Research Article

An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values

Kumarmangal Roy ¹, **Muneer Ahmad** ¹, **Kinza Waqar** ¹, **Kirthanaah Priyaah** ¹,
Jamel Nebhen ², **Sultan S Alshamrani** ³, **Muhammad Ahsan Raza** ⁴, and **Ihsan Ali** ¹

¹Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

²Prince Sattam Bin Abdulaziz University, College of Computer Engineering and Sciences, P.O. Box 151, Alkharj 11942, Saudi Arabia

³Department of Information Technology, College of Computer and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

⁴Department of Information Technology, Bahauddin Zakariya University, Multan 60000, Pakistan

Correspondence should be addressed to Ihsan Ali; ihsanalichd@siswa.um.edu.my

Received 19 March 2021; Revised 7 April 2021; Accepted 12 April 2021; Published 6 July 2021

Academic Editor: M. Irfan Uddin

Copyright © 2021 Kumarmangal Roy 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.

Diabetes is one of the most common metabolic diseases that cause high blood sugar. Early diagnosis of such a condition is challenging due to its complex interdependence on various factors. There is a need to develop critical decision support systems to assist medical practitioners in the diagnosis process. This research proposes developing a predictive model that can achieve a high classification accuracy of type 2 diabetes. The study consisted of two fundamental parts. Firstly, the study investigated handling missing data adopting data imputation, namely, median value imputation, K-nearest neighbor imputation, and iterative imputation. Consequently, the study validated the implications of these imputations using various classification algorithms, i.e., linear, tree-based, and ensemble algorithms, to see how each method affected classification accuracy. Secondly, Artificial Neural Network was employed to model the best performing imputed data, balanced with SMOTETomek ensuring each class is represented fairly. This approach provided the best accuracy of 98% on the test data, outperforming accuracies achieved in prior studies using the same dataset. The dataset used in this study is concerned with gender and population. As a prospect, the study recommends adopting a larger population sample without geographic boundaries. Additionally, as the developed Artificial Neural Network model did not undergo any specific hyperparameter tuning, it would be interesting to explore tuning on top of normalized data to optimize accuracy further.

1. Introduction

Diabetes mellitus, or more commonly known as only diabetes, is a global health problem affecting people worldwide. According to [1], 463 million people have diabetes globally, and 88 million are from South East Asia. Diabetes is the general term for separate but related chronic diseases. It occurs when the human body is incapable of producing or properly using the hormone known as insulin [2]. Insulin is responsible for converting glucose from the food humans consume into the energy they need for daily life; thus, its imbalance is detrimental

to human health. The cause of diabetes remains a mystery, but studies showed genetics and environmental factors [3].

Additionally, for most countries, diabetes has evolved with dietary changes, aging populations, reduced physical activity, and other unhealthy lifestyle choices [4]. Essentially, there are two types of diabetes. Type 1 diabetes occurs when the body's cells responsible for producing insulin are destroyed, causing insulin deficiency. The more prevalent, type 2 diabetes occurs when insulin production is not sufficient or adequate to control blood glucose level, causing relative insulin deficiency [5].

Diabetes increases the risk of being susceptible to other diseases as its development often leads to heart attack, stroke, heart disease, and many other severe consequences [6]. The first step in the development of type 2 diabetes is prediabetes. Prediabetes means that the blood sugar level is higher than usual but not high enough to be type 2 diabetes. However, the long-term damage of diabetes, especially to the heart, blood vessels, and kidneys, may already start, making an early diagnosis of diabetes of vital importance [7]. Assessing symptoms of diabetes is a challenging task whereby 50% of patients with type 1 and type 2 symptoms remain underdiagnosed due to no noticeable signs of disease and its complex interdependence to other conditions [2]. There is a need for accurately identifying patterns and relationships from patient data. With the increase in the amount of data in medical databases, it has become too time-consuming and costly to assign staff to conduct traditional manual analysis to identify data patterns and recognize anomalies in patients' medical records.

In the survey conducted by National Diabetes and Diabetic Retinopathy, it is stated that the prevalence of diabetes in India is at 11.8% [8]. According to the National Health and Nutrition Examination Survey data, more people died from the consequences of diabetes than of HIV/AIDS, malaria, and tuberculosis combined. The World Health Organization has predicted that by the year 2030, diabetes will be the seventh leading cause of death [9]. By 2040 around 482 million people will be affected globally by prediabetes and an estimated 642 million by diabetes, according to the International Diabetes Federation [10]. These statistics show that type 2 diabetes does not discriminate. It has an impact on everyone across all classes, societies, countries, and continents. Unless effective diagnosis methods are utilized, this problem will remain prevalent. This issue will stress the healthcare systems, challenge economies, and predominantly affect life quality.

A practical solution is to employ data mining techniques for examining data and seeking useful information from medical databases that help clinical staff diagnose type 2 diabetes more accurately [11]. The healthcare industry generates a vast amount of valuable data comprising patient information, electronic medical records, treatment history, etc. Conducting knowledge extraction on these data can identify patterns and support decision-making and cost reduction by developing intelligent medical systems. Early detection can prevent diabetes with higher accuracy [12].

Researchers have utilized a variety of classification algorithms to achieve high-performing models. This research builds on previous literature by developing a classification model using data mining and laboratory information of the patients while applying different data imputation approaches to enhance model performance. Besides, we consider the imbalance in data distribution between diabetic patients and ordinary people. It helped in training the classification model to ensure that each class was represented fairly. By identifying the best performing data preprocessing methods, the machine learning model's performance accuracy was optimized, making it the highest among previous studies.

