Hindawi
Mobile Information Systems
Volume 2022, Article ID 6521532, 16 pages
https://doi.org/10.1155/2022/6521532

Research Article

Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques

Zaigham Mushtaq,[1] Muhammad Farhan Ramzan,[1] Sikandar Ali,[2] Samad Baseer,[3] Ali Samad,[1] and Mujtaba Husnain[1]

[1] *Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*
[2] *Department of Information Technology, The University of Haripur, Haripur 22620, Khyber Pakhtunkhwa, Pakistan*
[3] *Department of Computer System Engineering, University of Engineering and Technology, Peshawar 25000, Khyber Pakhtunkhwa, Pakistan*

Correspondence should be addressed to Sikandar Ali; sikandar@cup.edu.cn

Received 2 January 2022; Revised 23 January 2022; Accepted 14 February 2022; Published 19 March 2022

Academic Editor: Ahmed Farouk

Diabetes mellitus is a hyperglycemia-like chronic condition that is a troublesome disease. It is estimated that, according to the growing morbidity, by 2040, the world will cross 642 million diabetic patients. This means that each one of the ten adults will be diabetes-affected. Diabetes can also lead to other illnesses such as heart attacks, kidney damage, and even blindness. The prediction of diabetes in advance motivates us to develop a machine learning-based model. A dataset was obtained from the online repository for this work. The obtained dataset was imbalanced. An imbalanced dataset presents a challenge that is needed to be balanced for prediction using multiple machine learning like Tomek and SMOTE. These techniques remove necessary outliers that are incomplete in the provided dataset. These outliers are also managed using the IQR method. Additionally, this research employed a two-stage model selection methodology. In the first stage, logistic regression, Support Vector Machine, k-nearest neighbors, gradient boost, Naive Bayes, and Random Forests were applied to determine the efficiency of prediction based on patients' preconditioning. At this stage, Random Forest was found to be the best with an accuracy of 80.7% after applying SMOTE oversampling technique to balance the dataset. In the second stage, three better-performing models were used by utilizing a voting algorithm. The results were encouraging, and the model obtained 82.0% accuracy with the default dataset and 81.7% accuracy with the balanced dataset. Naive Bayes Theorem, Gradient Boosting Classifier, and Random Forest were used as inputs to the voting algorithm.

1. Introduction

Imbalanced data detection is still one of the main difficulties in the area after several years of research into machine learning. The basic learning algorithm assumes that classes are roughly balanced within the dataset of the training. Learning output metrics most sometimes presume that the classes inside the dataset are of similar significance. Unfortunately, in real-life situations, balanced datasets are uncommon, and the underrepresented class usually has higher misclassification. Consider the differential designation of the United Kingdom (UK) people as having or not having diabetes, for example. The latest figures suggest that 4.6 percent of the population has diabetes, leaving 95.4 percent of cases with nondiabetes. A prediction model that correctly classifies all the majority classes and incorrectly classifies all the minority classes will have a very high yet deceptive 95.4 percent accuracy. The cost of misclassifying individuals with diabetes will lead to severe repercussions [1–6].

Another main challenge in machine learning is the classification of outliers. This is because clusters of data seldom follow a consistent trend. Such data samples can sometimes be different from other data of the same class and therefore far away from the data mass of that class.

Machine learning can learn from past results to make constructive choices on current cases which are previously invisible. In this research, machine-learning classifiers are used for classification. Machine-learning classifiers are trained with the dataset to predict diabetes itself. The classifiers which are used for this research are logistic regression, Support Vector Machine, k-nearest neighbors, gradient boost, Naive Bayes, Random Forests, and voting classifier [1].

Diabetes is a hereditary illness that develops while the pancreas does not contain enough insulin or when the body does not produce enough insulin. Insulin is a hormone that is named controls sugar in the blood. Hyperglycemia or high blood sugar levels, a typical consequence of uncontrolled and over uncontrolled diabetes time, contribute to substantial damage to many of the body's structures, nerves, and blood vessels. Health care services are built solely to address the demands of a growing global population. Citizens around the world are affected by various kinds of deadliest diseases [7–11].

Diabetes is a significant cause of blindness, kidney failure, heart problems, etc., among the various types of widely available diseases. Systems for the control of health services for multiple illnesses and symptoms are available around the world. There have been major advances in health care services due to the rapid advancement in the areas of information and communication technology. Various machine-learning algorithms are being proposed that simplify the health systems' operating model and improve the accuracy of disease prediction. Diabetes, particularly diabetes type 2, is one of the most common chronic diseases in the United States and affects millions of people's health. To identify risk factors for type 2 diabetes, we sought to develop predictive models that could help promote early detection and intervention and minimize medical expenses [12].

Diabetes is a disorder in which the levels of blood glucose or blood sugar are too high. The glucose comes from the carbohydrates that you drink. Diabetes is a condition that makes it possible for glucose to enter the cells to offer them energy. The graphical representation for the explanation of abnormality is shown in Figure 1. The dotted line shows a threshold limit in Figure 1. If the moving line crossed a threshold limit, then it is an abnormal body or disease. Below the threshold limit, it represents a normal body. On the left side, the majority of clusters of points showed a regular pattern. In contrast, one outlier point showed irregularity in the body tissue [13].

In today's world, diabetes is a major health challenge world. It is a group of syndromes that results in too much sugar in the blood. It is a protracted condition that affects the way the body mechanizes blood sugar. Prevention and prediction of diabetes mellitus are increasingly gaining interest in medical sciences. Diabetes mellitus diseases are critical, and numerous people are suffering from this disease. Diabetes is a public, long-lasting disease. Diagnosis of diabetes at a primary stage is a challenging task. Hence, an automated and accurate system is required for diabetes class and disease prediction. Following are the objectives in this research:

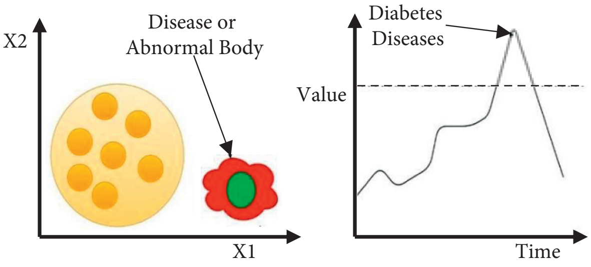

Figure 1: Basic description of diabetes in the human body.

Data cleaning and data preprocessing, such as data balancing and handling outliers on datasets have been implemented.

Implementation of Machine-Learning Predictive system for diabetes diagnosis at an early stage.

To implement ensemble learning techniques, i.e., voting classifier, to enhance traditional machine-learning models' results.

2. Literature Review

This section briefly explains the previous literature studies related to diabetic predictive systems based on machine-learning techniques.

Nnamoko and Korkontzelos [14] proposed a new technique for prediction. The goal is to match the dataset of the training while managing the impact of outliers. The experiments show that SMOTE is empowered by such selective oversampling, eventually leading to enhanced classification efficiency.

Mujumdar and Vaidehi [15] predict a new approach for diabetes with machine learning. It proposed an improved diabetes classification diabetes model that involves a few external factors responsible for diagnosing diabetes along with natural factors such as glucose, BMI, age, and insulin. In comparison to the current data collection, the precision of the classification is improved. In addition, a diabetes prediction pipeline model was created to improve classification accuracy.

Tigga and Garg [16] proposed a new method for the prediction of type 2 diabetes. 952 cases were obtained via an online and offline survey to experiment, featuring 18 questions related to fitness, lifestyle, and family history. The same methods were also added to the database for PIMA Indian diabetes. On both datasets, the output of the Random Forest Classifier is the most reliable.

Zhu et al. [17] proposed a new method for diabetes prediction. K-means are easy and can be used for a wide range of types of details. The initial locations of the cluster centers that decide the outcome of the cluster are very sensitive. Using patient electronic health records information, the model is seen to be useful for automatically forecasting diabetes.

Mahboob Alam et al. [18] proposed a new technique for the early prediction of diabetes. Diabetes is predicted using major attributes in this research paper, and the interaction of the various features is also defined. The results suggest a clear correlation between diabetes and the index of body mass (BMI) and the level of glucose.

