variables than Observations*, Citeseer, 2009.
- [10] S. Johansen and B. Nielsen, "An analysis of the indicator saturation estimator as a robust regression estimator," *Castle, and Shephard (2009)*, vol. 1, pp. 1–36, 2009.
- [11] S. Johansen, D. F. Hendry, and C. Santos, "Selecting a Regression Saturated by Indicators," *CREATES Research Paper 2007-36*, 2007.
- [12] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Model selection when there are multiple breaks," *Journal of Econometrics*, vol. 169, no. 2, pp. 239–246, 2012.
- [13] J. L. Castle, J. A. Doornik, D. F. Hendry, and F. Pretis, "Detecting location shifts during model selection by step-indicator saturation," *Econometrics*, vol. 3, no. 2, pp. 240–264, 2015.
- [14] S. Muhammadullah, A. Urooj, and F. Khan, "A revisit of the unemployment rate, interest rate, GDP growth, and inflation rate of Pakistan: whether structural break or unit root?," *iRASD Journal of Economics*, vol. 3, no. 2, pp. 80–92, 2021.
- [15] F. Pretis, J. J. Reade, and G. Sucarrat, "Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks," *Journal of Statistical Software*, vol. 86, no. 3, 2018.
- [16] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [18] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.
- [19] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [20] C. Epprecht, D. Guégan, Á. Veiga, and J. Correa da Rosa, "Variable selection and forecasting via automated methods for linear models: LASSO/adaLASSO and Autometrics," *Communications in Statistics: Simulation and Computation*, vol. 50, no. 1, pp. 103–122, 2021.

- [21] F. Khan, A. Urooj, S. A. Khan, A. Alsubie, Z. Almaspoor, and S. Muhammadullah, "Comparing the forecast performance of advanced statistical and machine learning techniques using huge big data : evidence from Monte Carlo experiments," *Complexity*, vol. 2021, Article ID 6117513, 11 pages, 2021.
- [22] F. Khan, A. Urooj, S. A. Khan, S. K. Khosa, S. Muhammadullah, and Z. Almaspoor, "Evaluating the performance of feature selection methods using huge big data: a Monte Carlo simulation approach," *Mathematical Problems in Engineering*, vol. 2022, Article ID 6607330, 10 pages, 2022.
- [23] S. Muhammadullah, A. Urooj, F. Khan, M. N. Alshahrani, M. Alqawba, and S. Al-marzouki, "Comparison of weighted lag adaptive LASSO with Autometrics for covariate selection and forecasting using time-series data," *Complexity*, vol. 2022, Article ID 2649205, 10 pages, 2022.
- [24] M. Hauskrecht, I. Batal, C. Hong et al., "Outlier-based detection of unusual patient-management actions: an ICU study," *Journal of Biomedical Informatics*, vol. 64, pp. 211–221, 2016.
- [25] B. Wilson, C. L. Tseng, O. Soroka, L. M. Pogach, and D. C. Aron, "Identification of outliers and positive deviants for healthcare improvement: looking for high performers in hypoglycemia safety in patients with diabetes," *BMC Health Services Research*, vol. 17, no. 1, pp. 1–10, 2017.
- [26] P. N. Jyothi, D. R. Lakshmi, and K. V. S. N. Rama Rao, "A supervised approach for detection of outliers in healthcare claims data," *Journal of Engineering Science & Technology Review*, vol. 13, no. 1, pp. 204–214, 2020.
- [27] H. Noma, M. Goshio, R. Ishii, K. Oba, and T. A. Furukawa, "Outlier detection and influence diagnostics in network meta-analysis," *Research Synthesis Methods*, vol. 11, no. 6, pp. 891–902, 2020.
- [28] G. Jenkinson, Y. I. Li, S. Basu, M. A. Cousin, G. R. Oliver, and E. W. Klee, "LeafCutterMD: an algorithm for outlier splicing detection in rare diseases," *Bioinformatics*, vol. 36, no. 17, pp. 4609–4615, 2020.
- [29] R. Sakurai, M. Ueki, S. Makino et al., "Outlier detection for questionnaire data in biobanks," *International Journal of Epidemiology*, vol. 48, no. 4, pp. 1305–1315, 2019.
- [30] M. Verbanck, C. Chen, B. Neale, and R. Do, "Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases," *Nature Genetics*, vol. 50, no. 5, pp. 693–698, 2018.
- [31] K. D. Hoover and S. J. Perez, "Data mining reconsidered: encompassing and the general-to-specific approach to specification search," *The Econometrics Journal*, vol. 2, no. 2, pp. 167–191, 1999.
- [32] H. M. Krolzig and D. F. Hendry, "Computer automation of general-to-specific model selection procedures," *Journal of Economic Dynamics and Control*, vol. 25, no. 6-7, pp. 831–866, 2001.
- [33] D. F. Hendry and H. M. Krolzig, "We ran one regression," *Oxford Bulletin of Economics and Statistics*, vol. 66, no. 5, pp. 799–810, 2004.
- [34] F. Pretis, L. Schneider, J. E. Smerdon, and D. F. Hendry, "Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation," *Journal of Economic Surveys*, vol. 30, no. 3, pp. 403–429, 2016.
- [35] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, 2011.
- [36] E. Konzen and F. A. Ziegelmann, "LASSO-type penalties for covariate selection and forecasting in time series," *Journal of Forecasting*, vol. 35, no. 7, pp. 592–612, 2016.
- [37] J. A. Doornik and D. F. Hendry, "Statistical model selection with "big data"," *Cogent Economics & Finance*, vol. 3, no. 1, article 1045216, 2015.
- [38] S. Lee, "An additive sparse penalty for variable selection in high-dimensional linear regression model," *Communications for Statistical Applications and Methods*, vol. 22, no. 2, pp. 147–157, 2015.
- [39] P. Chatterjee, N. Nagi, A. Agarwal et al., "The 2019 novel coronavirus disease (COVID-19) pandemic: a review of the current evidence," *The Indian Journal of Medical Research*, vol. 151, no. 2, pp. 147–159, 2020.
- [40] R. M. Elshazli, E. A. Toraih, A. Elgaml et al., "Diagnostic and prognostic value of hematological and immunological markers in COVID-19 infection: a meta-analysis of 6320 patients," *PLoS One*, vol. 15, no. 8, article e0238160, 2020.
- [41] Y. Li, W. Zhou, L. Yang, and R. You, "Physiological and pathological regulation of ACE2, the SARS-CoV-2 receptor," *Pharmacological Research*, vol. 157, article 104833, 2020.
- [42] D. Stringer, P. Braude, P. K. Myint et al., "The role of C-reactive protein as a prognostic marker in COVID-19," *International Journal of Epidemiology*, vol. 50, no. 2, pp. 420–429, 2021.
- [43] N. Chen, M. Zhou, X. Dong et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," *Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.
- [44] J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *Mathematical Intelligence*, vol. 27, no. 2, pp. 83–85, 2005.
- [45] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.

Research Article

Complex Survival System Modeling for Risk Assessment of Infant Mortality Using a Parametric Approach

Hang Chen,¹ Maryam Sadiq ,² and Zishen Song ¹

¹Department of Electronics and Information, Xi'an Jiaotong University, Shaanxi 710049, China

²Department of Statistics, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan

Correspondence should be addressed to Zishen Song; zishens@sina.com

Received 6 December 2021; Revised 6 March 2022; Accepted 31 March 2022; Published 19 April 2022

Academic Editor: Diego Pinto

Copyright © 2022 Hang Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pakistan is still one of the five countries contributing to half of the child deaths worldwide and holds a low ratio of infant survival. A high rate of poverty, low level of education, limited health facilities, rural-urban inequalities, and political uncertainty are the main reasons for this condition. Survival models that evaluate the performance of models over simulated and real data set may serve as an effective technique to determine accurate complex systems. The present study proposed an efficient extension of the recent parametric technique for risk assessment of infant mortality to address complex survival systems in the presence of extreme observations. This extended method integrated four distributions with the basic algorithm using a real data set of infant survival without extreme observations. The proposed models are compared with the standard partial least squares-Cox regression (PLS-CoxR), and higher efficiency of these proposed algorithms is observed for handling complex survival time systems for risk assessment. The algorithm is also used to analyze simulated data set for further verification of results. The optimal model revealed that the mother's age, type of residence, wealth index, permission to go to a medical facility, distance to a health facility, and awareness about tuberculosis significantly affected the survival time of infants. The flexibility and continuity of extended parametric methods support the implementation of public health surveillance data effectively for data-oriented evaluation. The findings may support projecting targeted interventions, producing awareness, and implementing policies planned to reduce infant mortality.

1. Introduction

Strong statistical survival techniques are the demand of the era for authentic and reliable results for deeply examining complex survival and mortality patterns. Nonparametric survival techniques including the Kaplan-Meier product-limit method [1], the Gehan's generalized Wilcoxon test [2], and the log-rank test [3] were extensively used in older times. The Cox's regression model remained the most popular and widely used semiparametric survival technique if the proportional hazards assumption is fulfilled [4]. In recent times, flexible parametric models (FPM) are considered as a better alternative to nonparametric and semiparametric methods as they produce estimates with higher efficiency and lower standard errors [5]. In addition, these models consider full likelihood to draw more precise inferences

and easily interpretable results. So far, the FPM has been employed various probability distributions to estimate survival functions. The exponential probability distribution supports as the baseline to handle survival time. The Weibull, Gompertz, generalized gamma, and generalized F-distribution are commonly practiced too. The FPM is also able to efficiently investigate the relationship of covariates with survival response [5]. The partial least squares-Cox regression (PLS-CoxR) integrates PLS with the Cox model to address survival time response with collinear covariates [6] since the Cox regression is restricted with inflexible estimates of the cumulative hazard and survival functions as being incomplete. Hence, the PLS-CoxR model is restricted in the long-term estimation with unsmooth functions.

The flexible parametric models (FPMs) are recommended to compute hazard and cumulative hazard

functions for covariates to extrapolate the survival model. The FPM can estimate continuous survival and hazard functions instead of a step representation due to its flexibility [7].

Despite considerable improvement towards increasing infant survival, nearly six million child deaths are recorded every year, before attaining their fifth birthday [8]. By the end of 2015, a minor proportion of developing countries have met the fourth target of Millennium Development Goal (MDG) which is intended to increase the child survival rate by two-thirds [9]. The recently described Sustainable Development Goals (SDG) seek to forward the objectives originated by the MDG. The third SDG is to reduce the under-five mortality rate (U5MR) to 25 deaths per 1000 live births by 2030 [10]. Previous literature evidenced that five countries including China, Congo, Nigeria, India, and Pakistan possess nearly half of under-five mortality in the world [11]. Pakistan has the sixth largest population in the world with 188 million people [12]. In 2018, Pakistan's infant mortality rate (IMR) was 61 deaths per 1000 live births. Due to political instability, civil conflicts, poverty, lower educational level, unavailability of health facilities, and disparities regarding the area in Pakistan, 70% MDG targets were not achieved [13]. Understanding the factors affecting infant mortality is significantly informative to health professionals, practitioners, and health policymakers for the improvement of population health status through effective interventions.

Within this line, the partial least squares flexible parametric model (PLS-FPM) is developed to analyze the complex survival systems in the presence of extreme observations for risk and hazard assessment [14]. The present study extended the PLS-FPM to collinear predictors having moderate trend observations using four alternative probability distributions.

The results exposed the flexible dynamics of the extended method to obtain smooth survival and hazards estimates in the presence of multicollinearity. This model can be implemented in the field of genetics, biology, engineering, medicine, social sciences, or behavioral sciences for system reliability and risk assessment. The formal statements of the problem are the following:

- (i) Selection of optimum model by execution of four distribution integrated with the PLS-FPM oversimulated and real data set having collinear predictors and moderate observation
- (ii) Identification of significant risk factors of infant mortality in Pakistani

2. Methodology

The PLS-CoxR model is considered as the benchmark method in the present study, and the PLS-FP model with four different distributions is the proposed technique.

2.1. The Cox Regression Model. The Cox model has the form

$$\lambda(t) = \lambda_o(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p) = \lambda_o(t) \exp[\beta' X], \quad (1)$$

where $\lambda_o(t)$ represents the baseline hazard function, β is the vector of regression estimates, and X denotes a $(n * p)$ matrix of predictors.