The positive impact of the proposed solution is on both the patients and the medical practitioners. The countries with a high rate of diabetic patients can employ an accurate early prediction system of diabetes. The more precise the prediction, the more accurate the severity estimation of the underlying medical condition [13]. Detecting diabetes at an early stage and treating it is an essential step towards keeping people with diabetes healthy. Early detection will benefit the patients and medical staff as they can make use of computer-based systems rather than relying on traditional manual analysis of data, which is costly, time-consuming, and prone to human errors [14].

The rest of the article is organized as follows: Section 2 covers a summary of prior literature, Section 3 presents the data and methods used to adopt the proposed solution. Section 4 outlines the results and discussion of the findings, followed by a conclusion in Section 5.

2. Literature Review

Diabetes mellitus is a common chronic disease found in numerous countries around the world. While it can seriously deteriorate an individual's quality of life, its early diagnosis can minimize the long-term adverse effects [15]. With the abundance of medical data from electronic records, patient information, treatment data, etc., researchers have utilized data mining techniques to discover knowledge through this data. Useful knowledge extraction can support decision-making and cost reduction of intelligent diabetes diagnosis systems. Table 1 outlines extensive research conducted in the early prediction of diabetes with a summary of selected 20 prior studies. For a fair comparison, all selected previous studies have used the same diabetes dataset publicly available on UCI Machine Learning Repository to conduct their investigation on the development of a diabetes predictive model [30]. Additionally, we noted data preprocessing in handling missing data and suggested solutions related to class imbalance problems for these studies.

Existing literature shows various data mining and machine learning algorithms such as neural networks, tree-based algorithms, rule-based fuzzy classifiers, linear classifiers, and hybrid models used to output a highly accurate predictive model. Artificial Neural Network (ANN) was the most popular classifier where 11 out of 20 prior studies used different ANN types in their experiments. More than half of the preliminary studies result in prediction accuracy within the 80%–90% range. Wu et al. [22] used a hybrid predictive model to get 95.42% accuracy in prediction, the highest in the literature. Medical datasets are likely to contain many missing data values. We reviewed their suggested solutions that influenced the classification accuracy. Earlier studies showed that using mean value imputation [2, 20, 22] led to better accuracy results than when missing data were deleted [11, 16, 27].

Consequently, most prior literature chose to delete missing values, with only 5 out of the 20 studies using imputation to replace missing values with meaningful values based on central tendency measures. Another aspect of data preprocessing in binary classification is often found in the

TABLE 1: Summary of machine learning solutions for diabetes classification.

Study	Preprocessing approach	Algorithm	Significance
Guldogan et al. [16]	Missing data deleted	Multilayer perceptron (MLP) and radial basis function (RBF)	MLP giving an accuracy of 78.1%
Alam et al. [17]	Median value imputation	ANN, random forest, K-means clustering	ANN giving an accuracy of 75.7%
Wang et al. [18]	naïve Bayes imputation, ADASYN oversampling	Random forest	87.10% accuracy
Sarwar et al. [19]	Method details not stated	K-nearest neighbors, naïve Bayes, support vector machine, decision tree, logistic regression, random forest	K-nearest neighbors and support vector machine giving accuracy 77%
Woldemichael and Menaria [20]	Mean value imputation	Backpropagation, support vector machine, J48, naïve Bayes	Backpropagation giving an accuracy of 83.11%
Sisodia and Sisodia [21]	Method details not stated	Decision tree, support vector machine, naïve Bayes	Naïve Bayes giving an accuracy of 76.30%
Wu et al. [22]	Mean value imputation	K-means + logistic regression	95.42% accuracy
Komi et al. [13]	Method details not stated	Gaussian mixture model, support vector machine, logistic regression, extreme learning machine, ANN	ANN giving an accuracy of 89.0%
AlThunayan et al. [15]	Method details not stated	Bayesian, naïve Bayes, J48, random tree, random forest, REP tree, FT tree, CART, SMO	Random forest giving an accuracy of 84.11%
Vaishali et al. [23]	Missing data kept unchanged	Naïve Bayes, J48, MLP, MOE fuzzy classifier	MOE fuzzy giving accuracy of 83.04%
Geman et al. [24]	Missing data kept unchanged	Hybrid adaptive neural network	84.27% accuracy
Ramesh et al. [25]	Missing data kept unchanged	Recurrent deep neural Network (RNN)	0.145 error rate
Soltani and Jafarian [14]	Missing data kept unchanged	ANN types: PNN, MLP, RBF, GRNN	PNN giving an accuracy of 81.49%
Ahmad et al. [2]	Mean value imputation	MLP, ID3, J48	J48 giving an accuracy of 89.3%
AlJarullah [11]	Missing data deleted	J48	78.17% accuracy
Marcano-Cedeño et al. [26]	Missing data deleted	Artificial metaplasticity on multilayer perceptron algorithm (AMMLP)	89.93% accuracy
Patil et al. [3]	Missing data deleted	K-means + C4.5	92.38% accuracy
Lekkas and Mikhailov [27]	Missing data deleted	Rule-based fuzzy classification	79.37% accuracy
Kahramanli and Allahverdi [28]	Method details not stated	ANN + fuzzy neural network	84.24% accuracy
Han et al. [29]	Missing data deleted	ID3, decision tree	ID3 giving an accuracy of 80.0%

medical domain to solve the class imbalance problem. The data contain larger samples of class labeled as not having diabetes than having diabetes. It can lead to a biased predictive model where the accuracy applies only to the majority class. Among prior studies, only [18] addresses both missing values and the class imbalance problem. The authors used naïve Bayes imputation to replace missing values with meaning values and use ADASYN oversampling to get equal representation of both classes in training the predictive model and achieve 87.10% accuracy.