Cahn et al. [19] proposed a new prediction of diabetes progression. They identify those at a high risk of prediabetes progression a priori. Diabetes may allow for tailor-made programmed intervention while avoiding the burden of preventing and caring for people with low risk. They investigated the possibility of improving the estimated incident diabetes using the electronic medical records for patient information using a machine-learning model.

Balcázar et al. [20] proposed a new method for diabetes management where the mission is to analyze the current state of ML in different areas of diabetes treatment and to identify crPIMAal obstacles to be addressed to exploit ML to its maximum potential.

Hasan et al. [21] proposed a new diabetes prediction technique. A robust diabetes prediction approach has been presented in this literature. Application is made of machine-learning (ML) classifications (K-nearest neighbor, random, decision-making trees, and forests). The predictions should also be improved by Forest, AdaBoost, Naive Bayes, XGBoost), and Multilayer Perceptron (MLP). Their approach to diabetes prediction outweighs the other systems.

Dinh et al. [22] suggested an approach based on data to predict diabetes. They suggested that survey-based machine-learning models provide patients at risk of diabetes and cardiovascular disorders with an automatic recognition mechanism. They also consider significant contributors to this prospect, and their influence on electronic health records can be further investigated.

Kaur and Kumari [23] proposed a new approach to predictive modeling for diabetes. In a recent study, the compilation of information on PIMA Indian diabetes uses machine-learning techniques to recognize trends and patterns of risk factors using the tool for *R* manipulation. To identify patients in diabetes and nondiabetic conditions, five separate predictive models were created and analyzed using the *R* data managing framework. For this reason, we used controlled machine learn algorithms, i.e., the vector machine support kernel linear (SVM-linear), the support for the radius base function (RBF), the neighboring k-nearest (k-NN), and the multifactor dimension reduction (MDR).

Ahmad et al. [24] proposed a new method Interpretable Machine Learning in Healthcare. We cover a range of applications in which healthcare needs to view competent machine-learning models and how they can be applied. In addition, we analyze the environment of recent innovations to resolve the complexities of healthcare model interpretation capacity and explain how to select the correct machine-learning interpretation algorithm for a given healthcare problem.

Choudhury and Gupta [25] proposed a new diabetes diagnostic technique. The key emphasis of this paper is the study of diabetes detection by machine-learning techniques. In comparison, there is PIMA India. In profound learning strategies such as artificial neural networks, decision-makers, random forestry, naive bays, k-nearest neighbors, vector support machines, and logistic regression, diabetic data sets are used. The findings have their benefits and drawbacks discussed.

Rodriguez-Romero et al. [26] suggested a new technique to predict type 2 diabetes nephropathy. Our findings showed that the best efficiency among the evaluated algorithms was seen in the Random Forest and Simple Regression techniques. A DN forecast is defined as DN predictors GFR, urinary creatinine, urinary albumin, potassium, cholesterol, low-density lipoproteins, and urinary albumin. The baseline values for early predictors in Month 4 were GFR, systolic blood pressure, plasma glucose fasting (FPG), and potassium fasting. The late development predictors were per year GFR, FPG, and triglycerides improvements. In conclusion, DN predictive factors were successfully identified in patients with T2DM by ML-based methods.

Faruque et al. [27] suggested a modern technique for diabetes prediction. Four popular machine-learning algorithms (Support Vector Machine, Naive Bayes, K-Nearest Neighbor, KNN, and C4.5 Decision Tree) predicted adult population data in this work, which are specifics of the population (DT). The author's experimental findings indicate that the decision tree C4.5 is more accurate than other machine-learning techniques.

Islam Ayon et al. [28] proposed a new diabetes prediction tool. By applying its properties in a 5- and 10-fold cross-validation mode, we propose a diabetes diagnostic technique using a deep neural network. The PIMA Indian diabetes (PID) data set is taken from the PIMA learning machine repository servers.

Liu et al. [29] suggested a recent gestational diabetes forecast. This research was designed to build a machine-based learning prediction model for early pregnancy gestational diabetes mellitus (GDM) in Chinese women.

Bettini et al. [30] proposed a new method for type 2 diabetes predictions. In this research, with precise ML classifiers, we performed experiments to predict diabetes in PIMA Indian women. The current research on PIDD, using cross-validation techniques, aimed to define an optimal ML model. The AUC was 0.83 for LR, 0.82 for RF, and 0.81 for NB. All three have been listed as the best models for predicting whether a patient is diabetic.

Makino et al. [31] proposed a new method to diagnose diabetic kidney disease. We also created a new predictive model for diabetic kidney disease with AI, natural language processing, and longitudinal data processing with Big Data Machine Learning. With 71 percent precision, AI could forecast DKD aggravation. AI's latest model may define DKD development and can lead to more efficient and reliable intervention.

Perveen et al. [32] suggested a new approach to diabetes prediction. Data from 172,168 patients in primary care were used to assess diabetes risk in a person using HMM for eight years using the Electronic Medical Record (EMR). Our analysis sample for 911 individuals with risk factors and monitoring data is based on 86 of the area under the recipient operating trait curve (AROC).

Segar et al. [33] proposed new diabetes prediction techniques. A new, master-learning risk score that combines readily available clinical, laboratory, and electrocardiographic variables was developed and validated for HF risk prediction among ambulatory patients with T2DM.

Sneha and Gangil [34] proposed a method for the analysis of diabetes mellitus. The goal of the proposed approach is to use predictive analysis to select the characteristics present in the early detection of diabetes Miletus.

Sonar and JayaMalini [35] proposed a new technique for diabetes prediction. They create a method that could predict the patient's level of diabetic risk with greater specificity. The creation of models is based on the decision tree, ANN, Naive Bayes, and SVM algorithms for categorization. For the decision tree, the models have 85 percent accuracy, 77 percent for Naive Bayes, and 77.3 percent for Support Vector Machine. Outcomes reflect the major reliability of the procedures.

Yuvaraj and SriPreethaa [36] proposed a new technique for diabetes prediction. A branch of Artificial Intelligence, Machine Learning, is used to evaluate and construct a model for diabetes prediction. To predict the possibility of diabetes, a data study from PIMA Indians was obtained. With 92 percent precision, this model is perfect for forecasting diabetes.

Jeevan Nagendra Kumar et al. [37] proposed a new technique for diabetes prediction. They intend to apply the resampling technique of bootstrapping to improve accuracy and then apply Naive Bayes, Decision Trees, and (KNN) and compare their efficiency.

Vigneswari et al. [38] proposed a new technique for diabetes prediction. They compare the efficiency of the classifiers of the machine-learning tree in predicting diabetes mellitus. The Logistic Model Tree (LMT) obtained a 79.31 percent higher accuracy and a 0.739 True Positive Score (TPR). Table 1 shows the comparative analysis of previous research based on performance parameters.

Table 1 shows the comparative analysis of previous research based on performance parameters.

3. Methodology

Here is how we did it. Feature engineering, model creation (machine-learning algorithm based), and performance evaluation are explained for acquired datasets. Figure 2 shows the study's operating flow.

The following are the details of the block diagram:

(i) PIMA diabetes data has been used to generate the CSV file

(ii) Data balance and handling outliers have been used in data preprocessing

(iii) The results have been validated using cross-validation

(iv) Machine-learning models have been applied

(v) Finally, ensemble learning approaches have been introduced after the best classification models were selected based on accuracy

(vi) Ensemble learning is used to create a hybrid model using the voting classifier

3.1. Dataset. Data from the PIMA diabetes prediction model can be downloaded for free [1]. The National Institute of Diabetes and Digestive and Kidney Disease provided this information. The dataset contains one (dependent) variable and several (independent) factors related to medical prediction.

3.1.1. Dataset Descriptions. Data concerning PIMA datasets are depicted in Table 2. Patient symptoms, such as the number of pregnancies, blood sugar and insulin levels, and thickness of the patient's skin, are all included in the dataset, as is their BMI, their family history of diabetes, their age, and whether or not they would become diabetic in the future.