2.2. The Partial Least Squares-Cox Regression Model. The PLS-CoxR model is employed as the reference method in the present study. Suppose the survival time is represented by t and $x_j = x_{1j}, x_{2j}, \dots, x_{nj}$ be the vector of p correlated covariates with n samples. The model estimates k components for p correlated predictors and assumes the hazard estimate as

$$\lambda(t) = \lambda_o(t) \exp(\beta_1 S_1 + \beta_2 S_2 + \dots + \beta_p S_k) = \lambda_o(t) \exp[\beta' S], \quad (2)$$

where S represents a $(n * k)$ matrix of components.

2.3. Flexible Parametric Survival Model (FPSM). Let T represent a nonnegative continuous survival response and let X is the vector of predictors x_1, \dots, x_p over a sample of size n . The survival function is the probability of being alive at time t and is represented by $S(t) = \Pr(T > t)$ for a vector of covariates at time t with the cumulative distribution function $F(t) = \Pr(T \leq t)$. Then the cumulative hazard or risk function is

$$\Lambda(t) = \int_0^t \lambda(x) dx. \quad (3)$$

Any distribution ranges over $t \in [0, \infty]$, and it may serve as survival distribution. The survival distributions included in this study as FPSM are as follows:

2.3.1. The Gompertz Distribution. A survival response T following a Gompertz distribution with parameters $(b > 0, \eta > 0)$ exhibits the survival function

$$S(t) = \exp\left(-\frac{b}{\eta}(e^{\eta t} - 1)\right), \quad (4)$$

and the cumulative hazard function as

$$\Lambda(t) = \frac{b}{\eta}(e^{\eta t} - 1). \quad (5)$$

The Gompertz distribution is also an extreme value distribution with increasing hazard function.

2.3.2. The Generalized Gamma Distribution. The generalized gamma distribution with parameters (β, σ, κ) has survival function as

$$S(t) = 1 - \Gamma \kappa^{-2} \left(e^{-\beta t} t^{\kappa/\sigma}; \kappa^{-2} \right). \quad (6)$$

The hazard function of the generalized gamma function is increasing, decreasing, bathtub, and arc-shaped [15].

2.3.3. *The Generalized F-Distribution.* The density function of generalized F-distribution with $2\nu_1$ and $2\nu_2$ is

$$f(t) = (\nu_1 e^t / \nu_2)^{\nu_1} (1 + \nu_1 e^t / \nu_2)^{-(\nu_1 + \nu_2)} \beta(\nu_1, \nu_2)^{-1}, \quad (7)$$

where $\beta(\nu_1, \nu_2)$ is the beta function and then the survival function is

$$S(t) = \int_0^{\nu_2(\nu_2 + \nu_1 e^t)^{-1}} \chi^{\nu_2 - 1} (1 - \chi)^{\nu_1 - 1} \beta(\nu_2, \nu_1)^{-1} dx, \quad (8)$$

where χ denotes the chi-square distribution. This distribution is useful for testing different parametric forms as it includes other distributions as limiting or special cases.

2.3.4. *The Exponential Distribution.* The survival time T has an exponential distribution with rate parameter λ having density function

$$f(t) = \lambda \exp(-\lambda t), \quad (9)$$

then the survival function is

$$S(t) = \exp(-\lambda t), \quad (10)$$

and the cumulative hazard function is

$$\Lambda(t) = \lambda t. \quad (11)$$

Several other probability distributions can be employed in FPM. The interpretation for regression coefficients of FPM is the same as for semiparametric models. The FPM provides a more stabilized cumulative hazard function than the semiparametric model. For instance, the Weibull models produce the hazard function as a continuous straight trend. The PLSR model integrated with FPM addressing generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution is included in the present study for improved model performance for multicollinear covariates.

2.4. *The Partial Least Squares Flexible Parametric (FP) Model.* The proposed model assumes the occurrence of an event e at time t in the presence of censoring, and let X be the matrix of p correlated predictors x_1, \dots, x_p for a sample of size n . The method computes the FP model for S components (as $S \leq p$) computed from PLSR for survival response and X as a matrix of predictors. The PLS-FP model assumes that some A is equal to the number of components to be predicted (where $A \leq p$), then for $a = 1, 2, \dots, A$, the algorithm runs:

(1) Loading weights are computed by

$$w_a = X'_{a-1} t_{a-1}. \quad (12)$$

Loading weights are normalized to have length equal to 1 by

$$w_a \leftarrow \frac{X'_{a-1} t_{a-1}}{\|X'_{a-1} t_{a-1}\|}. \quad (13)$$

(2) Score vector s_a is computed by

$$s_a = X_{a-1} w_a. \quad (14)$$

The risk function for FPSM is computed as

$$\Lambda(t) = \int_0^t \lambda(s) ds. \quad (15)$$

(3) If $a < A$ return to 1

The PLS-FP model is a two-stage procedure. At the first stage, the PLS-FP regression model computes components of PLS regression with time as response outcome and correlated covariates as predictors. Then, it executes the FP model with survival time as response and components of PLSR as explanatory factors at the later stage. This method produces efficient estimates with increased accuracy for collinear predictors. Hence, it is recommended to use in the case of collinear data as it is a conjugate of PLS and FP models. The PLSR model is also coupled with a filter-based factor selection method, namely, "loading weights" to identify the significant factors [16, 17].

2.5. *Simulated Survival Data Generated from Gompertz Distributions.* The R-package namely "simsurv" is used for the generation of simulated survival data [18] with moderate observation and collinear predictors. The data follows Gompertz distribution with 0.1 and 0.1 scale and shape parameters, respectively. The correlation among predictors is established as (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0) for 100 samples with 30 predictors.

2.6. *Infant Survival Times Data.* This study used secondary data, obtained from the Demographic and Health Surveys (DHS), gathered during 2012-2013 from Pakistan. Hence, no ethical concerns are required to conduct this study [19]. The present analysis used data set of infants aged 1-12 months in Pakistan. Due to missing and incomplete information, infants dead within one month of birth are excluded from the analysis. A total of 697 infants belonging to Pakistan and 83 predictor variables are included.

3. Results

The PLS-FPM parameterized with generalized gamma, generalized F, exponential, and Gompertz distribution are modeled on simulated data generated from Gompertz distribution to observe the variation in efficiency for multicollinear data. The left panel of Figure 1 showed the efficiency of models established by AIC and indicated that coupled with PLSR, the FPM models showed the higher efficiency over simulated

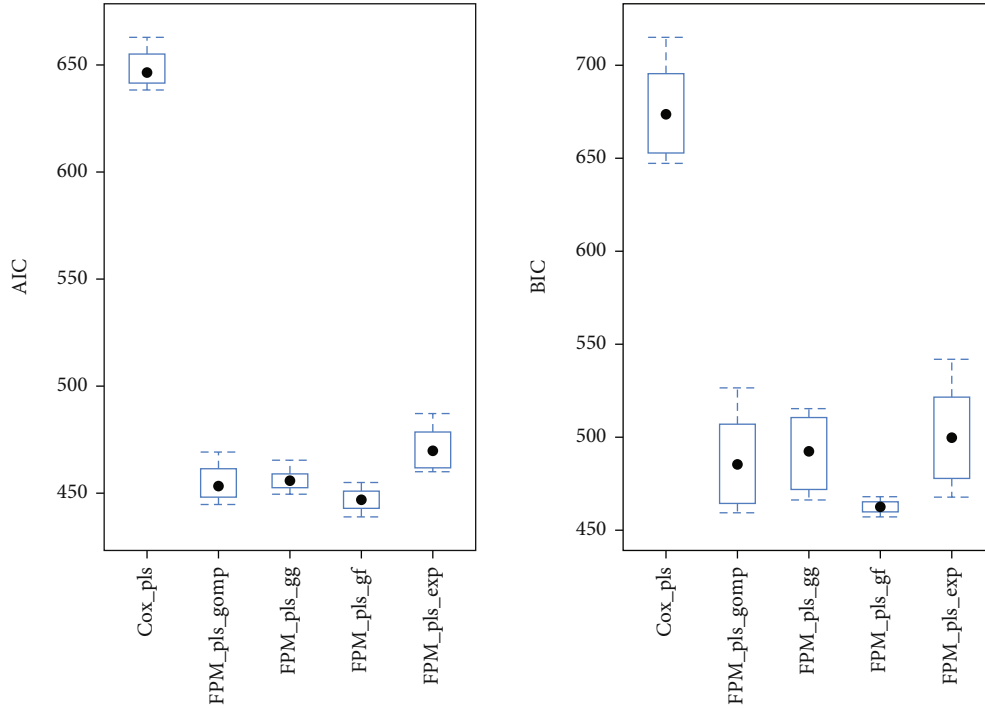


FIGURE 1: The comparison of the PLS-Cox model with the PLS-FPM parameterized over generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution for simulating survival response generated from Gompertz distribution based on AIC and BIC.

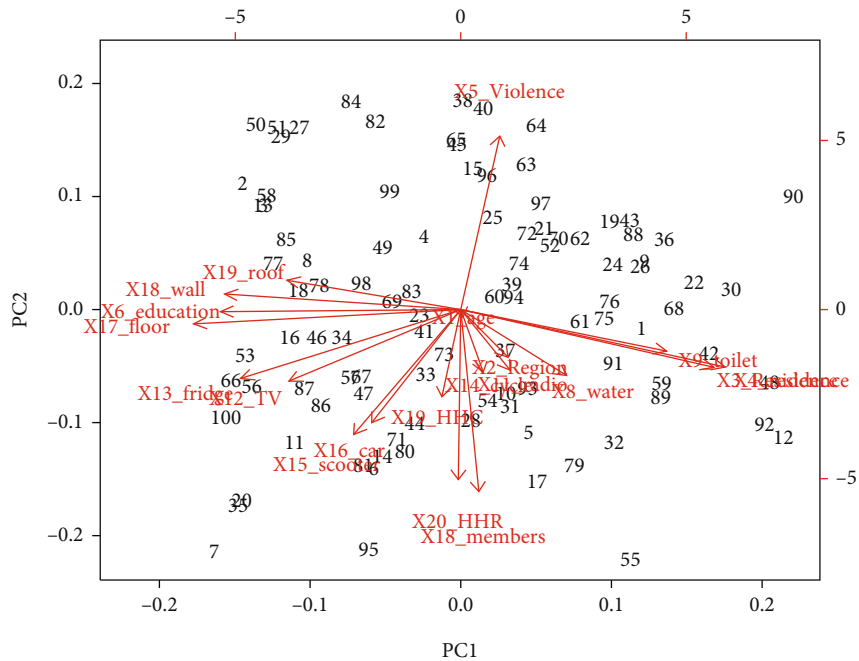


FIGURE 2: A biplot visualizing the correlations between the covariates on the first two principal components for infant survival data set.

data having known correlation structure. Similar results based on BIC, as shown in Figure 1(b), are observed. The simulation analysis demonstrated that the proposed models are efficient and reliable in terms of performance for the corresponding distributions. The analysis over simulation recommended the practical application of proposed models to examine sur-

vival response along with correlated covariates in a more flexible manner.

Before analyzing the real data set, multicollinearity among covariates is verified to justify the application of PLS. For this purpose, correlations structure for infant survival data is examined. The biplot for infant survival data

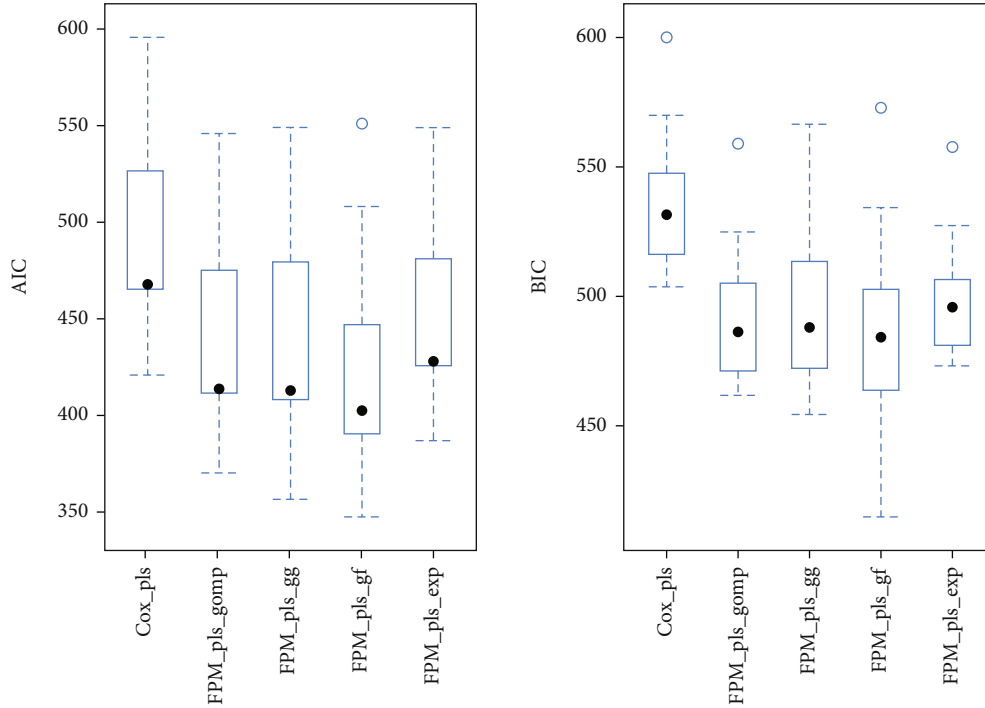


FIGURE 3: The comparison of reference model with the PLS-FPM parameterized over generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution on the basis of AIC and BIC for infant survival are presented.

presented in Figure 2 clearly portrayed the correlation between covariates showing close points of occurrence.