A high-quality classification model is one whose accuracy reflects its capability of generalizing unseen data samples. The development of such a model essentially requires more data representing both classes fairly. While deleting missing data values is a simple solution, it risks losing data points with valuable information reducing the model's ability to generalize [31]. Furthermore, skewed class distribution will ensure that predictive performance applies to both majority and minority classes. Therefore, despite the extensive reported research, prediction of diabetes remains a challenging problem that needs to be further explored,

especially in terms of investigating the effect of preprocessing approaches on classification accuracy. There is a need to experiment with various data imputation methods and solve class imbalance problems to ensure that the predicting model's resulting accuracy is of high quality.

3. Methodology

Optimal classification performance is dependent on efficient data preparation and data preprocessing methods. The proposed algorithm fills the research gap identified in the literature review. Handling missing values of attributes through a meaningful approach can substantially improve a machine learning model's performance [32]. Initially, the investigation focused on a comparative study of three different data imputation methods to identify the optimal way of handling missing values in the dataset. Consequently, data from the highest performing method was employed to achieve optimized performance in modeling.

Figure 1 presents the proposed framework for type 2 diabetes prediction. It has two phases where the first phase

FIGURE 1: Proposed framework for type 2 diabetes prediction.

consists of identifying optimal imputation technique. We apply three different imputation techniques after loading the data, namely, median, iterative, and KNN imputations. We handle each attribute outliers and split instances into a proportion of 70% training and 30% testing, modeling using seven different machine learning algorithms. Table 2 outlines these algorithms below.

We chose the highest performing outlier-treated imputed dataset for the second phase. In the second phase, we adopt a random sampling technique to produce some noise in the data. It then further preprocess to handle the class imbalance problem using undersampling and oversampling. It gives an equal representation of each class in the machine learning model's training data. The preprocessed data are then split into 70% train and 30% test datasets. Standard scaling smoothens and prepares the training data for modeling. ANN is chosen as the machine learning model based on its common use in past literature. ANN model architecture is created with 1 initial layer, 5 hidden layers, and 1 final layer. RELU activation function is applied to each layer except the final layer. Instead, the sigmoid activation function is used in the final layer. Adam optimizer is used for model compiling and "binary_crossentropy" is used as a loss function. After a few trials, we could conclude that 10 epochs

showed good results in terms of "accuracy"; a primary evaluation metric for this model.

Figure 2 shows the ANN model architecture with an input layer, all hidden layers, activation function, and the output layer.

Figure 3 represents the various steps taken to achieve the research objectives, i.e., to obtain high classification accuracy of the type 2 diabetes predictive model. Once the dataset is loaded, we create three duplicate copies of the data for three separate implementation approaches.

Let us consider that the dataset loaded is X .

$$X \longrightarrow X_1, X_2, X_3, \quad (1)$$

where (X_1, X_2, X_3) are the duplicate datasets created.

$$X_1 * \text{Median Value Imputation} \longrightarrow X_{1-\text{new}}, \quad (2)$$

$$X_2 * \text{Iterative Imputation} \longrightarrow X_{2-\text{new}}, \quad (3)$$

$$X_3 * K - \text{Nearest Neighbor Imputation} \longrightarrow X_{3-\text{new}}. \quad (4)$$

Equations (2), (3), and (4) show individual datasets, with X_1, X_2, X_3 being imputed. $X_{1-\text{new}}, X_{2-\text{new}}$ and $X_{3-\text{new}}$ are the datasets created, respectively, for datasets X_1, X_2, X_3 .

TABLE 2: Classification machine learning algorithms.

Algorithm	Short description
Logistic regression (LR)	It uses a logistic function to model the probability of a specific class
K-nearest neighbor (KNN)	It uses similarity of features/attributes to predict new data points based on how closely it resembles the points in the training set
Decision tree (DT)	It uses a tree structure in which each internal node represents an attribute, each branch represents the outcome of a condition set on the feature, and each leaf node represents a class label
Random forest (RF)	It uses multiple decision trees to train and output the class that is the mode of the classes from decision trees
Support vector machine (SVM)	It uses the kernel to transform data and identify an optimal boundary between the possible outputs
Gradient boosting (GB)	It uses a combination of many weak learning models together to create a robust predictive model
Light gradient boosting machine (LGBM)	It uses tree-based learning algorithms as a GB framework

FIGURE 2: ANN model architecture.

FIGURE 3: The sequence of methodology steps.

$$X_{1-new}, X_{2-new}, X_{3-new} * \text{Outlier Treatment} \longrightarrow X'_{1-new}, X'_{2-new}, X'_{3-new} \quad (5)$$

We process individual imputed datasets with outlier treatment in equation (5). X_{1-new} , X_{2-new} , and X_{3-new} are then split into 70% train set and 30% test set individually. Each dataset's training set is modeled using 10-fold cross-validation across the seven machine learning algorithms. We then use different metrics and accuracies to

determine model performance. Test set enables further evaluation of performance, leading to the conclusion of which imputed dataset provides the best accuracy.

$$X'_{1-new} * \text{Random Sampling} \longrightarrow X''_{1-new} \quad (6)$$

X_{1-new} , the outlier-treated dataset, is then exposed to random sampling to introduce some noise in the data, as shown in equation (6).

$$X''_{1-new} * \text{SMOTETomek} \longrightarrow X'''_{1-new}. \quad (7)$$

SMOTETomek link is used to create under-oversampling of the data. Equation (7) shows that X'''_{1-new} created is our current modified dataset. This step helps to convert an imbalanced dataset to a balanced dataset. X'''_{1-new} is then split into test-train data of ratio 70%–30%. The training data are then scaled for processing in ANN:

$$X'''_{1-new} * \text{Standard Scaling} \longrightarrow X'''_{1-new}. \quad (8)$$

ANN model's performance is judged based on different metrics. Prediction outcome is drawn out of the stable model for better understanding.