3.1.2. Dataset Attribute Statistics. The histogram for each attribute in the dataset can be seen in Figure 3. It provides statistics on dataset parameters such as age, body mass index (BMI), blood pressure, diabetes pedigree function, blood glucose, insulin, result, pregnancies, and skin thickness (thickness of the skin).

For each target, Figure 4 shows the total count (0 or 1). If the number is 0, the patient does not have diabetes; if the number is 1, they do. While 65.1 percent of the population will not get diabetes, 34.9 percent of patients are at risk.

3.2. Raw Data Processing. For analyzing and testing the proposed method, the PIMA diabetes dataset has been utilized. There is a wide range of disorders in PIMA's database. For preprocessing and extraction of features, the raw data is converted into a CSV file format [40–45]. Diabetic law is defined by the methodology outlined in this article. The maximal normal patient is distinct from the patient.

3.2.1. Data Cleaning. The information obtained from [1] was unprocessed. Because of this, several strategies including removing duplicates and null values have been used to clean the data.

3.3. Data Preprocessing. This technique is used in data mining to transform raw data into a format that can be understood. Data in the real world is often partial, mismatched, and/or missing in some behaviors and trends. Preprocessing the data can be done in a variety of ways.

3.3.1. Data Balancing. Prediction modeling is made more difficult by classifications that are not evenly distributed. Classification machine-learning algorithms often start with the same amount of instances for each class they are trying to learn. This results in inaccurate models, especially for minorities. This is a problem since the minority group is more significant and, as a result, more susceptible to classification errors than the majority group. So we have eliminated the outliers from this investigation's data set. Resampling approaches have evolved significantly as a result of this research. For example, we can aggregate the majority of class records and do undersampling by extracting records from each cluster. As an alternative to making exact replicas of minority class data, we can include modest adjustments to these versions through oversampling. Figures 5 and 6 depict

Table 1: Comparative analysis of previous research based on performance parameters.

Ref. Paper	Techniques	Datasets	Outcome	Accuracy
[39]	SVM, KNN	PIMA	Prediction of diabetes mellitus	78%, 79%
[37]	SVM, KNN, RF	PIMA	Prediction of diabetes mellitus	76%, 75%, 76%
[23]	ANN, RF, DT	Type-2 diabetes datasets	Prediction of diabetes mellitus and type 2	80%, 77%, 76%
[29]	SVM	PIMA	Prediction of diabetes mellitus	80%
[21]	SVM, KNN, RF, NB, DT, ANN	Type-2 diabetes datasets	Prediction of diabetes mellitus and type 2	78%, 78%, 78%, 79%, 79%, 78%
[27]	SVM, KNN, RF, NB, DT	Type-2 diabetes datasets	Prediction of diabetes mellitus and type 2	76%, 75%, 75%, 75%, 75%

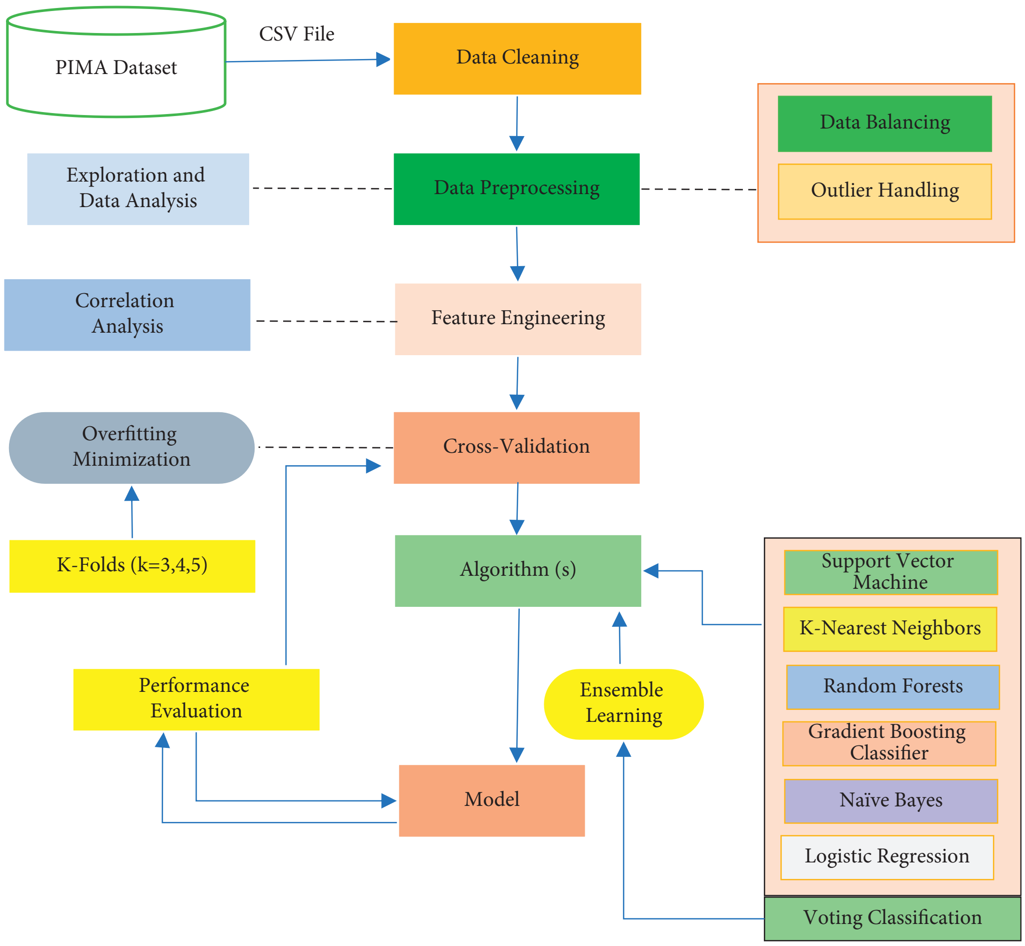

Figure 2: Block diagram.

Table 2: Dataset information.

	Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Diabetes pedigree function	Age	Outcome
Count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
Mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
Std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
Max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