Real data set of infant survival with 12 months of censoring is considered in this analysis. Discarding outliers, 83 covariates measured over 577 observations (infants) were included in the final sample to compare survival models. The data set is randomly split into testing (30%) and training data (70%) for reliable results. After verification of multicollinearity among covariates, the PLS-FPM parameterized over Gompertz, generalized gamma, generalized F, and exponential distribution are analyzed. The PLS-Cox model for survival time is considered as the reference method. Figure 3 showed the efficiency of models measured by AIC and BIC which demonstrated the higher performance of modified models compared to the PLS-Cox over infant survival data. The proposed models based on the parametric approach performed better due to their additional flexibility. Flexible parametric models integrated with PLSR parameterized with generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution showed increased accuracy compared to the Cox model integrated with PLS.

The Gompertz distribution is modeled into the innovation-imitation paradigm, and its hazard function works as a convex function. These properties develop their flexibility to use as flexible parametric distribution in survival models. Hence, it increased the performance of the model incorporated with PLS compared to the semiparametric model, due to its flexible nature. Based on AIC and BIC, it is concluded that the PLS-FPM parameterized over generalized F (GF) is the best-fitted model and hence further executed for influential factor selection. PLS-FP model based on generalized F-distribution with location parameter μ is found to be the most

TABLE 1: A description of corresponding parameter of each distribution used in the PLS-FPM for infant survival data set is presented.

Model	Parameter			
	Location	Scale	Shape	Rate
<i>FPM_pls_gomp</i>	—	—	0.231	-7.20
<i>FPM_pls_gg</i>	4.32	-0.32	0.56	—
<i>FPM_pls_gf</i>	4.37	-1.44	-0.11, 2.96	—
<i>FPM_pls_g</i>	—	—	0.77	-3.56

efficient model over infant survival times data. In this model, covariates on the corresponding parameter represent the accelerated failure time (AFT) model which speeds up or slows down the passage of time. A detailed illustration of PLS-FP model parameterization is presented in Table 1 to describe the corresponding location, scale, shape, and rate parameter of the associated distribution.

Figure 4 showed the cumulative hazards regression estimates for the reference method and the PLS-FPM integrated with generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution for infant mortality data. The proposed PLS-FPM delivered smooth regression coefficients of the hazard functions extrapolated to a time of 12 months showing consistent estimates. The reference model showed unsmooth hazard trends with odd fluctuations for certain time intervals shown in Figure 4.

For modeling the survival time data, the PLS-FPM parameterized over generalized F (GF) is applied, and a well-known factor selection method of PLS, namely, loading weights, is used to estimate the regression coefficients of

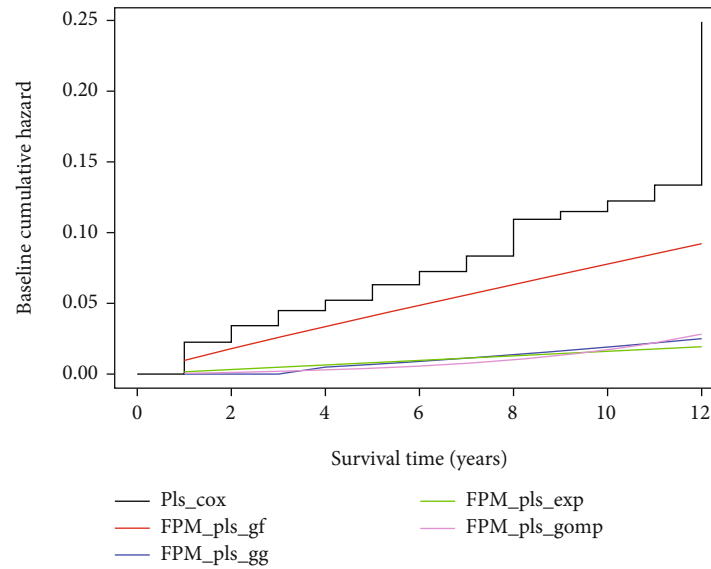


FIGURE 4: The cumulative hazard estimates of the PLS-Cox and the PLS-FPM parameterized over generalized gamma (GG), generalized F (GF), exponential, and Gompertz distribution for infant survival data.

TABLE 2: The coefficient estimates of influential factors for infant survival obtained by the PLS-FPM coupled with GF-distribution.

Factor	Coefficient
Mother's age	-0.2599634
Province	0.1152137
Type of place of residence	-0.2259600
Selected for domestic violence module	-0.1314814
Mother's educational level	0.1150733
Type of toilet facility	0.1306607
Household has: television	0.1123271
Main roof material	-0.1824950
Relationship to household head	-0.1462302
Sex of household head	0.1263040
Toilet facilities shared with other households	0.1128781
Wealth index	-0.2221866
Total children ever born	0.1440127
Sons died	0.1444297
Daughters died	0.1380171
Used contraceptive methods	0.1255910
Have mosquito bed net for sleeping	-0.1472814
Getting medical help for self: getting permission to go	0.2765390
Getting medical help for self: getting money needed for treatment	0.1091858
Getting medical help for self: distance to health facility	0.2711795
Getting medical help for self: having to take transport	0.1405939
Getting medical help for self: not wanting to go alone	0.1932727
Heard of tuberculosis or TB	-0.2165038
Person who usually decides on visits to family or relatives	-0.1273039
Preceding birth interval (in months)	-0.1647803
Duration of breastfeeding	-0.1697750
Blood relation with husband	-0.1285042
Total pregnancy outcomes	-0.1926369

significant factors. The estimates of important predictors associated with infant mortality are presented in Table 2.

After analysis, 28 influential factors out of 80 which significantly affect infant survival in Pakistan are observed. A negative relationship of mother's age, region, selection for domestic violence, main roof material, relationship to household head, wealth index, availability of mosquito bed net, awareness about tuberculosis (TB), decision power to visit family, preceding birth interval, duration of breastfeeding, blood relation with husband, and total pregnancy outcomes are found for infant survival. Furthermore, positive association of province, mother's education, toilet facility, availability of television, sex of household head, shared toilet, number of total children, number of dead son and daughters, use of contraception, availability of permission, money, transport, and attendant for medical facility and distance to a medical facility was observed.

4. Discussion

Estimating the hazard and survival functions that flexibly explain complex systems remained a hard and computationally challenging task. Hence, the candidate models are usually limited in studies to allow for evaluations and comparisons. However, nonparametric and semiparametric survival methods can speculate model structures as unsmooth estimates are evaluated. The present study extended the PLS-FPM [14] to correlated predictors having moderate trend observations using four alternative probability distributions. The PLS-FPM extends previous survival approaches that either perform semiparametric analyses or use nonparametric methods, while analysis of all previous methods was limited due to their inflexible nature. To administrate all four shaped hazard functions, distribution fitting is implemented over defined simulated survival data set.

Most previous literature used the Cox regression model for infant survival analysis [20]. Very few recent studies used FPM to examine infant survival analysis [21]. The PLS-FPM is compared with the reference method for both simulated and a real data set for collinear covariates. A previous study proposed the PLS-FPM integrated with Gamma, Weibull, log-logistic, and log-normal distributions for data with extreme observations to examine four real data sets of breast cancer survival time and identify the significantly associated gene signatures for each data set. The study found that the PLS-FPM has higher performance than the traditional PLS-Cox model [14]. Consistent with the previous study, the present study found the higher efficiency of the PLS-FPM compared to the PLS-Cox regression method for data sets with moderate observations. The PLS-FPM coupled with Gompertz distribution is found to be the optimum model to estimate hazard functions using AIC for simulated survival data following Gompertz distribution. The efficiency of the algorithms flexibly increases the model accuracy to a greater extent even considering correlated predictors. This accuracy suggested that hazard, as well as survival functions, can be accurately computed by smooth trends for the survival response. A recent study proposed the partial least squares spline modeling approach by inte-

grating PLS with restricted cubic spline model and compared it with the PLS-Cox model [22]. The study estimated the risk factors of infant mortality in Pakistan by using the PLS-spline model based on the odds scale with one knot. This study also examine the important factors of infant mortality by executing the optimal model, namely, the PLS-FPM parameterized over generalized F (GF), and identified the influential factors which are also determined by various previous studies. Consistent with the recent literature, the present study evidenced that mother's age, region, selection for domestic violence, relationship to household head, wealth index, awareness about tuberculosis (TB), decision power to visit family, preceding birth interval, and blood relation with husband [22] are significantly associated with infant mortality. Some other previous literature also supported the association of main roof material [23, 24], availability of mosquito bed net [25], duration of breastfeeding [26], and total pregnancy outcomes [27] with infant survival similar to the present study.

Various previous studies also observed the positive association of province [28], mother's educational level [29], type of toilet facility [30], availability of television [31], sex of household head [32], shared toilet [33], number of total children [34], number of died son and daughters [35], use of contraception [36], and availability of permission, money, transport, distance and attendant for medical facility [37, 38] with infant survival which is consistent with the current study. Last but not least, the PLS-FPM not only can extrapolate survival response besides the availability of follow-up information but also sponsors variant hazard shapes. The PLS-FPM is suggested as a helpful parametric addition for the estimation and prediction of survival response. This model is recommended to use in reliability theory for risk assessment.

Data Availability

Data is freely available at <https://dhsprogram.com/>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] E. T. Lee and O. T. Go, "Survival analysis in public health research," *Annual review of public health.*, vol. 18, no. 1, pp. 105–134, 1997.
- [2] E. A. Gehan, "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, no. 1-2, pp. 203–224, 1965.
- [3] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 2, pp. 185–198, 1972.
- [4] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [5] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Science & Business Media, 2006.