3.1. Data Description. The dataset used in this study is taken from the National Institute of Diabetes and Digestive and Kidney Diseases. As discussed by [23], this is a publicly available dataset at [30]. A detailed description of all attributes is given in Table 3. It contains several medical predictors: the independent variables and one dependent variable, the "Outcome."

Figure 4 illustrates the histogram of each attribute, explaining the distribution and skewness, and shows which attributes have missing values. Based on the histograms, the attributes age (A), diabetes pedigree function (B), insulin (C), pregnancies (D), and skin thickness (E) show that most of the data are on the left side of the histogram, which makes it skewed to the right. This means that when data are positively skewed and the mean is larger than the median.

Figure 5 shows histograms for BMI (A), blood pressure (B), and glucose (C) are symmetric; i.e., they have a normal distribution. The data are symmetrically distributed where most of the observations are around the central peak.

Most healthcare data contain the missing value, noisy and inconsistent data. This diabetes dataset is no exception where its attributes/features include missing values. Namely, glucose, blood pressure, skin thickness, insulin, and BMI have zero (0) values. We observed several candidates with a zero reading/value, for instance, five candidates with a zero glucose level, 35 candidates with a zero blood pressure, 227 candidates with a zero skin thickness, 374 candidates with a zero insulin level, and 11 candidates with a zero BMI. It reflects an inaccurate scenario; i.e., the zero values present in the data are incorrect and will probably lead to wrong results for diabetes prediction. Thus, we need an imputation of these particular zero values with something more meaningful for each parameter. As any machine learning model's performance heavily depends on the data, we performed necessary preprocessing techniques to handle the missing information.

Figure 6 represents the percentage of missing data related to a specific attribute. 48.69% of data are missing for insulin. 29.55% of information is missing for the skin thickness feature. 4.55% of data are missing for the blood pressure attribute. 1.43% of data are missing for BMI, and 0.65% of data are missing for the glucose attribute.

3.2. Data Preprocessing. In this research, we consider data preprocessing as a significant component. In the first phase of the proposed framework, we conducted an investigation on data imputation methods. After we determine the optimal method, we handled the data's outliers and the class imbalance problem.

3.2.1. Handling Missing Values. For the first phase, we applied three different imputation techniques to fill in the missing information. Method-I refers to imputation where we group by each missing variable median value under the outcome variable and assign those values to miss-values. Method-II refers to the iterative imputation technique. Method-III refers to the KNN imputation technique.

(1) Method-I. In order to fill missing values, we are grouping each missing variable's median values in accordance with the outcome class and assign those values to missing values. For example, the median value of glucose parameter is 107 for outcome class 0 and 140 for outcome class 1. All the missing values with 0 outcome class will be replaced with value 107 and all the missing values with 1 outcome class will be replaced with value 140.

The correlation matrix for this imputed dataset is shown in Figure 7. The highest positive correlation is 0.57, between skin thickness and BMI. The highest correlated factor for output class is glucose at 0.50, followed by insulin at 0.38 and BMI at 0.32.

(2) Method-II. Iterative Imputation. A sophisticated practice to find and replace missing values of a dataset is iterative imputation. This involves modeling to predict each feature as a function of all other features and keep repeating the process of estimating feature values multiple times. The repetition process allows the estimated values for other features to be used as input criteria for predicting missing values during subsequent iterations.

This is basically a data transformation method. Bayesian Ridge model is used as a function to all other input parameters by default. All parameters are considered in ascending order, filling the parameter with the least missing values first and then proceeding to the next parameters having the next minimum number of missing values. Different regression algorithms can be used to estimate the missing values for each parameter, although linear methods are often used for simplicity.

The correlation matrix for this imputation method is shown in Figure 8. The highest correlation factor is between skin thickness and BMI (0.71). The highest correlated factor for output class glucose is 0.50, followed by other factors such as insulin (0.36) and BMI (0.32).

(3) Method-III. KNN Imputation. Another good practice to find and replace missing values is through KNN imputation using the KNN machine learning model. In this method, the k-nearest neighbor algorithm is applied to impute a missing value by finding the samples in the training set "closest" to it and averaging these nearby points to fill in the value.

TABLE 3: Dataset attributes description.

No.	Attribute name	Attribute description	Attribute type
1	Pregnancies	Number of pregnancies of a woman	Numeric
2	Glucose (mg/dl)	Plasma glucose concentration for 2 hours in an oral glucose tolerance test	Numeric
3	Blood pressure (mmHg)	Diastolic blood pressure	Numeric
4	Skin thickness (mm)	Triceps skinfold thickness	Numeric
5	Insulin (μ U/ml)	2-hour serum insulin	Numeric
6	BMI (kg/m^2)	Body mass index (weight in $\text{kg}/(\text{height in m})^2$)	Continuous
7	Diabetes pedigree function	Diabetes pedigree function	Continuous
8	Age	Age in years	Numeric
9	Outcome	Target variable (0 or 1) 0 for negative, 1 for positive	Binary

Similar to the last method, KNN imputer is another data transformation method. KNN imputation requires numeric input values, which in our case is the ideal scenario. It is a lazy learner as it learns on the fly. It checks the nearest observation/neighbor to discover its path. Configuration of KNN imputation often requires selecting the distance measure and the number of nearest neighbors for each prediction. The default distance measure is Euclidean distance and the number of neighbors selected is 5.