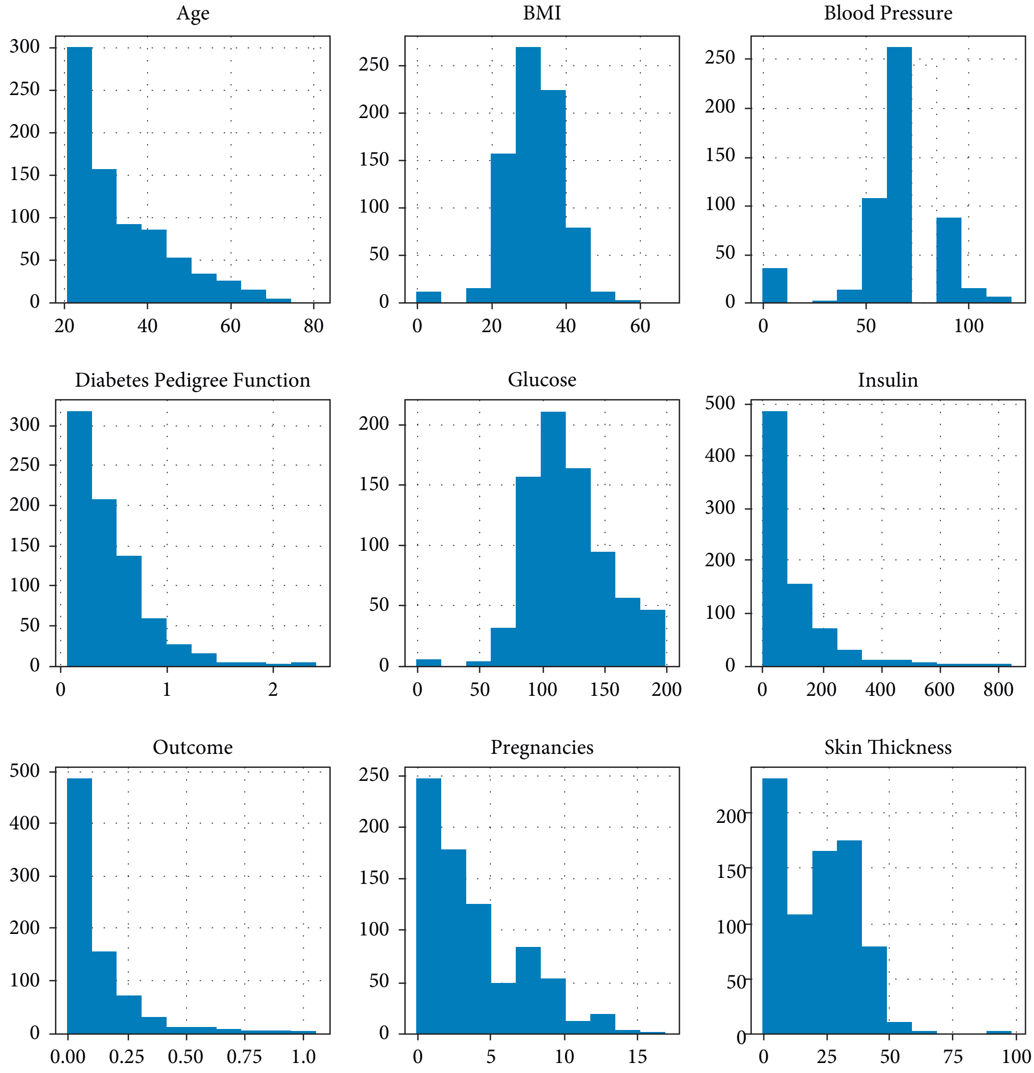

Figure 3: Histogram of each attribute.

the results when undersampling and oversampling approaches are applied. Figure 7 shows the relation between count and outcome with outliers. The outlier depiction of glucose and insulin levels in terms of age, BMI, blood pressure, skin thickness, and pregnancies is shown in Figures 8–13.

Data balancing procedures employed in this study include the following:

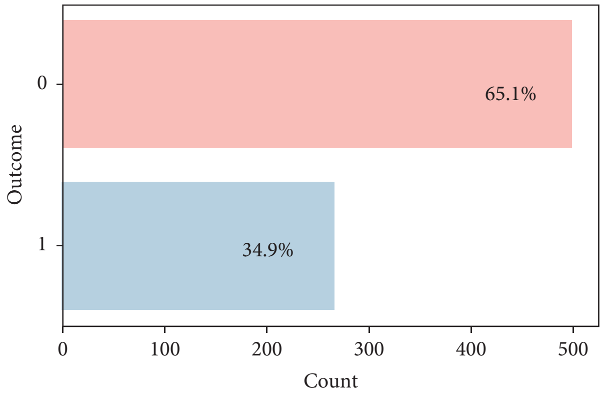

Figure 4: Relation between count and outcomes.

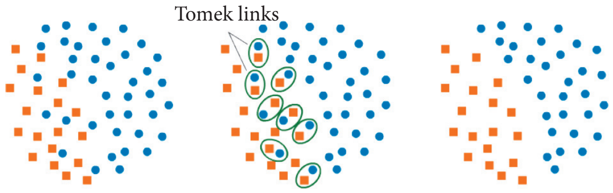

Figure 5: Tomek Links: undersampling.

(a) *Random Undersampling (imblearn).* RandomUnderSampler is a method for aligning datasets with disequilibrium. Using this method, you may quickly and easily verify the information. Data is selected for each target group at random. With or without substitution, choose random samples to test the plurality class(es).

(b) *Random Oversampling (imblearn).* Create new samples of minorities in order to address the problem of bias. Replacement of existing samples with new ones via random sampling is the most simplistic approach.

(c) *Undersampling (Tomek links).* Similarities between Tomek linkages and pairs of opposing groupings can be found. The region between the two classes is increased, making classification easier, by removing the higher class occurrences from each pair. Assuming the two samples are located near each other, Tomek's link is relevant.

(d) *Oversampling using (SMOTE).* The methodology based on this methodology provides fake data for the minority group. By using SMOTE, SMOTE generates a random point from the minority community (synthetic minority oversample technique). This point's neighbors are also calculated. The synthetic points shown in Figure 6 are added between the specified point and its neighbors.

Figure 7 shows the relation between count and outcome with outliers.

In order to eliminate outliers, the IQR approach is utilized when the boxplot data exceeds a specific range. The difference between the upper and lower quartiles is measured by the interquartile range (IQR). The interquartile range measures the difference between the upper and lower quartiles (IQR). Statistical methods such as IQR, Z-Score, and Data Smoothing were utilized to identify outliers in the data in this study. To calculate the IQR, the first and third quartiles of a data set, or the 25th and 75th percentiles, are used, and then Q1 is subtracted from Q3 to get the final IQR. The outlier depiction of glucose and insulin levels in terms of age, BMI, blood pressure, skin thickness, and pregnancies is shown in Figures 8–13.

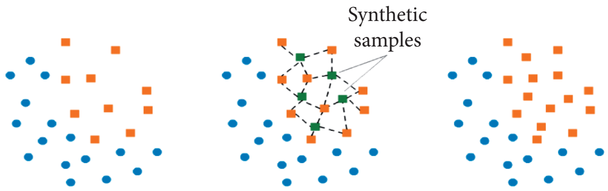

Figure 6: Synthetic minority oversampling.

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}. \tag{1}$$

A total of 10% of the dataset's samples have been deemed outliers and deleted.

3.4. Feature Engineering. This is the process of using data from a certain domain to develop functions that may be used by learning machines. It is the process of extracting and transforming raw data into machine-learning representations that is called for. A correlation matrix is employed in this study to discover the relationships between various variables.

3.4.1. Correlation Matrix. Correlation matrices are just a covariance matrix. The correlation sums up the linear association's strength. The frequency and direction of a straight-line connection between two quantitative variables are summarized by the concept of correlation. Values between 1 and 1 are represented by *r*. Pregnancies and age have a negative correlation, as seen in the chart below, whereas skin thickness has no effect on either of these variables.

3.5. Cross-Validation. With machine learning, cross-validation is the process of reassessing models in a small sample of data. A single parameter, *k*, governs the procedure, and it specifies how many groups of data should be formed from a given sample. K-fold cross-validation is another name for this method.

3.5.1. K-Fold Cross-Validation. Using a value between 5 and 10, depending on the quantity of the data, randomly divide the entire dataset into K-folds. Validate the model with the remaining Kth fold by fitting it with folds K-1 (K minus 1).

3.6. Classification Algorithms. Predicting diabetes utilizing the PIMA dataset, six efficient classifiers are used. There are six classifiers that are used for classification: SVM, Nave