- [6] P. Bastien, V. E. Vinzi, and M. Tenenhaus, "PLS generalised linear regression," *Computational Statistics & data analysis.*, vol. 48, no. 1, pp. 17–46, 2005.
- [7] P. Royston and P. C. Lambert, *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*, vol. 347, Stata press College Station, TX, 2011.
- [8] J. Bryce, R. E. Black, and C. G. Victora, "Millennium development goals 4 and 5: progress and challenges," *BMC medicine.*, vol. 11, no. 1, pp. 1–4, 2013.
- [9] R. E. Black, C. Levin, N. Walker et al., "Reproductive, maternal, newborn, and child health: key messages from _Disease Control Priorities 3rd Edition_," *The Lancet.*, vol. 388, no. 10061, pp. 2811–2824, 2016.
- [10] W. A. Rosa, *New Era in Global Health: Nursing and the United Nations 2030 Agenda for Sustainable Development*, Springer Publishing Company, 2017.
- [11] A. Helova, K. R. Hearld, and H. Budhwani, "Associates of neonatal, infant and child mortality in the Islamic Republic of Pakistan: a multilevel analysis using the 2012–2013 demographic and health surveys," *Maternal and child health journal.*, vol. 21, no. 2, pp. 367–375, 2017.
- [12] W. H. Organization, *World Health Statistics 2016: Monitoring Health for the SDGs Sustainable Development Goals*, World Health Organization, 2016.
- [13] National Institute of Population Studies (Pakistan), Macro International, Institute for Resource Development, and Demographic, & Health Surveys, *Pakistan Demographic and Health Survey*, National Institute of Population Studies, 2012.
- [14] M. Sadiq and T. Mehmood, "A flexible and robust approach to analyze survival systems in the presence of extreme observations," *Mathematical Problems in Engineering.*, vol. 2021, pp. 1–11, 2021.
- [15] C. Cox and M. Matheson, "A comparison of the generalized gamma and exponentiated Weibull distributions," *Statistics in medicine.*, vol. 33, no. 21, pp. 3772–3780, 2014.
- [16] M. Sadiq, T. Mehmood, and M. Aslam, "Identifying the factors associated with cesarean section modeled with categorical correlation coefficients in partial least squares," *PLoS One*, vol. 14, no. 7, article e0219427, 2019.
- [17] T. Mehmood, M. Sadiq, and M. Aslam, "Filter-based factor selection methods in partial least squares regression," *IEEE Access.*, vol. 7, pp. 153499–153508, 2019.
- [18] S. L. Brilleman, R. Wolfe, M. Moreno-Betancur, and M. J. Crowther, "Simulating survival data using the simsurvR package," *Journal of Statistical Software.*, vol. 97, no. 3, pp. 1–27, 2021.
- [19] P. Demographic, "Health Survey 2012–13," in *Islamabad and Calverton, MA: National Institute of Population Studies and ICF International; 2013, 2015*, <https://dhsprogram.com/data>.
- [20] A. K. Iddrisu, A. Alhassan, and N. Amidu, "Survival analysis of birth defect infants and children with pneumonia mortality in Ghana," *Public Health*, vol. 2019, pp. 1–7, 2019.
- [21] R. K. Saroj, K. H. H. V. S. S. N. Murthy, M. Kumar, R. Singh, and A. Kumar, "Survival parametric models to estimate the factors of under-five child mortality," *Journal of Health Research and Reviews*, vol. 6, no. 2, 2019.
- [22] M. Sadiq, D. K. F. Alnagar, A. T. Abdulrahman, and R. Alharbi, "The partial least squares spline model for public health surveillance data," *Computational and Mathematical Methods in Medicine.*, vol. 2022, pp. 1–7, 2022.
- [23] S. A. Adebowale, O. M. Morakinyo, and G. R. Ana, "Housing materials as predictors of under-five mortality in Nigeria: evidence from 2013 demographic and health survey," *BMC pediatrics.*, vol. 17, no. 1, p. 30, 2017.
- [24] E. Bendavid, "Changes in child mortality over time across the wealth gradient in less-developed countries," *Pediatrics*, vol. 134, no. 6, pp. e1551–e1559, 2014.
- [25] D. Meekers and J. O. Yukich, "The association between household bed net ownership and all-cause child mortality in Madagascar," *Malaria journal.*, vol. 15, no. 1, p. 475, 2016.
- [26] D. Phukan, M. Ranjan, and L. Dwivedi, "Impact of timing of breastfeeding initiation on neonatal mortality in India," *International breastfeeding journal*, vol. 13, no. 1, 2018.
- [27] S. W. Masho and P. W. Archer, "Does maternal birth outcome differentially influence the occurrence of infant death among African Americans and European Americans?," *Maternal and child health journal.*, vol. 15, no. 8, pp. 1249–1256, 2011.
- [28] S. A. Collins, P. Surmala, G. Osborne et al., "Causes and risk factors for infant mortality in Nunavut, Canada 1999–2011," *BMC pediatrics*, vol. 12, no. 1, 2012.
- [29] G. T. Kiross, C. Chojenta, D. Barker, T. Y. Tiruye, and D. Loxton, "The effect of maternal education on infant mortality in Ethiopia: a systematic review and meta-analysis," *PLoS One*, vol. 14, no. 7, p. e0220076, 2019.
- [30] O. Ezeh, K. Agho, M. Dibley, J. Hall, and A. Page, "The impact of water and sanitation on childhood mortality in Nigeria: evidence from demographic and health surveys, 2003–2013," *International journal of environmental research and public health.*, vol. 11, no. 9, pp. 9256–9272, 2014.
- [31] T. Dendup, Y. Zhao, and D. Dema, "Factors associated with under-five mortality in Bhutan: an analysis of the Bhutan National Health Survey 2012," *BMC Public Health*, vol. 18, no. 1, p. 1375, 2018.
- [32] R. Adhikari and C. Podhisita, "Household headship and child death: evidence from Nepal," *BMC international health and human rights*, vol. 10, no. 1, 2010.
- [33] J. Golding, R. Greenwood, A. McCaw-Binns, and P. Thomas, "Associations between social and environmental factors and perinatal mortality in Jamaica," *Paediatric and perinatal epidemiology.*, vol. 8, no. s1, pp. 17–39, 1994.
- [34] A. van Soest and U. R. Saha, "Relationships between infant mortality, birth spacing and fertility in Matlab, Bangladesh," *PLoS One*, vol. 13, no. 4, article e0195940, 2018.
- [35] B. B. Gubhaju, "The effect of previous child death on infant and child mortality in rural Nepal," *Studies in Family Planning.*, vol. 16, no. 4, pp. 231–236, 1985.
- [36] U. R. Saha and A. van Soest, "Contraceptive use, birth spacing, and child survival in Matlab, Bangladesh," *Bangladesh. Studies in family planning*, vol. 44, no. 1, pp. 45–66, 2013.
- [37] S. Adebowale and E. Udjo, "Maternal health care services access index and infant survival in Nigeria," *Ethiopian journal of health sciences.*, vol. 26, no. 2, pp. 131–146, 2016.
- [38] D. Kadobera, B. Sartorius, H. Masanja, A. Mathew, and P. Waiswa, "The effect of distance to formal health facility on childhood mortality in rural Tanzania, 2005–2007," *Global health action.*, vol. 5, no. 1, p. 19099, 2012.

Research Article

The Partial Least Squares Spline Model for Public Health Surveillance Data

Maryam Sadiq ¹, Dalia Kamal Fathi Alnagar ^{2,3}, Alanazi Talal Abdulrahman,⁴
and Randa Alharbi²

¹Department of Statistics, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan

²Department of Statistics, University of Tabuk, Saudi Arabia

³Department of Statistics, Omdurman Islamic University, Sudan

⁴Department of Mathematics, College of Science, University of Ha'il, Saudi Arabia

Correspondence should be addressed to Maryam Sadiq; hussainulamad@gmail.com

Received 6 December 2021; Revised 24 December 2021; Accepted 31 December 2021; Published 27 January 2022

Academic Editor: David Becerra-Alonso

Copyright © 2022 Maryam Sadiq et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Factor discovery of public health surveillance data is a crucial problem and extremely challenging from a scientific viewpoint with enormous applications in research studies. In this study, the main focus is to introduce the improved survival regression technique in the presence of multicollinearity, and hence, the partial least squares spline modeling approach is proposed. The proposed method is compared with the benchmark partial least squares Cox regression model in terms of accuracy based on the Akaike information criterion. Further, the optimal model is practiced on a real data set of infant mortality obtained from the Pakistan Demographic and Health Survey. This model is implemented to assess the significant risk factors of infant mortality. The recommended features contain key information about infant survival and could be useful in public health surveillance-related research.

1. Introduction

Survival approach is a common regression modeling method used for prognostic analysis as it examines the relationship between the covariates, the response, and the time until the occurrence of an event. The framework for survival analysis is based on the Cox proportional hazard (PH) model due to its ease of computing the hazard ratio (HR) without needing to estimate the baseline hazard function. The Cox PH model maximizes the partial likelihood function which estimates the regression parameters but not the baseline hazard function. Consequently, the survival probability and the hazard rates can be estimated only at the event times and not for the long-term evaluations [1].

Parametric survival models specify the probability distribution to estimate the absolute measure of effect in time to event response. A common specification is the Weibull dis-

tribution in these models to estimate the baseline hazard $h_o(t)$. A parametric survival model with a scale parameter ($\lambda > 0$), a shape parameter ($\gamma > 0$), and time (t) is defined as $h_o(t) = \lambda\gamma t^{\gamma-1}$. For the absolute measure of effect, the Weibull distribution can generally facilitate accurate predictions for a constant, monotonically decreasing or monotonically increasing hazards. However, for more complex hazard functions, the parametric survival model specifying a Weibull function will lead to inaccurate predictions [2].

The Royston and Parmar model is an advanced type of flexible parametric survival model featuring a restricted cubic spline to model more complex hazard shapes and to estimate a continuous function [3]. This model considers the baseline log cumulative hazard function on the log timescale. For Weibull distribution, this function is $\ln(H(t) | z_i) = \ln(\lambda) + \gamma \ln(t) + \beta z_i$ where $\ln(\lambda)$ and $\gamma \ln(t)$ represent the baseline hazard with respect to log time and βz_i denotes the

```

1: function PLS model  $\mathbb{X}, t, e, a$  where  $\mathbb{X}$  is the covariate matrix,  $t$  is the time,  $e$  is the event, and  $c$  is the number of components.
2:    $w_{(c)} = \mathbb{X}_{(c-1)}^t \mathbb{T}_{(c-1)}$  loading weights
3:    $w_{(c)} \leftarrow w_{(c)} / \|w_{(c)}\|$  normalized loading weights
4:    $s_{(c)} = \mathbb{X}_{(c-1)} w_{(c)}$  score vector
5:    $p_{(c)} = \mathbb{X}_{(c-1)}^t (s_{(c)} / s_{(c)}^t s_{(c)}) \mathbf{X}$ -loadings
6:    $q_c = \mathbb{T}_{(c-1)}^t (s_{(c)} / s_{(c)}^t s_{(c)}) t$ -loadings            $\triangleright$  repeat the above steps until  $c < C$ 
7:   forc = 1 to Cdo
8:      $\{\text{RP}\{\text{Surv}(t, e)\}\}^c \sim \sum_{c=1}^C s^c$             $\triangleright$  Royston and Parmar (RP) restricted cubic spline model on PLSR components.

```

ALGORITHM 1: Partial least squares spline (PLS-spline) model.

vector of predictors. This function can be generalized as $\ln(H(t) | z_i) = \ln[H_o(t)] + \beta z_i$ where $\ln[H_o(t)]$ describes a general baseline log cumulative hazard function. Royston and Parmar used a restricted cubic spline to model the baseline hazard function on the log timescale. A restricted or natural cubic spline has an additional restriction featuring the first and last subfunctions beyond the boundary knots as linear instead of cubic. A restricted cubic spline can be mathematically expressed as [15] $s(z) = \eta_0 + \eta_1 x_1 + \eta_2 x_2 + \dots + \eta_{K-1} x_{K-1} K$, where K denotes the number of knots, x_i represents derived variables, and η_i describes the coefficients for these variables. This spline has the ability to fit complex shapes of baseline log cumulative hazard functions improving the stability of the function [4].

Multivariate survival regression models assume that there is no multicollinearity among covariates. Most of the survival methods are not appropriate to model large data with correlated covariates. The partial least squares (PLS) regression is considered as a good alternate of traditional regression methods in the presence of multicollinearity [5, 6].

Therefore, the partial least squares-Cox (PLS-Cox) regression model was developed to analyze survival systems in the presence of multicollinearity [7]. Due to several limitations of the PLS-Cox regression model, the PLS flexible parametric (PLS-FP) survival regression model is proposed to estimate smooth hazard ratios of predictors and corresponding cumulative hazard functions and to extrapolate the survival model [2]. However, the major limitation of the PLS-FP model is that it is not appropriate for all complex shapes of hazard function. The motivation of this research was to develop a survival model that has the ability to model complex shapes in the presence of multicollinearity. The proposed method is developed by integrating partial least squares with the Royston and Parmar restricted cubic spline model, hence the named as the partial least squares spline (PLS-spline) model. This model has the ability to fit more complex shapes of baseline log cumulative hazard functions. The efficiency of the partial least squares spline (PLS-spline) model is tested using simulated data by examining its performance on different scales with various spline knots. The proposed model is applied to a real data set of infant mortality to estimate the hazard function and regression coefficients. The analyses based on different scales using simulated and real data set reveal the efficiency of these models to estimate baseline log cumulative hazard functions in the presence of multicollinearity.

2. Materials and Methods

2.1. The Cox Proportional Hazard Model. For the occurrence of an event at time t , the Cox model assumes the hazard function in the presence of censoring

$$\lambda(t) = \lambda_o(t) \exp[\beta' \mathbb{X}], \quad (1)$$

where $\lambda_o(t)$ is the baseline hazard function, β is the vector of coefficients, and \mathbb{X} is a $(n * p)$ matrix of covariates. In this model, the baseline hazard function is unspecified.