The correlation matrix for this imputation method is shown in Figure 9. The highest correlation factor is between skin thickness and BMI (0.64). The highest correlated factor for output class glucose is 0.50, followed by other factors such as insulin (0.32) and BMI (0.31). Overall, all three imputed dataset correlation matrices show that the most correlated features with the target variable Outcome are glucose, insulin, and BMI.

3.2.2. Handling Outliers. The next preprocessing step was to handle outliers in all three imputed datasets. We set a threshold to decide on which data points constituted as outliers. We calculate the first quartile, third quartile, and interquartile range. We set the upper limit as third quartile $+1.5 \times \text{interquartile_range}$ of the variable, and the lower limit as quartile 1 $-1.5 \times \text{interquartile_range}$ of the variable.

$$\begin{aligned} \text{up_limit} &= \text{quartile3} + 1.5 \times \text{interquartile_range}, \\ \text{low_limit} &= \text{quartile1} - 1.5 \times \text{interquartile_range}. \end{aligned} \quad (9)$$

The results shown below illustrate the median value imputed dataset before and after outliers are handled.

Figure 10 shows outliers for blood pressure feature. We can see that there are multiple data points, which are out of bound. Blood pressure is treated with both upper limit and lower limit outlier correction. All these data points are corrected and shown in Figure 10(b) after the removal of outliers.

Figure 11 shows outliers for the diabetes pedigree function (DPF) feature. The DPF has multiple data points, which are beyond the upper limit bound of the boxplot. Upper limit outlier treatment is applied to DPF. Figure 11(b) shows the DPF feature after the removal of outliers.

Figure 12 shows outliers for the skin thickness feature. There seems to be only one data point that is beyond the upper limit of the boxplot. Upper limit outlier treatment is

applied to skin thickness. Figure 12(b) shows skin thickness features after the removal of outliers.

Figure 13 shows outliers for the insulin feature. Insulin has multiple data points, which are beyond the upper limit bound of the boxplot. Upper limit outlier treatment is applied to insulin. Figure 13(b) shows insulin features after the removal of outliers.

Figure 14 shows outliers for the pregnancy feature. We can see that there are three data points that are out of bounds before outliers are handled. These particular data points can cause erroneous outcomes if not treated properly. Pregnancy feature is treated with upper limit outlier treatment. These data points were corrected and shown in Figure 14(b) after the removal of outliers.

Figure 15 shows outliers for the BMI feature. We can see that there are multiple data points that are out of bound. Most of the data points are beyond the upper limit. There is only one data point that is beyond the lower limit. BMI is treated with both upper limit and lower limit outlier correction. Figure 15(b) shows BMI feature after the removal of outliers.

Figure 16 shows outliers for the glucose feature. We can see that there is only one data point that is out of bound. Only glucose is treated with lower limit outlier treatment as it has data points breaching the lower limit. This data point is corrected and shown in Figure 16(b) after the removal of outliers.

3.2.3. Handling Class Imbalance. The final preprocessing technique was to handle class imbalance and scale the imputed data. The imbalance in the number of samples that were diabetic compared to those that did not have diabetes was 268 by 500. Data sampling techniques reduced getting biased results in predictive performance and got more data samples for learning. According to [33], a combination of synthetic minority oversampling techniques (SMOTE) and random under- or over-sampling performs better than when SMOTE is used by itself. Therefore, in this research, we used a combination of random oversampling and SMOTETomek. Firstly, we used random oversampling to create more data samples and then transform them by SMOTETomek. SMOTE-Tomek combines over- and undersampling using SMOTE and Tomek links. SMOTE randomly increases minority class examples by synthesizing new minority instances

FIGURE 4: Dataset attribute distribution (right-skewed distribution). (a) Age, (b) diabetes pedigree function, (c) insulin, (d) pregnancies, and (e) skin thickness.

between existing minority instances. Tomek links are pair of examples that belong to different classes and are each other's nearest neighbors. After class imbalance has been

dealt with, we normalize data features using standard scaling and prepare them as input into the ANN machine learning model.

FIGURE 5: Dataset attribute distribution (normal distribution). (a) BMI, (b) blood pressure, and (c) glucose.

FIGURE 6: Percentage of missing data per attribute.

FIGURE 7: Median value imputed data correlation matrix.

FIGURE 8: Iterative imputed data correlation matrix.

FIGURE 9: KNN imputed data correlation matrix.

4. Results and Discussion

Implementing the proposed framework's first phase generates three datasets based on each imputation method. We applied seven classification algorithms on each dataset, and the results were slightly different as each algorithm's working criteria are different. We used 10-fold cross-validation to evaluate the training set performance. The primary evaluation metric used throughout this study is accuracy:

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{all samples}}, \quad (10)$$

The results of all classifiers were evaluated based on different metrics as follows:

- (a) True positive (TP): subject with disease and model correctly predicts that
- (b) True negative (TN): subject with no disease and model correctly predicts no disease
- (c) False positive (FP): subject with no disease but model predicts incorrectly that the subject is diabetic
- (d) False negative (FN): subject with the disease, but the model predicts incorrectly that the subject is not diabetic

For the evaluation of the accuracy of the classification model, the values from the confusion matrix are used. Evaluation metrics, i.e., F1-score, precision, and recall, are calculated.