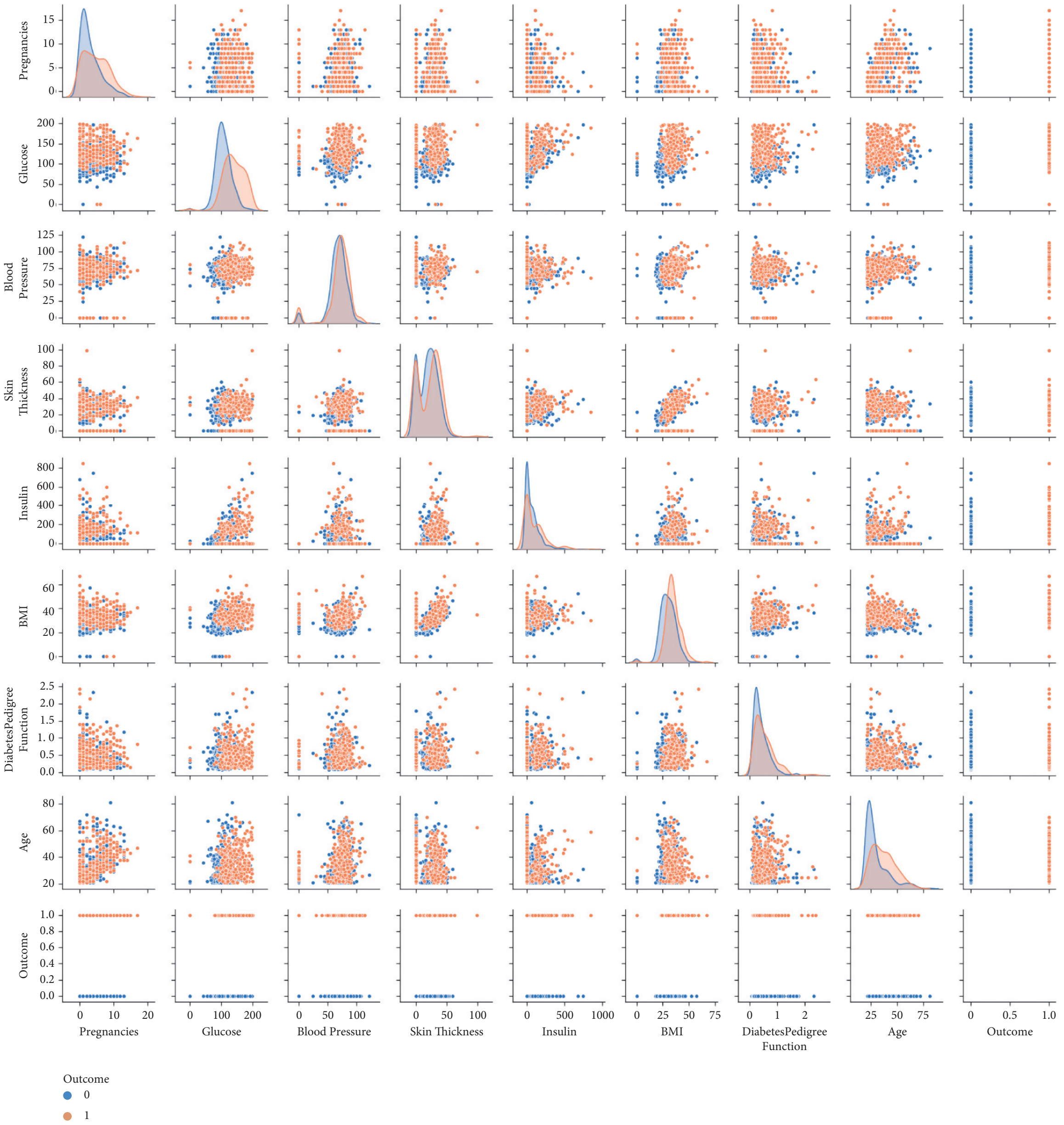

Figure 7: The relation between count and outcome with outliers.

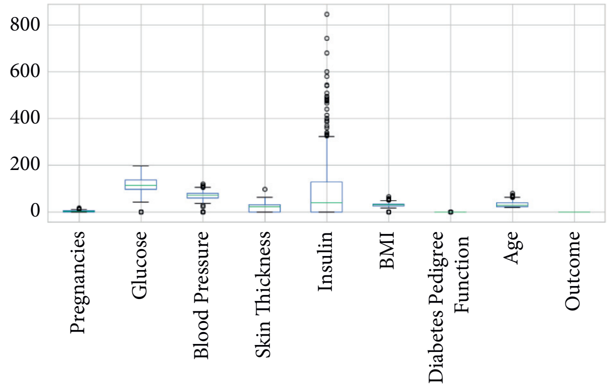

Figure 8: Outlier visualization.

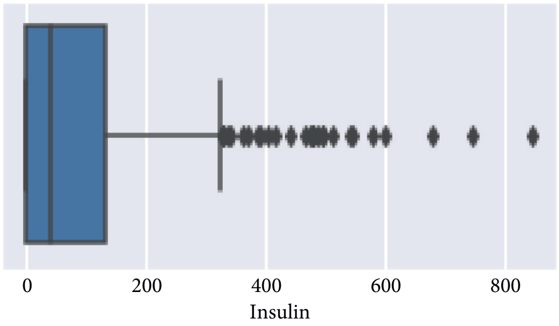

Figure 9: Insulin induction rate per patient.

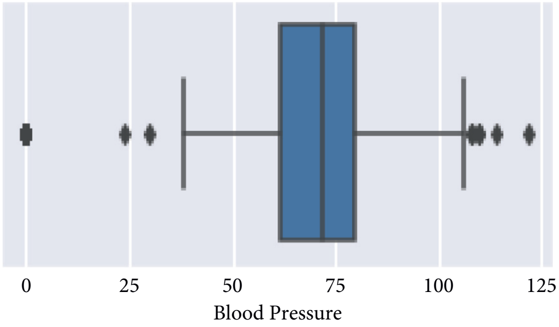

Figure 10: Boxplot (blood pressure range).

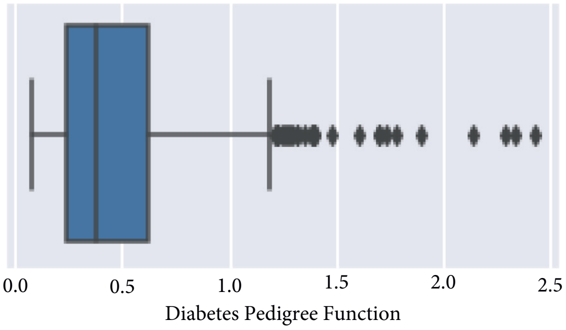

Figure 11: Diabetes pedigree function and its range.

Bayes, KNN, logistic regression, RF, and Gradient Boosting Classifier. Three of the most accurate classifiers from among these six are combined to assess the voting classifier.

3.6.1. Support Vector Machine (SVM). In a Support Vector Machine (SVM), the hyperplane is built between distinct classes or objects in order to classify the data. The hyperplane is generated by calculating the dimensions of the problem space. Additionally, dimensionality reduction is possible in SVM to balance data dimensions. In order to create a gap between the classes, the marginal distance is calculated from the hyperplane's center using the class corner points and the support vectors. Kernels, C coefficients, and intercepts are some of the parameters used in SVM. Kernels are the most important aspect of SVM. These kernels have been fine-tuned based on the type of data they process. Linear and Gaussian Kernels are useful in this study because the data is linear to RBF [46, 47].

3.6.2. Naive Bayes. Naive Bayes is the best algorithm for classification in machine learning. Based on the Naive Bayes Theorem, each object's likelihood was estimated, allowing it to foretell its appearance in any given class [11, 41–50]. Naive Bayes theorem states the following:

$$P(B) = \frac{P(A/B)}{P(A)}. \quad (2)$$

Gaussian Naive Bayes was utilized in this study. Data ambiguity is removed to generate significantly more accurate results using this method.

3.6.3. K-Nearest Neighbors (KNNs). Regression and classification problems in machine learning can be solved using k-nearest neighbors (KNNs). According to KNN algorithms, new data points can be discovered based on comparisons between existing data points (e.g., distance function). Neighbors are categorized by a simple majority vote. Means of the nearby training samples in feature space are used in KNN classification. Some of the PIMA diabetes and disease information is provided via some criteria. KNN is a well-known classification algorithm that falls within the classification supervised learning category [34–39].

3.6.4. Logistic Regression (LR). Regression analysis is used to categorize data into distinct categories. In logistic regression, the dependent variable is usually either true or false. In our scenario, a positive or negative diagnosis of diabetes is represented by a value of 1 or 0. Instead of a regression model, the term "logistic regression" refers to a linear classification model. Additionally, logistic regression can be thought of as a log-linear classifier. The logistical importance of this model is shown in the probability of characterizing a single test's potential outcomes. It presupposes that the data are Gaussian, which is not the case, as well as the variation in each of the learned traits. It is also a classification that is not observed. This classifier, however, has higher effectiveness in determining the disease kind. Machine-learning algorithms like logistic regression are very widespread. Despite its simplicity, it is a useful tool in a wide range of situations. The binary variable must be analyzed using linear regression as the regression analysis. The data is characterized and the link between one or more independent binary variables is clarified using logistic regression [12–15].