2.2. The Partial Least Squares-Cox (PLS-Cox) Regression Model. Partial least squares-Cox (PLS-Cox) regression model is used as a benchmark model in this study. Let t represent the survival time and $\mathbb{X} \in \mathbb{R}^{n * p}$. The partial least squares model computes k latent components for p correlated covariates; then, the Cox model assumes the baseline hazard function as

$$\lambda(t) = \lambda_o(t) \exp[\beta' S], \quad (2)$$

where $\lambda_o(t)$ is the unspecified baseline hazard function, β is the vector of coefficients, and S is a $(n * k)$ matrix of components. The hyperparameters are found by maximum likelihood estimation method.

2.3. The Royston-Parmar Spline Model. In the context of the PH model, the Royston-Parmar (RP) model can be expressed as

$$\ln(H(t) | x_i) = s(\ln(t) | \eta, k_o) + \beta x_i, \quad (3)$$

where $s(\ln(t) | \eta, k_o)$ describes a restricted cubic spline that is a function of the derived variables η and the number of knots k_o . Generally, three different scales, hazard, odds, or normal, are used to model the RP spline model. When no knots are specified, the restricted cubic spline reduces to the Weibull distribution if the scale is hazard. For odds and normal scales, no knots give log-logistic and lognormal models, respectively.

2.4. Partial Least Squares Spline (PLS-Spline) Survival Regression Algorithm. Let $\mathbb{X} \in \mathbb{R}^{n * p}$ denote the matrix of p correlated covariates x_1, \dots, x_p for a sample of size n . The algorithm executes the FP model based on the C components (as

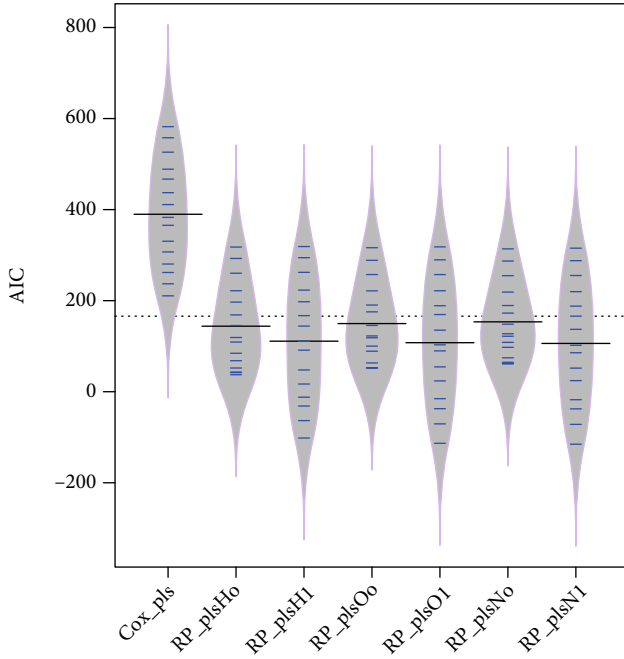


FIGURE 1: The efficiency of benchmark and proposed survival methods for simulated data set based on AIC is presented.

$C \leq p$) of PLSR computed with time T as a response variable and \mathbb{X} as a matrix of covariates for $c = 1, 2, \dots, C$. The pseudo-code for the proposed PLS-spline model is expressed as follows.

2.5. Data Simulation. Simulated data is generated using the `simsurv` R-package to evaluate the efficiency of existing and proposed survival models. The simulated data set is generated from Weibull distribution for the scale parameter ($\lambda = 0.1$) and shape parameter ($k = 1.5$) over 5 years of censoring. The correlation structure between 200 covariates ranged from 0 to 0.9 over 100 samples.

2.6. Real Data Set. This study used publically available secondary data, borrowed from the Demographic and Health Survey (DHS), collected during 2012-13 from Pakistan with the support of the United States Agency for International Development and ICF International. Therefore, there are no ethical concerns involved in this work, and no ethics review is required for this study [8]. The secondary data of infants from birth to aged 12 months born to ever married women aged 15-49 years in Pakistan is used in this study. The outcome of interest was infant survival within 12 months after first month of birth. The sample consists of 80 infants belonging to Pakistan, and 86 covariates are included.

3. Results

3.1. Simulation-Based Results. Using Weibull distribution, the high dimensional simulated data set having multicollinearity is generated. The constructed data is then split into test and training sets with 70 : 30 to train and evaluate the performance of benchmark and proposed methods. The hazard, odds, or normal scales are modeled each with zero and one knot.

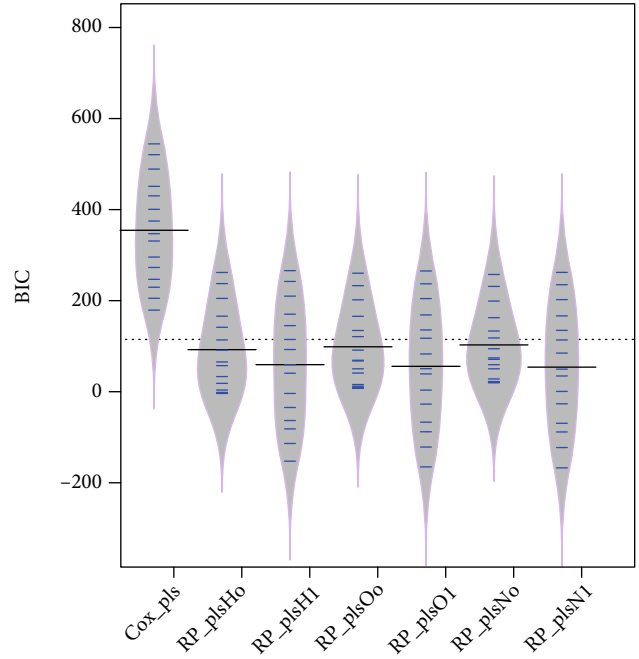


FIGURE 2: The efficiency of existing and proposed survival methods for simulated data set based on BIC is presented.

The PLS-spline model with different knots measured on different scales is fitted over the simulated data set generated from Weibull distribution to access the performance of models based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Figure 1 shows the comparison between the standard, PLS-Cox regression model, and six PLS-spline models with different knots based on various scales. The proposed PLS-spline models based on the hazard scale with zero knot and one knot are symbolized as RP_plsH_0 and RP_plsH_1 , respectively. Similarly, RP_plsO and RP_plsN stand for odds and normal scales accordingly. Figure 1 shows that the PLS-spline model based on all three scales with one knot has the highest performance compared to the PLS-Cox and PLS-spline models with zero knot. But it is also clear from Figure 1 that the PLS-spline model having zero knot showed even higher efficiency than the benchmark PLS-Cox method. Figure 2 shows the efficiency comparison based on the BIC defending performance based on AIC.

3.2. Application

3.2.1. Infant Survival Time Data Set. A cluster heat map presented in Figure 3 is used to show the magnitudes of correlation among covariates. Negative correlations are shown in blue color, and positive correlations are presented in red. High intensity of colors shows higher correlation among corresponding variables. Only 36 covariates are selected for examining multicollinearity for comprehensible visualization. Figure 3 clearly portrays the correlation between covariates showing intense colors.

The presence of multicollinearity is evident in the heat map. Hence, the existence of multicollinearity among covariates in high dimensional survival data is detected visually.

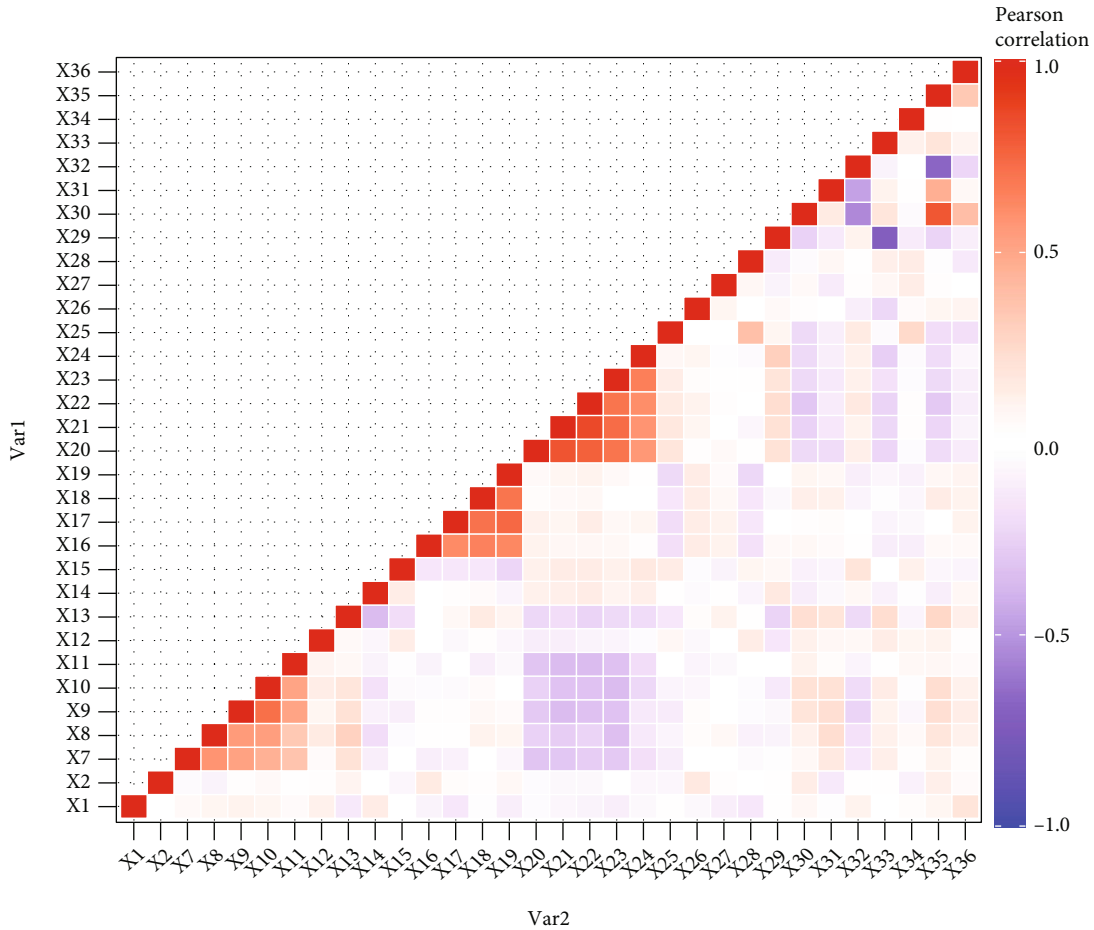


FIGURE 3: The heat map for infant survival time data.

The high dimensional infant survival data set having multicollinearity is used for comparison of models and identification of risk factors of infant mortality. The sample data is split into test and training sets with 70:30 to evaluate the efficiency of PLS survival methods.

The PLS-spline models with zero and one knot are fitted over the real data set to access the performance of models based on different scales using AIC and BIC. Figure 4 shows the comparison presenting the higher efficiency of all proposed methods compared to PLS-Cox based on AIC. Also, the highest performance of RP_plsO_1 is observed in Figure 4 compared to other RP_pls methods. This result showed that the proposed PLS-spline model based on the odds scale with one knot is the optimal model for the observed data.

Figure 5 shows the comparison of models based on BIC. The visual representation showed that the PLS-spline model based on the odds scale with zero and one knot has nearly the same efficiency. On the basis of both model assessment criteria, we may conclude that the PLS-spline model based on the odds scale is the best fitted model for the observed data. For identification of significant risk factors, the PLS-spline model based on the hazard scale with one knot is executed as being best fitted.