- (a) F1-score:

$$\text{F1 - score} = \frac{2TP}{2TP + FP + FN}. \quad (11)$$

- (b) Precision:

$$\text{precision} = \frac{TP}{TP + FP}. \quad (12)$$

- (c) Sensitivity (recall):

$$\text{recall} = \frac{TP}{TP + FN}. \quad (13)$$

- (d) Specificity:

$$\text{specificity} = \frac{TN}{TN + FP}. \quad (14)$$

4.1. Data Imputation Results. Various evaluation metrics are used to assess the performance of each method of imputation technique applied via 10-fold cross-validation for seven machine learning classifiers.

4.1.1. Prediction Accuracy. Accuracies achieved for each method of imputation technique applied via 10-fold cross-validation using different classifiers are discussed as follows.

FIGURE 10: Handling outliers for blood pressure feature. Before (a) and after (b).

Figure 17 shows that GB is the best performing classifier. GB classifier achieves an accuracy of 91.06%, followed by the LGBM classifier, which achieves an accuracy of 90.69%. LR, KNN, DT, and RF classifiers all show good accuracy but not as good as GB. SVM performs the weakest.

Figure 18 shows the accuracy performance measure of all the classifier algorithms applied to Method-II imputation. LR and RF classifiers have almost similar accuracies, with 78.585% and 78.588%, respectively. GB classifier also performs well with 78.03% accuracy. Like in Method-I, SVM classifier performs the weakest.

Figure 19 shows the accuracy performance measure of all the classifier algorithms applied to Method-III imputation. LR performs the best and achieves an accuracy of 78.21%, followed by GB classifier scoring accuracy of 77.65%. KNN and DT classifiers have achieved the same accuracies of 71.89%. SVM performs the weakest in this method as well.

4.1.2. Precision, Recall, and F1-Score. The above figures conclude that Method-I imputation resulted in the highest accuracy. To further validate that Method-I is indeed the optimal data imputation technique among the three methods, performance was observed for other evaluation metrics as well, namely F1-score, Precision, and Recall. As our data is an imbalanced dataset, we are considering a weighted average for the scoring.

Table 4 describes the performance of each classification model evaluated with the help of mentioned metrics. These scores were achieved by applying 10-fold cross-validation. The precision score is similar for both GB and LGBM classifiers, whereas for recall and precision, GB has a better score compared to LGBM.

Table 5 describes the performance of each classification model evaluated with the help of the mentioned metrics for imputation technique Method-II. LGBM classifier has precision, recall, F1-score at 0.74 for each. This is the best, and as per this score, we can see that this model outperforms GB classifier marginally. SVM has the least scores for the individual metrics.

Table 6 describes the performance of each classification model evaluated, with the help of the mentioned metrics for imputation technique Method-III. LGBM and LR have precision, recall, F1-score at 0.73 for each. SVM has the least scores for the individual metrics.

4.1.3. Area under ROC (Receiver Operating Characteristic) Curve. Furthermore, we also considered Receiver Operating Characteristics (ROC) as part of the performance evaluation. We know that the greater the area under the curve, the better the model. An ideal model should be having a TP rate = 1 and a FP rate = 0. So, any curve which is more aligned to the Y-axis is more fit deemed to perform better.

Figure 20 shows the AUC that we have achieved for the Method-I imputation technique. We can see that a solid red line with “o” marker is our curve for GB classifier, strictly followed by the green dashed line, LGBM classifier.

Figure 21 shows the AUC that we have achieved for the Method-II imputation technique. The green dashed line has more area under its curve than that of other models. Thus, it indicates that the LGBM classifier has a better model performance for the Method-II imputation technique.

Figure 22 shows the AUC that we have achieved for the Method-III imputation technique. The green dashed line has more area under its curve than that of other models. Thus, it indicates that LGBM classifier has a better model performance for Method-III imputation technique.

4.1.4. Summary. Table 7 summarizes the findings of the first phase results, presenting the highest numbers of evaluation metrics achieved by each method.

All methods result in favorable performance with more than 60% across all evaluation metrics. LGBM classifier shows up recurrently as the best performing model for all methods. Among the three imputation methods, Method-I can be seen outperforming the other two methods by a significant margin across all the evaluation metrics, while Method-II and Method-III have similar numbers in performance. The correlation matrix of the imputed data presented in Section 3.2.1 offers a

FIGURE 11: Handling outliers for DPF feature. Before (a) and after (b).

FIGURE 12: Handling outliers for skin thickness feature. Before (a) and after (b).

possible reason for this difference in performance. Method-II and Method-III imputation generated data that had a higher correlation between its variables than Method-I. Method-II and Method-III imputed data may be suffering from the problem of multicollinearity. When predictor variables are significantly dependent or associated with each other, it can hinder the model's ability to learn the contribution of each predictor variable individually as a change in one would cause a change in the other. A possible reason behind Method-I's high performance can thus be attributed to a low chance of multicollinearity in its imputed dataset.

4.2. ANN Results. The second phase consisted of implementing an ANN model on the highest performing imputed data, i.e., the Method-I imputed dataset. After handling the class imbalance and scaling the training dataset, an accuracy of 0.9817 was achieved. Upon evaluation of the test dataset, it produced an accuracy of 0.9809. Table 8 presents the classification report with relevant evaluation metrics.

Precision, recall, and F1-score achieved for outcome "0" are 0.99, 0.97, and 0.98, respectively. Precision, recall, and F1-score achieved for outcome "1" are 0.97, 0.99, and 0.98, respectively. We can also see that the macro average and the weighted average for precision, recall, and F1-score are 0.98. Even though hyperparameter tuning was not conducted for this ANN model, high-performance results were achieved. In the comparison of area under ROC curves, the ANN model outperformed the other seven machine learning algorithms.