3.6.5. Random Forest (RF). Decision trees are built using training time and the mean estimate of the individual trees in the Random Forests ensemble learning technique for classification and other tasks. At the categorization step, the majority voting technique is used to achieve effective outcomes in determining the kind of diabetic disorders. A

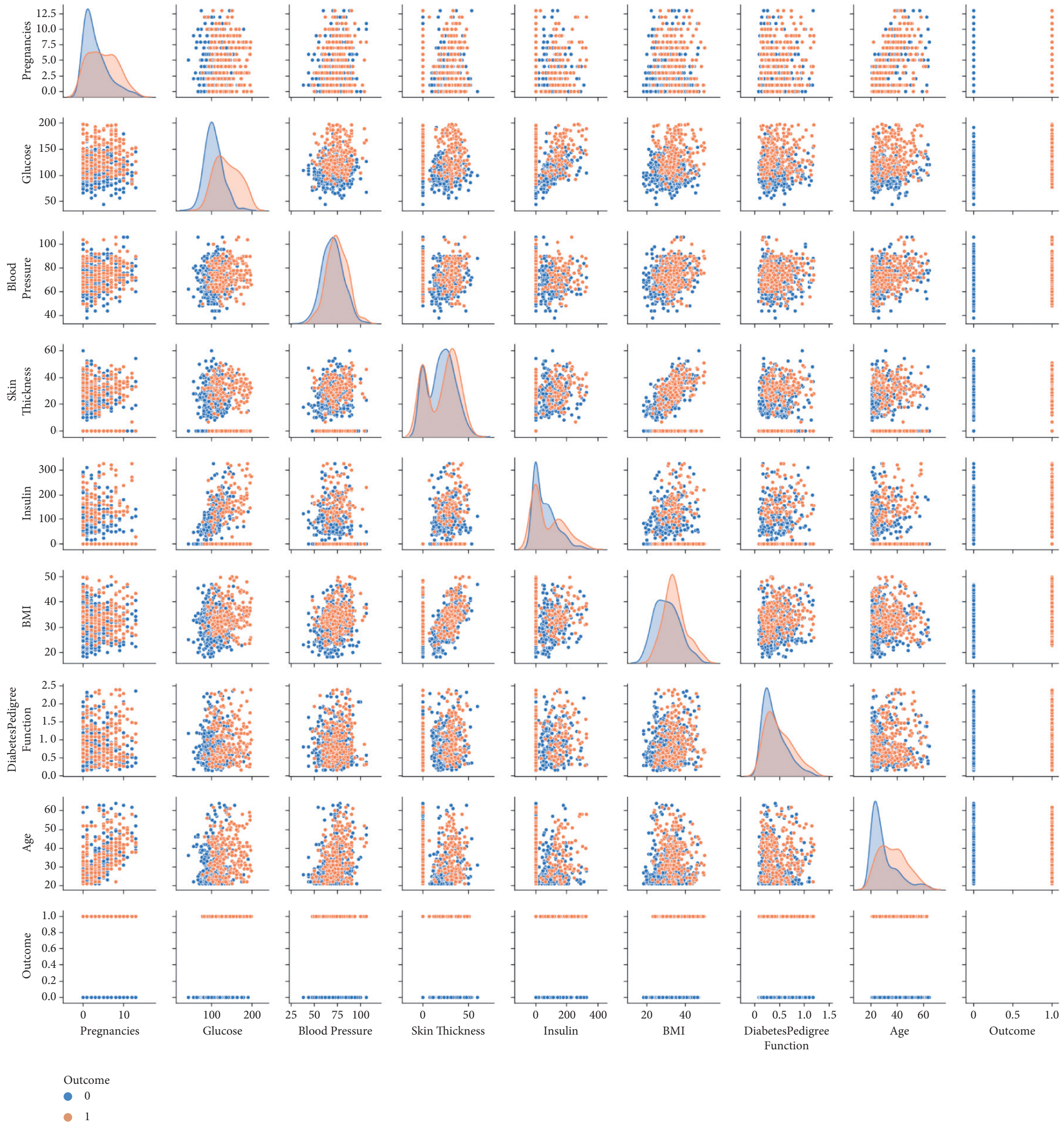

Figure 12: Scatter matrix after removing outlier.

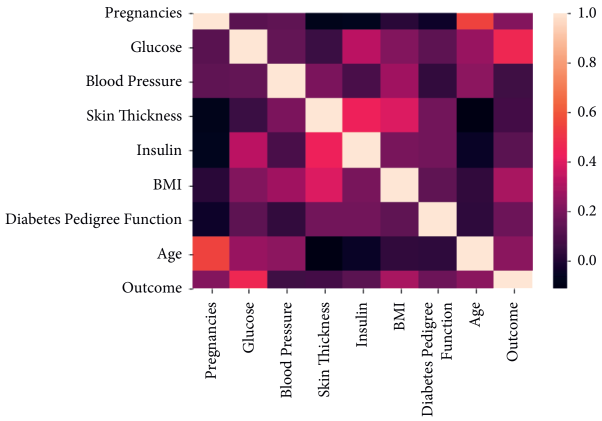

Figure 13: Correlation Matrices PIMA diabetes dataset.

normal patient's results might be categorized as either a normal or a diseased patient.

3.6.6. Gradient Boost Classifier (GBC). Groups of algorithms called Gradient Boosting Classifiers combine numerous poor learning models to achieve an effective prediction. It is usual practice to employ decision trees to increase the gradient. Regression and classification problems can be solved by using gradient boosting, a machine-learning technique that builds a predictive model from a collection of low-quality models. Decision trees with a weak learner are known as gradient trees, and they often outperform Random Forests in comparison. Rather than building the model sequentially, it applies it by minimizing an arbitrarily differentiable loss function, as other techniques do.

3.6.7. Ensemble Learning. The classification accuracy of the overall system can be improved by integrating a variety of different classifiers into a single platform. To improve classification accuracy, two or more machine-learning algorithms work together on the same topic.

3.6.8. Voting Classifier. Use a voting mechanism to select the best option from a list of several possibilities. As a result, numerous classifiers are able to select from a variety of options. A final decision is made in light of the choices made by the majority. If multiple algorithms are working on the same problem, a superior solution can be found. When employing ensembles in multiple categories, not everyone makes the same mistake.

3.7. Performance Parameters. Precision, sensitivity, specificity, and ROC are some of the performance metrics used to validate the suggested method. The following are the efficiency parameters calculated using the technique described above:

$$\begin{aligned} \text{specificity} &= \frac{\text{TN}}{(\text{TN} + \text{FP})}, \\ \text{accuracy} &= \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}, \\ \text{sensitivity} &= \frac{\text{TP}}{(\text{TP} + \text{FN})}, \\ \text{ROC} &= \frac{\text{sensitivity} + \text{specificity}}{2}. \end{aligned} \tag{3}$$

(i) False negative (FN): the patient sample result is 0 and there is a patient feature in the data file

(ii) True negative (TN): the feature result is 0 and the feature is not present in the diabetes data

(iii) A false positive (FP) occurs when the feature result is 1 and the feature is not present in the dataset

(iv) Features are present in this data file because it has a true positive (TP) result of 1

3.7.1. Accuracy. When it comes to accuracy, what counts is whether or not something can be proven to be true. In a matter of seconds, we will know whether or not the model has been adequately trained. The system is correct if the model is correctly educated. The accuracy of each method is to be compared in this work analysis. All models' efficacy must be evaluated.

3.7.2. Sensitivity and Specificity. Sensitivity is defined as the capacity to appropriately recognize a feature (true positive rate). Specificity, on the other hand, refers to the trial's ability to correctly classify people who do not have a syndrome.

4. Results

A CSV version of the PIMA data is created for use by ML. Patients are either healthy or sick. It is divided into two categories. Classification methods used to predict diabetes include the ones listed below.

4.1. KNN. (KNN) is a supervised ML method of this type. It can be utilized for both classifications and regression issues. However, it is mostly utilized in industry to deal with issues of categorization [23–25, 34].

K-NN classification relies on feature space nearby samples for classification. K-NN algorithm default performance is shown in Figure 14.

4.2. Logistic Regression. To represent the relationship between two variables, a logarithmic equation is utilized in the process of logistic regression. The dependent variable is referred to as such, whereas the independent variable is referred to as such. Overfitting is avoided by using the L1 and L2 regularization structures. To minimize overfitting, the coefficient values are decreased using L1 and L2 regularization.