Table 1 presents the selected influential risk factors of infant mortality by using the RP_plsO_1 as being the optimal

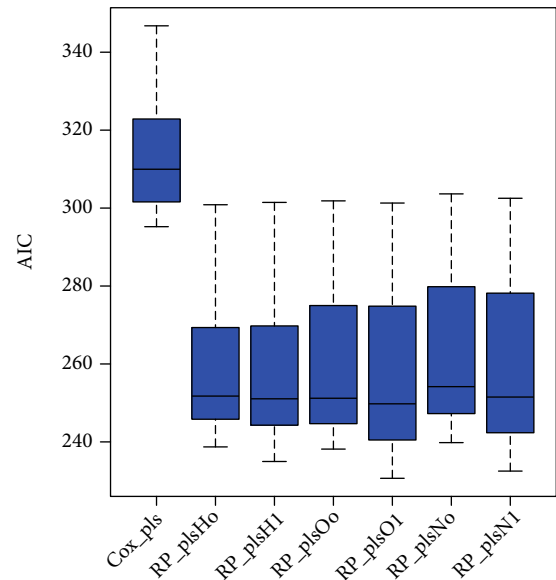


FIGURE 4: The efficiency of existing and proposed survival methods for infant survival data set based on AIC is presented.

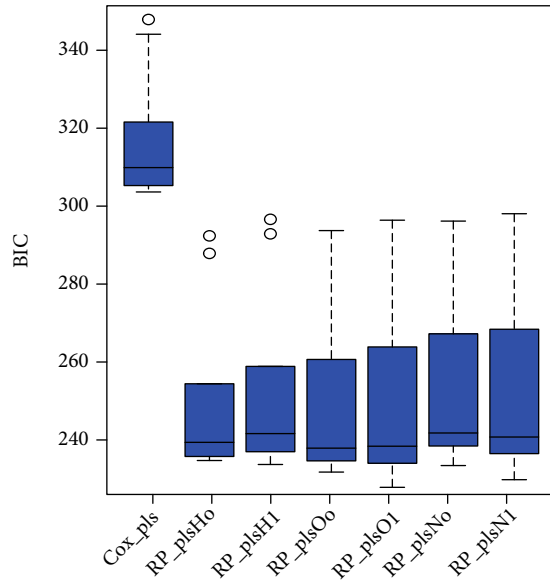


FIGURE 5: The efficiency of existing and proposed survival methods for infant survival data set based on BIC is presented.

model. After analysis, 27 influential factors are found significantly associated with infant mortality in Pakistan. The positive association of mother’s age, type of place of region, de facto place of residence, relationship of mother to household head, type of cooking fuel, number of births in last five years, distance, transport and accompany to health facility, mother’s occupation, person who usually decides on respondent’s health care, person who usually decides on visits to family or relatives, person who usually decides what to do with money husband earns, succeeding birth interval, and blood relation with husband is found for infant mortality. Furthermore, negative association of region, selection for domestic violence, household has motorcycle/scooter, reading newspaper or magazine, watching television, wealth index, awareness of tuberculosis and hepatitis, beating justified if wife neglects the children or argues with husband or if wife burns the food, and preceding birth interval is observed.

Figure 6 shows the estimates of the baseline cumulative hazards from the PLS-spline model measured on hazard, normal, and odds scales with zero and one knot for the data set of infant survival. All six PLS-spline models produce smooth estimates of the baseline cumulative hazards extrapolated to time of 12 months showing consistent estimates. The PLS-spline model based on the odds scale with one knot is represented by the red line in Figure 6 showing the lowest cumulative hazard for the first 4 months after birth, moderate increase in the fifth month, and maximum at the sixth month.

4. Discussion

Alongside advances in statistical techniques, several modifications are suggested for survival analysis to improve efficiency of the model. Yang et al. [9] introduced Deep-CoxPH, an estimation strategy based on deep learning and the Cox model which is proposed to improve the risk strat-

TABLE 1: Regression estimates of finally fitted PLS-spline model based on odds scale with one knot to select influential factors of infant mortality.

Selected factor	Estimate
Mother’s age	0.156
Region of residence	-0.191
Type of place of residence	0.257
De facto place of residence	0.258
Selected for domestic violence module	-0.164
Household has motorcycle/scooter	-0.133
Relationship of mother to household head	0.125
Reading newspaper or magazine	-0.108
Watching television	-0.222
Type of cooking fuel	0.133
Wealth index	-0.146
Number of births in last five years	0.103
Getting medical help for self: problem due to distance to health facility	0.197
Getting medical help for self: problem having to take transport	0.185
Getting medical help for self: not wanting to go alone	0.255
Awareness of tuberculosis	-0.126
Mother’s occupation	0.129
Person who usually decides on respondent’s health care	0.247
Person who usually decides on visits to family or relatives	0.170
Person who usually decides what to do with money husband earns	0.253
Beating justified if wife neglects the children	-0.191
Beating justified if wife argues with husband	-0.178
Beating justified if wife burns the food	-0.106
Preceding birth interval	-0.126
Succeeding birth interval	0.100
Blood relation with husband	0.153
Awareness about hepatitis	-0.147

ification for overall survival analysis. Rueda et al. [10] used discrete-time Markov chain theory and the Cox regression to predict survival function. The authors also employed a parametric analysis for comparison and variable selection. Another study developed an algorithm as a conjugate of the parametric model and partial least squares in the presence of extreme observations to enhance model performance [2]. In this study, the PLS-spline model is proposed to treat survival response with collinear predictors using the spline strategy based on different scales with various knots regarding better model performance and superior interpretation potential. To examine hazard function with higher accuracy, the PLS-spline model is proposed by integrating PLS and the Royston and Parmar spline model in the presence of multicollinearity. The proposed model is compared with the PLS-Cox model using simulated and real data sets for efficiency comparison. The PLS-spline model with one knot over hazard, odds, and normal scales turns out to be the best model

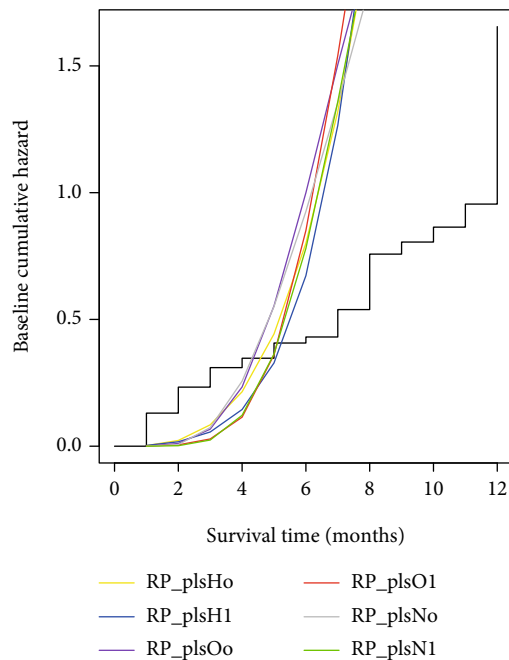


FIGURE 6: The estimates of the baseline cumulative hazard from PLS-spline model measured on different scales for infant survival data.

to estimate cumulative hazards based on AIC and BIC over simulated data generated from Weibull distribution. More importantly, for known simulated data, the PLS-spline model showed better performance than the PLS-Cox model. For the real data set of infant mortality, the PLS-spline model with one knot over the odds scale is observed to be optimal model. The finally selected model is used to identify the influential risk factors of infant mortality in Pakistan. Maternal age, occupation, and place of residence are found to be significant predictors of infant mortality in the present study. Previous studies observed that younger and older maternal ages are significantly associated with infant mortality [11]. Another study reported that the region of residence and working status of mother are statistically significant risk factors for stunted, underweight, and wasted children [12]. Consistent with literature, domestic violence is found to be significantly associated with infant mortality [13]. The present study observed that an increase in media awareness (watching television and reading newspaper) and wealth level could decrease the ratio of infant mortality. Literature described that media exposure and income level are associated with maternal outcomes [14, 12]. Availability and utilization of health facility is determined an important risk factor of mortality rate among infants. Several former studies verified that health expenditure potentially reduces maternal and infant mortalities across different countries [15, 16]. Closely similar to previous literature, birth interval and consanguineous marriage showed a significant association with infant mortality [17, 18]. The overall accuracy of the proposed algorithm enhances the model performance to a higher extent, considering collinear covariates. This efficiency suggests that survival function, hazard function, cumulative hazard function, and parameters of distribution for the survival time data with unknown distribution can

be estimated more efficiently in terms of smooth lines. The PLS-spline model is viewed as a useful addition to the toolbox of estimation and prediction of survival time response for the widely used PLS-Cox model in the survival settings.

5. Conclusion

The proposed PLS-spline model based on different scales with various knots is shown to be a better choice regarding model performance and superior interpretation potential. Using the PLS-spline model based on the odds scale with one knot, the influential factors identified as the important predictors of infant mortality are in agreement with other studies. So, the PLS-spline model has the potential as a multivariate survival technique in scientific research to treat high-dimensional correlated survival times data more efficiently.

Data Availability

Data are freely available at <http://www.dhs.org>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] R. B. D'Agostino, S. Grundy, L. M. Sullivan, P. Wilson, and Group, C. R. P, "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation," *Jama*, vol. 286, no. 2, pp. 180–187, 2001.
- [2] M. Sadiq and T. Mehmood, "A flexible and robust approach to analyze survival systems in the presence of extreme observations," *Mathematical Problems in Engineering*, vol. 2021, Article ID 9927377, 11 pages, 2021.
- [3] R. Ng, K. Kornas, R. Sutradhar, W. P. Wodchis, and L. C. Rosella, "The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review," *Diagnostic and prognostic research*, vol. 2, no. 1, pp. 1–15, 2018.
- [4] P. Royston and M. K. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in medicine*, vol. 21, no. 15, pp. 2175–2197, 2002.
- [5] T. Mehmood, M. Sadiq, and M. Aslam, "Filter-based factor selection methods in partial least squares regression," *IEEE Access*, vol. 7, pp. 153499–153508, 2019.
- [6] M. Sadiq, T. Mehmood, and M. Aslam, "Identifying the factors associated with cesarean section modeled with categorical correlation coefficients in partial least squares," *PLoS One*, vol. 14, no. 7, p. e0219427, 2019.
- [7] P. Bastien, V. E. Vinzi, and M. Tenenhaus, "PLS generalised linear regression," *Computational Statistics & Data Analysis*, vol. 48, no. 1, pp. 17–46, 2005.
- [8] P. Demographic, "Health survey 2012-13. Islamabad and Calverton, MA: National Institute of Population Studies and ICF International; 2013," 2015, Available at: <https://dhsprogram.com/data>.

- [9] C.-H. Yang, S.-H. Moi, F. Ou-Yang, L.-Y. Chuang, M.-F. Hou, and Y.-D. Lin, "Identifying risk stratification associated with a cancer for overall survival by deep learning-based CoxPH," *IEEE Access*, vol. 7, pp. 67708–67717, 2019.
- [10] L. Rueda, S. Sansregret, B. Le Lostec, K. Agbossou, N. Henao, and S. Kelouwani, "A probabilistic model to predict household occupancy profiles for home energy management applications," *IEEE Access*, vol. 9, pp. 38187–38201, 2021.
- [11] A. W. Ratnasiri, S. Lakshminrusimha, R. A. Dieckmann et al., "Maternal and infant predictors of infant mortality in California, 2007-2015," *PLoS one*, vol. 15, no. 8, p. e0236877, 2020.
- [12] S. J. Rahman, N. F. Ahmed, M. M. Abedin et al., "Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach," *PLoS One*, vol. 16, no. 6, p. e0253172, 2021.
- [13] P. Memiah, T. Bond, Y. Opanga et al., "Neonatal, infant, and child mortality among women exposed to intimate partner violence in East Africa: a multi-country analysis," *BMC Women's Health*, vol. 20, no. 1, pp. 1–16, 2020.
- [14] A. O. Igbino, E. O. Soola, O. Omojola, J. Odukoya, O. Adekeye, and O. P. Salau, "Women's mass media exposure and maternal health awareness in Ota, Nigeria," *Cogent Social Sciences*, vol. 6, no. 1, p. 1766260, 2020.
- [15] K. E. Agho, O. K. Ezeh, A. J. Ferdous, I. Mbugua, and J. K. Kamara, "Factors associated with under-5 mortality in three disadvantaged East African districts," *International Health*, vol. 12, no. 5, pp. 417–428, 2020.
- [16] P. A. Owusu, S. A. Sarkodie, and P. A. Pedersen, "Relationship between mortality and health care expenditure: sustainable assessment of health care system," *PLoS One*, vol. 16, no. 2, p. e0247413, 2021.
- [17] S. Anwar, J. Taslem Mouroso, Y. Arafat, and M. J. Hosen, "Genetic and reproductive consequences of consanguineous marriage in Bangladesh," *PLoS One*, vol. 15, no. 11, p. e0241610, 2020.
- [18] A. F. Dadi, "A systematic review and meta-analysis of the effect of short birth interval on infant mortality in Ethiopia," *PLoS One*, vol. 10, no. 5, p. e0126759, 2015.