Figure 23 shows an AUC curve plot. The figure is for comparative study among all the classification techniques applied. ANN performs thirteen times better than GB, which had performed highest in the data imputation phase.

Key findings of this research include the following:

- (i) Method-I: missing values imputed with median values showed better results than the KNN imputation technique and the iterative imputation

FIGURE 13: Handling outliers for insulin feature. Before (a) and after (b).

FIGURE 14: Handling outliers for the pregnancy feature. Before (a) and after (b).

FIGURE 15: Handling outliers for the BMI feature. Before (a) and after (b).

FIGURE 16: Handling outliers for glucose feature. Before (a) and after (b).

FIGURE 17: Model performance accuracy achieved for Method-I.

FIGURE 18: Model performance accuracy achieved for Method-II.

technique. A possible reason for this may be due to lower multicollinearity presence in Method-I imputed dataset than that of the other two methods.

(ii) SMOTETomek, along with the random sampling technique, was applied to the dataset created by the

Method-I imputation technique. This combined method for treating imbalanced data was better than using one technique individually.

(iii) ANN was implemented after balancing the dataset, and it produced an accuracy of 0.9809. This is by far

FIGURE 19: Model performance accuracy achieved for Method-III.

TABLE 4: Evaluation metrics for Method-I.

Models	Precision	Recall	Specificity	F1-score
LR	0.76	0.76	0.63	0.76
KNN	0.85	0.85	0.74	0.85
CART	0.83	0.83	0.73	0.83
RF	0.85	0.85	0.81	0.85
SVM	0.42	0.65	0.00	0.51
LGBM	0.86	0.86	0.77	0.86
GB	0.86	0.87	0.79	0.87

TABLE 5: Evaluation metrics for Method-II.

Models	Precision	Recall	Specificity	F1-score
LR	0.73	0.73	0.55	0.73
KNN	0.71	0.71	0.58	0.71
CART	0.67	0.68	0.51	0.67
RF	0.73	0.73	0.55	0.73
SVM	0.42	0.65	0.00	0.51
GB	0.72	0.71	0.60	0.72
LGBM	0.74	0.74	0.64	0.74

TABLE 6: Evaluation metrics for Method-III.

Models	Precision	Recall	Specificity	F1-score
LR	0.73	0.73	0.55	0.73
KNN	0.69	0.69	0.57	0.69
CART	0.68	0.67	0.53	0.67
RF	0.72	0.72	0.54	0.72
SVM	0.42	0.65	0.00	0.51
GB	0.69	0.7	0.55	0.7
LGBM	0.73	0.73	0.62	0.73

the best accuracy achieved in terms of classification compared to past studies done with the same dataset.

Apart from this, some interesting facts about the attributes' relationship with type 2 diabetes were found because of data mining, as follows:

FIGURE 20: AUC analysis for Method-I.

FIGURE 21: AUC analysis for Method-II.

- (i) As per the correlation matrix shown in Section 3, BMI is highly correlated with diabetic patients. So, it can be concluded that obesity has a high chance of causing type 2 diabetes.
- (ii) Insulin resistance has been identified as a major risk factor in the pathogenesis of type 2 diabetes and has

been thoroughly studied in the Arizona Pima. Insulin resistance appears to increase incrementally according to BMI levels primarily and body fat levels secondarily. As shown in the data, that obesity is causing high BMI, which is leading to chances of type 2 diabetes. Further, DPF, which indicates

FIGURE 22: AUC analysis for Method-III.

TABLE 7: Data imputation results summary.

Method	Accuracy		Precision		Recall		Specificity		F1-score		AUC	
	Model	Best	Model	Best	Model	Best	Model	Best	Model	Best	Model	Best
I	LGBM	0.91	GB	0.86	GB	0.87	GB	0.79	GB	0.87	LGBM	0.84
II	LR	0.78	LGBM	0.74	LGBM	0.74	LGBM	0.65	LGBM	0.74	LGBM	0.78
III	LR	0.78	LGBM	0.73	LGBM	0.73	LGBM	0.62	LGBM	0.73	LGBM	0.70

TABLE 8: Classification report for ANN model.

Model	Precision	Recall	Specificity	F1-score
ANN	0.98	0.98	0.99	0.98

FIGURE 23: AUC analysis for all models applied using Method-I imputation.

diabetes history in relatives and genetic relationship, shows high chances of diabetes.

5. Conclusion

With advancements in technology and computing power, intelligent machine systems can be utilized to provide a more efficient alternative to traditional manual analysis for medical decision support. The current solutions include classification models that are highly inaccurate, but they do not consider some essential data preprocessing steps that can positively impact overall accuracy achieved. This research adopted a framework whereby the first phase focused on identifying the optimal handling of missing data. The application of three different data imputation techniques was investigated and observed for its impact on classification model performance. In the second phase, the best performing imputation data were balanced using a combination of SMOTETomek and random oversampling and modeled using ANN.

There is a certain assumption that we have considered during data preprocessing. The upper boundary and lower boundary set for treating outliers were based on generic assumptions. ANN performs well with a bigger dataset. But for our case, the dataset was small. Thus, random oversampling applied to imputed data for building ANN is assumed to be close to real data. The number of nodes for starting and hidden layers of ANN is based on multiple trials, and the current network is assumed to work the best. Eight nodes were assumed for starting layer as the target variable is dependent on 8 variables. The next hidden layer had double the number of starting nodes, 16. For the consecutive following hidden layers, every time, the number of nodes was reduced down by 4. The final layer of the network had 1 node as a standard process.