Regularization occurs when one moment norm is equal to Euclidian Distance (which is |x1-x2|2), and for the other moment norm (L2), it is the absolute distance between the two points. By this, I mean that even if the "L1" can shrink all coefficients to 0, the "L2" does not do variable selection and instead contracts them all by similar amounts. In spite of the fact that all of the characteristics are linked to the tag/label, the ridge outperforms the lasso in ways such as the coefficients never being 0. A particular coefficient can be reduced to 0 in a lasso model if a subset of characteristics is connected with the tag/label. According to Figure 15, LR's default performance is shown.

4.3. Naive Bayes. It is a classifier based on probabilities. Figure 16 displays the naive Bayes algorithm's default performance.

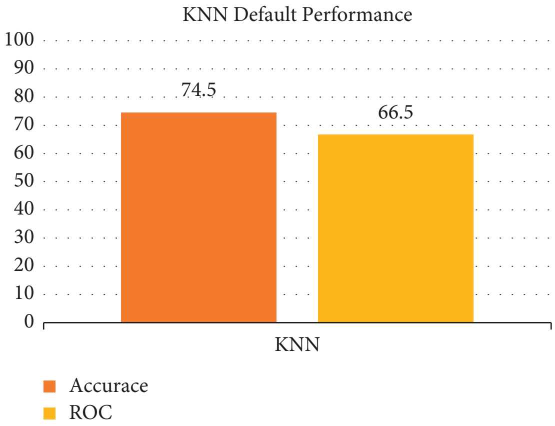

Figure 14: K-nearest neighbor performance.

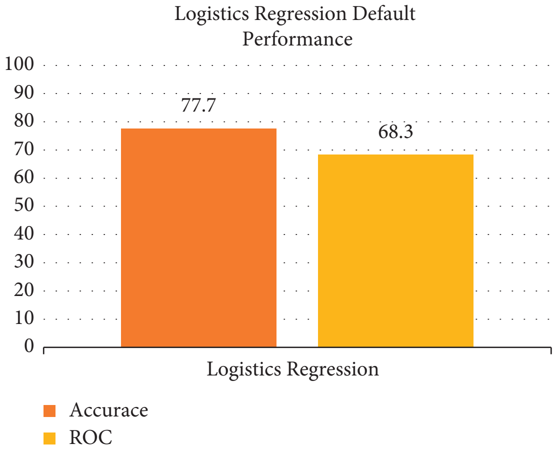

Figure 15: Performance of logistic regression.

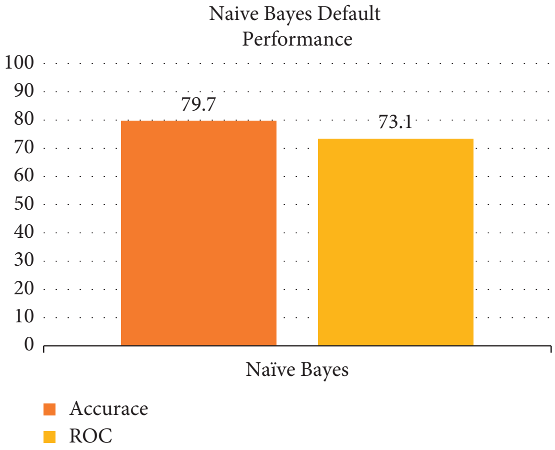

Figure 16: Performance of Naive Bayes.

4.4. Support Vector Machine. In the SVM classification, the classifier produces irregularity as a result of advanced dimensional characteristics in the original input datasets. SVM classifiers are more likely to notice anomalies because of these features. When it comes to anomaly classification, an SVM classificatory gets the job done quickly and accurately since it uses long-term estimations for specific classification parameters. Starting with more discriminatory traits and working your way down to less discriminating ones is how the distinction is earned. Homogeneity, comparison, similarity, mean, and likelihood characterize abnormality classification. Support Vector Machine is an example of a machine-learning method that is being closely scrutinized. Using it for classifications and regressions is a natural fit for the software. The influence of a particular sample distribution with low standards is represented by the gamma limit. Contrary to the evaluation surface's simplicity, the "C" parameter trades on incorrectly identifying training data. Figure 17 displays the SVM algorithm's default performance.

4.5. Random Forests. In essence, it is a divide-and-conquer method for group learning. This collection of decision tree classifiers is another name for the forest. Attribute collection predictors such as knowledge gain, profit ratio, and Gini index are integrated into each determined tree. Random selection is used to build each tree. In a classification problem, each tree votes and the most common class is the final outcome. Each tree output is averaged to arrive at a final result when using regression. This algorithm is smoother and faster than other nonlinear classification algorithms. Figure 18 shows the results with the default RF parameters:

4.6. Gradient Boost Classifier. Gradient boosting is a group of techniques that work together to build a predictive model for many poor learning models. Gradients are frequently bolstered through the usage of decision trees. As a result of their efficiency, gradient booster models have become increasingly common.

They are known as gradient booster classifiers because they are used to classify functions. The computer's learning method and the notion of value are based on features. In a mathematical setting, the features of the data set are the variables that are employed to answer the equation. The label or target, which is the group of instances, is the other component of the equation. During the training process, data should be separated into training and test sets since labels offer goal values for the classifier in machine learning. The training set has goals and labels, while the test set does not. Figure 19 displays the GBC algorithm's default performance.

4.7. Voting Classifier. A machine-learning algorithm is based on the highest probability of the chosen class. Performance is predicted using a variety of methods. In essence, it aggregates the results of any classed vote and forecasts the output class based on a large majority of votes. A single model, depending on the majority of votes in each class, is created instead of multiple models, each dedicated to a certain accuracy goal. There are two ways to cast a vote with Voting Classification.

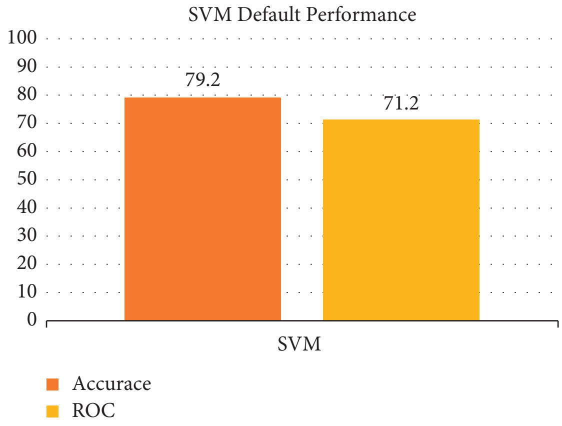

Figure 17: Performance of SVM.

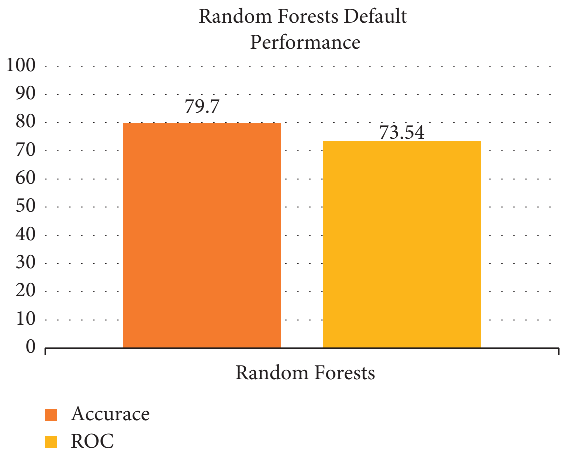

Figure 18: Random Forest.

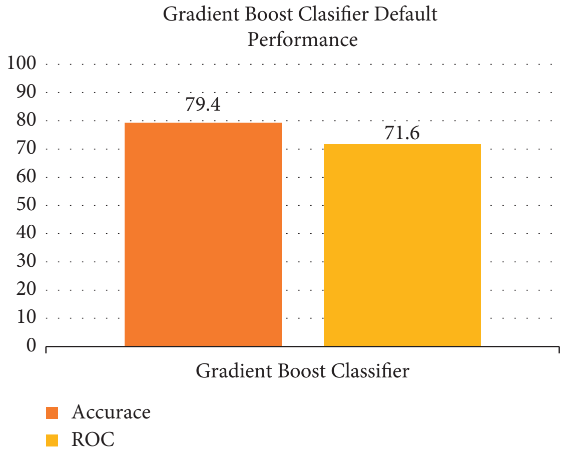

Figure 19: Performance of Gradient Boosting Classifier.