Research Article

Data Analysis and Computational Methods for Assessing Knowledge of Obesity Risk Factors among Saudi Citizens

Alanazi Talal Abdulrahman ¹ and Dalia Kamal Alnagar ^{2,3}

¹Department of Mathematics, University of Ha'il, Saudi Arabia

²Department of Statistics, University of Tabuk, Saudi Arabia

³Department of Statistics, Omdurman Islamic University, Sudan

Correspondence should be addressed to Alanazi Talal Abdulrahman; t.shyman@uoh.edu.sa

Received 4 August 2021; Revised 14 September 2021; Accepted 8 October 2021; Published 26 October 2021

Academic Editor: Miguel G. Torres

Copyright © 2021 Alanazi Talal Abdulrahman and Dalia Kamal Alnagar. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction. According to the World Health Organization (2020), obesity is a growing problem worldwide. In fact, obesity is characterized as an epidemic. **Objective.** The aim of this paper is to use a logistic regression model as one of the generalized linear models and decision tree as one of the machine learning in order to assess the knowledge of the risk factors for obesity among citizens in Saudi Arabia. **Methods and Materials.** A cross-sectional questionnaire was given to the general population in KSA, using Google forms, to collect data. A total of 1369 people responded. **Results.** The findings showed that there is widespread knowledge of risk factors for obesity among citizens in Saudi Arabia. Participants' knowledge of risk factors was very high (95.5%). In addition, a significant association was found between demographics (gender, age, and level of education) and knowledge of risk factors for obesity, in assessing variables for knowledge of the risk factors for obesity in relation to the demographics of gender and level of education. In addition, from decision tree results, we found that level of education and marital status were the most important variables to affect knowledge of risk factors for obesity among respondents. The accuracy of correctly classified cases was 95.5%, the same in logistic regression and decision tree. **Conclusion.** The majority of participants saw regular exercise and diet as an essential way to reduce obesity; however, awareness campaigns should be maintained in order to avoid complacency and combat the disease.

1. Introduction

According to the World Health Organization, obesity is a growing problem worldwide. In fact, obesity is characterized as an epidemic, affecting nearly 650 million adults worldwide. Specifically, about 13% of the adult population worldwide is considered obese [1].

In Saudi Arabia, obesity is a growing problem. A systematic analysis of data from 33 years before 2014 revealed that Saudi Arabia was one of the top seven countries with the most severe rises in both male and female obesity rates [2]. Nearly 35% of Saudi adults are obese, with females having higher rates than males (41% vs. 31%) [3]. There was evidence to suggest that obesity is set to worsen in the decades to come [4]. Obese persons are more likely to develop heart disease, hypertension, stroke, type 2 diabetes, and other

adverse health outcomes than nonobese persons. From an economic perspective, obesity costs the country about \$147 billion in 2008 American dollars annually, and the annual cost is rising substantially over the past 10 years or so.

A variety of studies addressed the topic of obesity and its causes, which provided statistical data that show its prevalence and the associated factors. Among these studies is a study by Mosli et al. [5], which discovered the association between educational level and income level with odds of being obese among adults in KSA. In contrast to participants with advanced education or higher, ignorant participants and those with rudimentary schooling had higher chances of corpulence. In any case, members with low pay had lower chances than members who had higher pay.

A study by Al-Raddadi et al. [6] is aimed at studying the relationship of demographic and way of life factors, recently

demonstrated to be related to overabundant weight in different populaces, and BMI in the grown-up populace of Jeddah of KSA. In the study, there were 1419 persons: 667 males and 752 females. 30.1% males and 35.6% females were the prevalence of overweight and obesity, the prevalence increased to 60 years, and it decreased in the older age group in both genders. In males, the risk of obesity increased with obtaining a postgraduate degree, and the rate decreases with increased physical activity, and in females, obesity increased the risk of prediabetes and diabetes; the risk of prediabetes, diabetes, dyslipidemia, and hypertension increases with increasing BMI.

The study of Al-Qahtani [7] was aimed at appraising the commonness of overweight and obesity among grown-ups going to essential medical service settings, southwestern district of the KSA. Data on BMI estimation was recorded for 1649 out of 1681 individuals (98.1%). The general mean weight was 74.1 ± 15.81 kg, and that for males was 77.69 ± 16.14 kg versus 69.37 ± 14.02 kg for females. The general predominance of overweight and stoutness was, individually, 38.3% and 27.6%. Smoking was not essentially connected with corpulence, though hypertension was altogether connected with weight. The danger of overweight or corpulence essentially expanded from the most elevated to the least month-to-month pay. High spread of obesity and overweight should be considered a public health worry to be trailed by explicit mediations at the network level with multidisciplinary exercises beginning from childhood as an early stage counteraction program.

The point of the study of Aljabri et al. [8] was to assess obesity and overweight in Saudi women of childbearing age. Age was 32.3 ± 9.1 years (least 15 and most extreme 49 years), 5.8% (165) were lean, 26.6% (759) were of typical weight, 27.6% (785) were overweight, 22.4% (637) were obese Grade I, and 11.1% (316) were large Grade II while 6.6% (187) were beefy beyond belief (obese Grade III). The recurrence of overweight and stoutness expanded with the advance of age gathering, and dismal corpulence was most elevated in the 40-long-term age gathering. Except if quick advances are taken to contain the expanding commonness of weight, the medical care costs for ongoing sicknesses will represent a colossal budgetary weight to the KSA.

Albin Saleh et al. [9] assessed obesity commonness among kids and teenagers in Al-Ahsa, KSA, for the year 2016 and decided the connected preventable danger factors. Obesity and overweight were 29.6% (10.8% overweight, 3.8% fat, and 15% obese large). The prevalence of obesity and overweight was altogether connected with youth weight, parental overweight, mother's work, family pay, fast food, actual dormancy, and time spent sitting in front of the TV. There is an earnest need to spread mindfulness about obesity, and the anticipation programs that include schools and families are the vital systems for controlling the current epidemic of overweight and obesity.

Alshammari and Elsbali [10] measured the prevalence paces of obesity in Hail City in KSA. 80.83% (1455/1800) have fully responded to all required parameters of the 1800, 52% (756/1455) were females and 48% (699/1455) were males, and 60.34% (878/1455) were found obese and overweight, with females' proportion more than males. Obe-

sity and overweight are common in Hail City in KSA with generally higher females' susceptibility.

Al-Hazzaa et al. [11] sought to give updated estimates of obesity and overweight prevalence from three main cities in Saudi Arabia, namely, Riyadh, Al-Khobar, and Jeddah, with members of 2,908 auxiliary school understudies aged 14 to 19 years; the prevalence of overweight was 19.5 percent in individuals and 20.8 percent in girls, while stoutness was 24.1 percent in males and 14 percent in females. The predominance of obesity in males and females was 35.9% and 30.3%, respectively. Such high pervasiveness of obesity and overweight is a significant public health concern.

Al-Ateeq and Al-Hargan [12] looked at the potential relationship between obesity and the method of transportation to neighborhood offices, social climate, type of work, and actual movement at neighborhood offices and at home. Of the participants, 33.7% were overweight and only 39.2% were obese. Most of the members traveled to work (98%), school (90.2%), shopping centers (95.7%), eateries (91.5%), social visits (84%), mosques (84.3%), and markets (50.2%). The rate of obesity was higher among members who drove (45%) than among those who walked (30%) to the market stores. Thus, the proposed paper's principal goal is to assess the knowledge of the risk factors for obesity among citizens in Saudi Arabia towards the risk factors for obesity.

Knowledge is the ability to learn, retain, and apply knowledge; it is a combination of comprehension, experience, discernment, and skill.

Many researches have reported the varied prevalence of obesity and overweight among Saudi residents, but there is little information on individuals' knowledge of the risk factors for obesity. As a result, the findings of this study are critical since they will assist in the management of obesity.

Nonetheless, there has been an expanded interest in defining new factual models or new groups of measurable models to give a superior depiction of the issues viable. For more details, we refer to Abdulrahman and Alamri [13] and Abdulrahman [14].

2. Materials and Methods

2.1. The Questionnaire. A questionnaire was used in this assessment to allow individuals to identify obesity risk factors. The questionnaire comprised two sections. Section one included questions on personal information, including gender, age, marital status, educational qualifications, and occupation. The other section contained ten questions on the causes of obesity. The participants were asked to determine the main cause of obesity among the following reasons:

- (i) Heredity is one of the most important causes of obesity
- (ii) Diet pattern has a significant impact on the causes of obesity
- (iii) Consumption of sugars and starches is a major cause of obesity
- (iv) Lack of exercise is a major cause of obesity

- (v) Lack of sleep of a person may expose him to obesity
- (vi) Some hormones such as leptin can increase obesity
- (vii) Taking some medications may expose a person to obesity
- (viii) As a person gets older, he is at risk of obesity
- (ix) Some diseases that affect a person are exposed to obesity
- (x) Mental state is one of the most important causes of obesity

The target populace was 748 male responders and 621 female respondents.

2.2. Data Analysis and Study Test. Table 1 shows demographic characteristics such as age, gender, education level, marital status, and work. All participants spoke Arabic fluently. SPSS version 25.0 was used to analyze the data. Quantitative analysis entailed calculating frequencies and percentages for demographic data, which were then tested using inferential statistics. Pearson's chi-squared test was used to evaluate the analyses' goodness of fit; homogeneity, to compare respondents (groups) in a specified variable; and independence, to determine whether respondent cohorts exhibited distinct answers.

2.3. Binary Logistic Regression Model. Binary logistic regression is a statistical method used to investigate a variety of subjects in medical research [15, 16]. It helps researchers to predict whether an event will occur or not based on predictor factors [17].

The odds ratios for each of the model's independent variables (age, gender, marital status, level of education, and work) were estimated using logistic regression. When the odds ratio is more than one, it shows a positive relationship, and when it is less than one, it suggests a negative correlation. To forecast a logit transformation of the likelihood of the presence of the attribute of interest, use the following formula:

$$\text{Logit}(p) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k. \quad (1)$$

Here, p is the probability of the occurrence of the property of interest.

The logged chances are defined as the logit transformation:

$$\text{Odds} = \frac{p}{1-p}. \quad (2)$$

Here, p indicates the probability of a characteristic's presence, $1-p$ represents the probability of a characteristic's absence, and

$$\text{Logit}(p) = \log \log \frac{p}{1-p}. \quad (3)$$

TABLE 1: The demographic information.

Variable	Frequency	Percent
Gender		
Male	748	54.6
Female	621	45.4
Age		
Less than 18	76	5.6
18-30	602	44.0
30-40	394	28.8
40 and more	297	21.7
Marital status		
Married	871	63.6
Unmarried	498	36.4
Level of education		
Primary and middle school	45	3.3
Secondary school	215	15.7
University student	923	67.4
Postgraduate	186	13.6
Work		
Government employee	616	45.0
Private sector	177	12.9
I do not work	576	42.1

The logit is a function that translates probability values from $(0, 1)$ to real numbers $(-\infty, \infty)$.

2.4. Decision Tree. A decision is a flowchart-like structure in which each internal node represents a test on an attribute, each branch of the tree represents a test outcome, and each leaf node stores a class label.

The Chi-square Automatic Interaction Detector (CHAID) method utilized in this research detects such differences by employing two tests to assess the relationship between the dependent and independent variables [18]. The CHAID process begins by identifying independent variables that have a statistically significant relationship with the dependent or target variable.

Decision tree techniques may be used to choose the most relevant input variables that should be utilized to build decision tree models, which can then be used to formulate clinical hypotheses and inform further research.

The data mining technique of decision tree analysis offers an alternative means of identifying specific variables affected by knowledge of the risk factors for obesity among the respondents, which included the model of participants' gender, age, marital status, education, level of education, and work.

The information gain may be used to select the appropriate attribute to utilize for data classification:

$$G(A) = I(p, n) - \sum_{i=1}^v \frac{P_{i+n_i}}{p+n} I(p_i, n_i), \quad (4)$$

where p is the probability that the tuple belongs to class V and n is the number of attributes in the class:

$$I(p, n) = \frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{p}{p+n} \log_2 \frac{n}{p+n}. \quad (5)$$

A binary outcome value for the i object is represented by n and p and takes zero and one values [17].