There are several limitations to this study. The data used in this study are of a specific region, restricted to the population of Native Americans. All patients considered in this study are females of age 21 and above. Future works should include data across different geographic regions and of various origins. It would be interesting to see how the model behaves with a mixed group of males and females. From a modeling point of view, hyperparameter optimization of ANN can also be adapted to study the outcome/accuracy.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful to Taif University Researchers Supporting Project (no. TURSP-2020/215), Taif University, Taif, Saudi Arabia. This work was also supported by the

Faculty of Computer Science and Information Technology, the University of Malaya under Postgraduate Research Grant (PG035-2016A).

References

- [1] International Diabetes Federation, IDF SEA Members., 2020, <https://idf.org/our-network/regions-members/south-east-asia/members/94-india.html>.
- [2] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, "Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus," in *Digital Information Processing and Communications*, pp. 537–545, Springer, Berlin, Germany, 2011.
- [3] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8102–8108, 2010.
- [4] J. C. Pickup, "Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes," *Diabetes Care*, vol. 27, no. 3, pp. 813–823, 2004.
- [5] D. W. Guthrie and R. A. Guthrie, "Nursing management of diabetes mellitus: a guide to the pattern approach," *Home Healthcare Nurse*, vol. 23, no. 2, p. 119, 2005.
- [6] E. I. Mohamed, R. Linder, G. Perriello, N. Di Daniele, S. J. Pöppel, and A. De Lorenzo, "Predicting Type 2 diabetes using an electronic nose-based artificial neural network analysis," *Diabetes, Nutrition & Metabolism*, vol. 15, no. 4, p. 215, 2002.
- [7] Mayo Clinic Staff, Prediabetes—symptoms and causes, 2020, <https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278>.
- [8] N. C. Sharma, Government survey found 11.8% prevalence of diabetes in India, 2019, <https://www.livemint.com/science/health/government-survey-found-11-8-prevalence-of-diabetes-in-india-11570702665713.html>.
- [9] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes—estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, 2004.
- [10] World Health Organization, *Global Report on Diabetes*, WHO, Geneva, Switzerland, 2016, <https://www.who.int/publications/i/item/9789241565257>.
- [11] A. A. AlJarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Proceedings of the 2011 International Conference on Innovations in Information Technology*, pp. 303–307, Abu Dhabi, UAE, April 2011.
- [12] U.S. Department of Health and Human Services, *Diabetes: A National Plan for Action. The Importance of Early Diabetes Detection*, U.S. Department of Health and Human Services, Washington, D.C, USA, 2004, <https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection>.
- [13] M. Komi, J. Li, Y. Zhai, and Z. Xianguo, "Application of data mining methods in diabetes prediction," in *Proceedings of the International Conference on Image, Vision and Computing Application*, pp. 1006–1010, IEEE, Chengdu, China, June 2017.
- [14] Z. Soltani and A. Jafarian, "A new artificial neural networks approach for diagnosing diabetes disease type II," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 89–94, 2016.
- [15] L. AlThunayan, N. AlSahdi, and L. Syed, "Comparative analysis of different classification algorithms for prediction of diabetes disease," in *Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing*, pp. 1–6, New York, NY, USA, 2017.

- [16] E. Guldogan, Z. Tunc, A. Acet, and C. Colak, "Performance evaluation of different artificial neural network models in the classification of type 2 diabetes mellitus," *The Journal of Cognitive Systems*, vol. 5, no. 1, pp. 5–9, 2020.
- [17] T. M. Alam, M. A. Iqbal, Y. Ali et al., "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, 2019.
- [18] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019.
- [19] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *Proceedings of the 24th International Conference on Automation and Computing*, pp. 1–6, IEEE, Newcastle Upon Tyne, UK, September 2018.
- [20] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques," in *Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 414–418, IEEE, Tirunelveli, India, May 2018.
- [21] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.
- [22] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.
- [23] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in *Proceedings of the International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1–5, IEEE, Lagos, Nigeria, October 2017.
- [24] O. Geman, I. Chiuchisan, and R. Todorean, "Application of adaptive neuro-fuzzy inference system for diabetes classification and prediction," in *Proceedings of the 2017 E-Health and Bioengineering Conference, EHB*, pp. 639–642, IEEE, Sinaia, Romania, June 2017.
- [25] S. Ramesh, R. D. Caytiles, and N. C. S. Iyenga, "A deep learning approach to identify diabetes," *Advanced Science and Technology Letters*, vol. 145, pp. 44–49, 2017.
- [26] A. Marcano-Cedeño, J. Torres, and D. Andina, "A prediction model to diabetes using Artificial Metaplasticity," in *New Challenges on Bioinspired Applications*, pp. 418–425, Springer, Berlin, Germany, 2011.
- [27] S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 117–126, 2010.
- [28] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 82–89, 2008.
- [29] J. Han, J. C. Rodriguze, and M. Beheshti, "Diabetes data analysis and prediction model discovery using RapidMiner," in *Proceedings of the 2008 Second International Conference on Future Generation Communication and Networking*, pp. 96–99, IEEE, Hainan, China, December 2008.
- [30] UCI Machine Learning, Pima indians diabetes database, 2016, <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.
- [31] K. Lakshminarayan, S. A. Harp, R. P. Goldman, and T. Samad, "Imputation of missing data using machine learning techniques," *KDD*, pp. 140–145, August 1996.
- [32] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1625–1657, 2007.
- [33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.