3. Results

The reliability analysis result showed that Cronbach's alpha was 0.68 for 10 items. Therefore, there was internal consistency of the scales. Hence, this instrument used in this study had a high reliability value (Alnagar et al. [19]).

The demographic information is shown in Table 1. Of the 1369 samples analyzed, $n = 748$ (54.6%) were male and $n = 621$ (45.4%) were female; the majority of the participants' ($n = 602$, 44.0%) ages range from 18 to 30 while the percentage of respondents with ages less than 18 was $n = 76$ (5.6%).

According to their marital status, $n = 871$ (63.6%) of respondents are married and $n = 498$ (36.4%) are unmarried. The majority of participants were university students ($n = 923$, 67.4%) while the percentage of respondents with postgraduate education levels were $n = 186$ (13.6%).

Table 2 shows that the participants' knowledge of the risk factors for obesity was very high (95.5%). Specifically, they were most knowledgeable about diet (99.3%), fast food (96.6%), heredity (74.3%), lack of exercise (93.5%), lack of sleep (82.8%), hormones (98.2%), increased age (67.3%), some diseases (90.4%), and mental stress (87.1%).

Table 3 shows the association between several demographic variables and knowledge of risk factors for obesity among respondents; females had significantly higher (97.1%) knowledge of risk factors for obesity than males (94.3%). Moreover, respondents of age ranging from 30 to 40 had significantly higher (96.7%) knowledge of risk factors for obesity than those of ages less than 18 (88.2%). Majority of the respondents married had 96.1% high knowledge of risk factors for obesity than those unmarried (94.9%), so there was no association between knowledge of risk factors for obesity and marital status.

University students had significantly higher (96.9%) knowledge of risk factors for obesity than primary and middle school students (94.3%). Private workers had high (96.6%) knowledge of risk factors for obesity than those that do not work (95.1%), so there is no association between knowledge of risk factors for obesity and work.

In sum, there were associations between knowledge of risk factors for obesity among the respondents and variables (gender, age, and level of education).

Table 4 shows a binary logistic regression model to estimate variables affected on knowledge of risk factors for obesity among respondents, including the model of participants' gender, age, marital status, education, level of education, and work. Items emerged as significant ($p \leq 0.05$) from the logistic regression analysis model; we found that gender and level of education were both variables affecting knowledge of risk

TABLE 2: Participants' knowledge of the risk factors for obesity.

Variable	Knowledge	Percent
Poor knowledge	61	4.5
Good knowledge	1308	95.5
Heredity	1018	74.3
Diet	1359	99.3
Fast food	1322	96.6
Lack of exercise	1280	93.5
Lack of sleep	1134	82.8
Some hormones	1344	98.2
Taking some medications	1300	95
The older a person	930	67.9
Some diseases	1238	90.4
Mental stress	1192	87.1

TABLE 3: Association between several demographic variables and knowledge of risk factors for obesity among respondents.

Variable	Poor	Good	p value
Gender			0.001
Male	43 (5.7%)	705 (94.3%)	
Female	18 (2.9%)	603 (97.1%)	
Age			0.007
Less than 18	9 (11.8%)	67 (88.2)	
18-30	29 (4.8%)	573 (95.2%)	
30-40	13 (3.3%)	381 (96.7%)	
40 and more	10 (3.4%)	287 (96.6%)	
Marital status			0.191
Married	34 (3.9%)	837 (96.1%)	
Unmarried	27 (5.4%)	471 (94.6)	
Level of education			0.001
Primary and middle school	8 (17.8%)	37 (82.2%)	
Secondary school	18 (8.4%)	197 (91.2%)	
University student	29 (3.1%)	894 (96.9%)	
Postgraduate	6 (3.2%)	180 (96.8%)	
Work			0.704
Government employee	27 (4.4%)	589 (95.6%)	
Private work	6 (3.4%)	171 (96.6%)	
I do not work	28 (4.9%)	548 (95.1%)	

factors for obesity among respondents. Gender as a variable showed a good odds ratio of 2.261 at 95% confidence interval (CI = 1.189, 4.301). There was a high knowledge of risk factors for obesity among respondents from those with level of education, with odds ratio = 2.054 (95% CI = 1.426, 2.957).

The classification results for the decision tree for knowledge of risk factors for obesity among respondents are shown in Table 5. The percentages of cases that were correctly classified were 95.5%, which demonstrates the accuracy of the decision tree model.

TABLE 4: Binary logistic regression model to estimate variables affecting knowledge of risk factors for obesity among respondents.

Variable	Standard error	p value	Odds ratio	95% CI for odds ratio	
				Lower	Upper
Age	0.189	0.061	1.424	0.983	2.063
Gender	0.328	0.013	2.261	1.189	4.301
Marital status	0.354	0.748	1.121	0.560	2.242
Level of education	0.186	0.001	2.054	1.426	2.957
Work	0.182	0.845	0.965	0.675	1.379
Constant	1.062	0.335	0.359		

TABLE 5: Classification table for the decision tree for knowledge of risk factors for obesity among respondents.

Observed	Predicted	Observed	Predicted
Poor knowledge	0	61	0.0%
Good knowledge	0	1308	100.0%
Overall percentage	0.0%	100.0%	95.5%

Growing method: Classification Regression Tree (CRT). Dependent variable: knowledge.

TABLE 6: Variable importance using CART methods.

Independent variable	Importance	Normalized importance
Age	0.002	100.0%
Level of education	0.002	90.7%
Marital status	0.001	35.5%
Work	0.000	22.0%
Gender	0.000	11.7%

Growing method: CRT. Dependent variable: knowledge.

Table 6 shows that decision trees were used to gain information to determine which variables are most important to affect knowledge of risk factors for obesity among respondents. Age, level of education, and marital status were the most important variables to affect knowledge of risk factors for obesity among respondents.

4. Discussion and Conclusion

The findings of this study are that there is widespread knowledge of risk factors for obesity among citizens in Saudi Arabia; it agrees with [20]. The participants' knowledge of obesity risk factors was generally high (95.5%). In addition, there was a high knowledge of risk factors for obesity among respondents from those with a level of education. A significant association was found between demographics (gender, age, and level of education) and knowledge of risk factors for obesity; it agrees with [21]. A decision tree was used to gain information to determine which variables are most important to affect knowledge of risk factors for obesity among respondents; the percentages of cases that were correctly classified are 95.5%, which demonstrates the accuracy of the decision tree model.

Accuracy of correctly classified cases was the same in two methods. However, the results are different in the logistic regression and decision tree; in the logistic regression analysis model, we found that both gender and level of education variables affected knowledge of risk factors for obesity among respondents. Age, level of education, and marital status were the most important variables to affect knowledge of risk factors for obesity among respondents.

Logistic regression is a statistical approach for modeling the probability p of an occurrence in terms of one or more predictor variables' values. The model is made up of two parts: a binary tree structure that depicts the data divisions and a series of basic linear logistic models, one for each partition. This is the division of model complexity that makes the model easy to interpret. In conclusion, the majority of participants saw regular exercise and diet as an essential way to reduce obesity; however, awareness campaigns should be maintained in order to avoid complacency and combat the disease.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

The authors would like to express their thanks to the University of Ha'il and the University of Tabuk. The Deanship of Research at University of Hail, Saudi Arabia, funded this project with project number BA-1903.

References

- [1] The World Health Organization, "Fact sheet, obesity & overweight," 2020, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
- [2] M. Ng, T. Fleming, M. Robinson et al., "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013," *The Lancet*, vol. 384, no. 9945, pp. 766-781, 2014.
- [3] World Health Organization, "Global health observatory data repository," 2018, <http://apps.who.int/gho/data/node.main.A900A?lang=en>.
- [4] A. Aljaadi and M. Alharbi, "Overweight and obesity among saudi children: prevalence, lifestyle factors, and health impacts," in *Handbook of Healthcare in the Arab World*, pp. 1-25, Springer Link, 2020.
- [5] H. H. Mosli, H. A. Kutbi, A. H. Alhasan, and R. H. Mosli, "Understanding the interrelationship between education, income, and obesity among adults in Saudi Arabia," *Obesity Facts*, vol. 13, no. 1, pp. 77-85, 2020.
- [6] R. Al-Raddadi, S. M. Bahijri, H. A. Jambi, G. Ferns, and J. Tuomilehto, "The prevalence of obesity and overweight, associated demographic and lifestyle factors, and health status

- in the adult population of Jeddah, Saudi Arabia,” *Therapeutic Advances in Chronic Disease*, vol. 10, 2019.
- [7] A. M. Al-Qahtani, “Prevalence and predictors of obesity and overweight among adults visiting primary care settings in the Southwestern region, Saudi Arabia,” *BioMed Research International*, vol. 2019, Article ID 8073057, 5 pages, 2019.
- [8] K. Aljabri, S. Bokhari, M. Alshareef, M. Khan, and B. Aljabri, “Overweight and obesity in Saudi women of childbearing age,” *EC Endocrinology and Metabolic Research*, vol. 3, pp. 53–62, 2018.
- [9] A. A. Albin Saleh, A. S. Alhaiz, A. R. Khan et al., “Prevalence of obesity in school children and its relation to lifestyle behaviors in Al-Ahsa district of Saudi Arabia,” *Global Journal of Health Science*, vol. 9, no. 12, p. 80, 2017.
- [10] E. M. Alshammari and A. M. Elsbali, “Obesity in Hai’l City, Kingdom of Saudi Arabia (KSA): is it a gender specific?,” *Egyptian Academic Journal of Biological Sciences. C, Physiology and Molecular Biology*, vol. 7, no. 1, pp. 83–90, 2015.
- [11] H. M. Al-Hazzaa, N. A. Abahussain, H. I. Al-Sobayel, D. M. Qahwaji, N. A. Alsulaiman, and A. O. Musaiger, “Prevalence of overweight, obesity, and abdominal obesity among urban Saudi adolescents: gender and regional variations,” *Journal of Health, Population, and Nutrition*, vol. 32, no. 4, 2014.
- [12] M. A. Al-Ateeq and M. H. Al-Hargan, “Relationships between overweight and obesity with preferred mode of transportation and use of neighborhood facilities in Riyadh, Saudi Arabia,” *Journal of Obesity & Weight Loss Therapy*, vol. 4, no. 4, 2014.
- [13] A. T. Abdulrahman and O. Alamri, “Robust estimation methods used to study the reasons behind increasing divorce cases in Saudi Society,” *Mathematical Problems in Engineering*, vol. 2021, Article ID 4027599, 6 pages, 2021.
- [14] A. T. Abdulrahman, “Methods for designing experiments to study the actual causes of the housing crisis,” *International Journal of Analysis and Applications*, vol. 19, no. 4, pp. 542–560, 2021.
- [15] H. Jr, W. David, S. Lemeshow, and X. Rodney, “Sturdivant,” in *Applied Logistic Regression*, pp. 4–20, John Wiley & Sons, 2013.
- [16] J. F. Hair, *Análisis Multivariante*, vol. 491, Prentice Hall, Madrid, Spain, 1999.
- [17] R. Alharbi, D. Alnagar, A. T. Abdulrahman, and O. Alamri, “Statistical methods to represent the anxiety and depression experienced in Almadinh KSA during Covid-19,” *JP Journal of Biostatistics*, vol. 18, no. 2, pp. 231–248, 2021.
- [18] A. Agresti, “On logit confidence intervals for the odds ratio with small samples,” *Biometrics*, vol. 55, no. 2, pp. 597–602, 1999.
- [19] D. Alnagar, R. Alharbi, O. Alamri, and O. Alamri, “Analysis of the female student academic performance using an exploratory factor analysis,” *Advances and applications in statistics*, vol. 69, no. 1, pp. 41–57, 2021.
- [20] S. F. Alfadhel, H. S. Almutairi, T. H. G. al Darwish, L. T. Almana, R. A. Aldosary, and A. H. Shook, “Knowledge, attitude, and practice of bariatric surgery among adult Saudi community, Saudi Arabia, 2019,” *Journal of Family Medicine and Primary Care*, vol. 9, no. 6, p. 3048, 2020.
- [21] T. X. Mangalathil, P. Kumar, and V. Choudhary, “Knowledge and attitude regarding obesity among adolescent students of Sikar, Rajasthan,” *IOSR Journal of Nursing and Health Science*, vol. 3, no. 2, pp. 44–48, 2014.