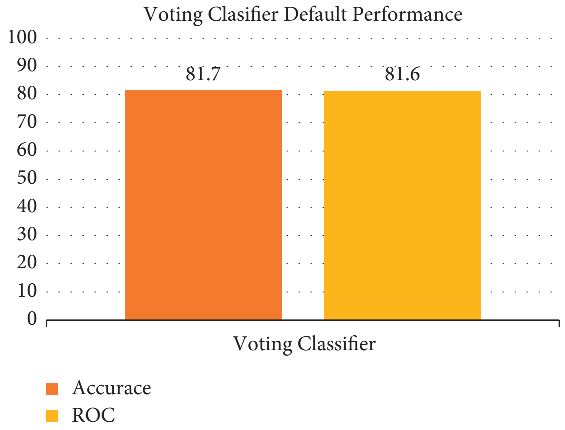

Figure 20: Voting classifier comparison using soft and hard voting.

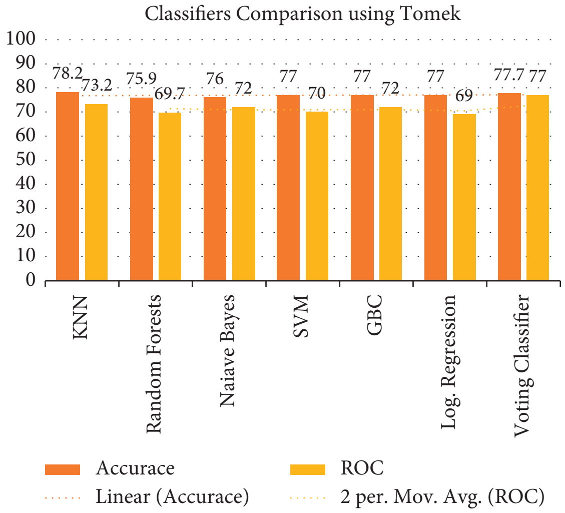

Figure 21: Classifiers' comparison using Tomek link undersampling.

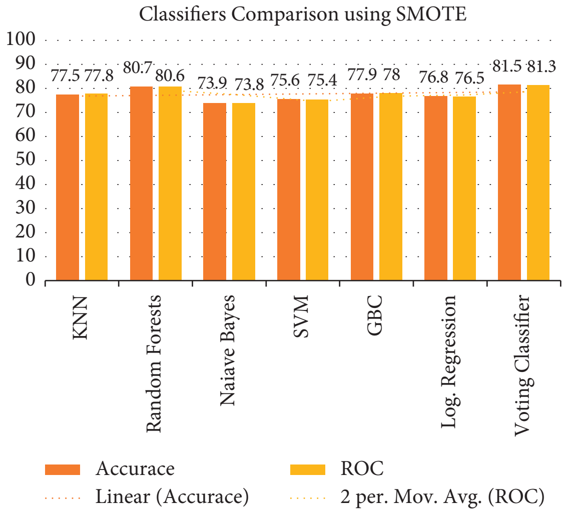

Figure 22: Classifiers' comparision using SMOTE oversampling.

In a hard voting scenario, the performance class that is most likely to be selected is the one with the most number of votes cast, i.e., the one that the classifier most accurately predicted. This class's average probability is used to predict the performance class in soft voting. As can be seen in Figure 20, the voting method performs as expected by default.

Table 3: Comparative analysis.

Classifier/algorithm	Technique	Accuracy %	ROC %
Logistic regression	Default	77.7	68.3
	Tomek- undersampling	77.0	69.7
	Smote- oversampling	76.7	76.5
SVM	Default	79.2	71.2
	Tomek- undersampling	77.7	70.2
	Smote- oversampling	7.2	70.3
KNN	Default	74.5	66.5
	Tomek- undersampling	78.2	73.3
	Smote- oversampling	77.5	77.9
Gradient boost	Default	79.4	71.6
	Tomek- undersampling	77.0	72.0
	Smote- oversampling	77.9	78.0
Naive Bayes	Default	79.7	73.1
	Tomek- undersampling	75.5	72.0
	Smote- oversampling	73.9	73.8
Random Forests	Default	79.7	73.5
	Tomek- undersampling	75.9	69.8
	Smote- oversampling	80.7	80.6
Voting classifier (using three best models as an input)	Default	**81.7**	**81.6**
	Tomek- undersampling	**77.7**	**76.5**
	Smote- oversampling	**81.5**	**81.5**

4.8. Undersampling of Dataset Using Tomek Links. Tomek has a mix of near and far-flung companions. It is easier to sort the two classes when the multiclass instances in each pair are gone. The Tomek linkages were employed for undersampling in this study. To check for undersampling, we used the Tomek linkages shown in Figure 21, and the following are the outcomes:

4.9. Oversampling of Dataset Using SMOTE. Synthetic data for the minority group is generated using this technique of research. A random point from the lower-class community is chosen and its nearest neighbors are calculated through SMOTE to determine this point's location. A set of synthetic points is placed in the vicinity of the currently selected point and those of its immediate neighbors. The SMOTE has been employed for oversampling in this study. There are SMOTE linkages applied to oversampling for each classification model as shown in Figure 22.

4.10. Comparative Analysis. The accuracy of each classifier changes when balancing procedures are applied to a dataset. Results and comparisons of all research methods are shown in Table 3.

5. Conclusions

Diabetes mellitus diseases is a critical and long-lasting disease. Detection of diabetes at the primary stage can lead to better-quality treatment. This research proposes a diabetes estimation model for accurate classification of diabetes that takes into account characteristics such as glucose, body mass index, age, and insulin. Predictive models face difficulties when faced with an unbalanced dataset. As a result, data balancing techniques (Tomek and SMOTE) were utilized to balance the dataset. Outliers have been removed from data to make it more useable. Additionally, this study compares various machine-learning algorithm-based classification models for predicting a patient's diabetic state at the earliest feasible stage. After balancing the dataset, the accuracy of classifiers was compared. Random Forest outperformed logistic regression, Support Vector Machine, k-nearest neighbors, Naive Bayes Theorem, and Gradient Boosting Classifier algorithms with an accuracy of 80.7 percent. Additionally, a voting classifier was evaluated and found to be 81.7 percent accurate on the original dataset and 81.5 percent accurate on the balanced dataset. Additionally, this work can be expanded to determine the likelihood that nondiabetic people would develop diabetes in the following several years based on a person's lifestyle and physical inactivity [47–49].

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 there are no conflicts of interest.

References

[1] S. Revathy, B. Bharathi, B. Bharathi, P. Jeyanthi, and M. Ramesh, "Chronic kidney disease prediction using machine learning models," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 6364–6367, 2019.

[2] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

[3] D. U. N. Qomariah, H. Tjandrasa, and C. Fatichah, "Classification of diabetic retinopathy and normal retinal images using CNN and SVM," in *Proceedings of the 2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pp. 152–157, IEEE, Surabaya, Indonesia, 18 Jul. 2019.

[4] S. Sharma, "Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients," *Environmental Science and Pollution Research*, vol. 27, no. 29, pp. 37155–37163, 2020.

[5] A. M. Syed, M. U. Akram, T. Akram, M. Muzammal, S. Khalid, and M. A. Khan, "Fundus images-based detection and grading of macular edema using robust macula localization," *IEEE Access*, vol. 6, pp. 58784–58793, 2018.

[6] N. Chakrabarty, "A deep learning method for the detection of diabetic retinopathy," in *Proceedings of the 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) 2018*, no. January, pp. 1–5, Gorakhpur, India, 2 November 2018.

[7] A. Narin, C. Kaya, and Z. Pamuk, *Department of Biomedical Engineering*Zonguldak Bulent Ecevit University, Zonguldak, Turkey. https://arxiv.org/abs/2003.10849, Article ID 67100, 2020.

[8] C. Lam, C. Yu, L. Huang, and D. Rubin, "Patches," 2018.

[9] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, and A. Elbaz, "Deep learning based method for computer aided diagnosis of diabetic retinopathy," *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, p. 19, 2019 –22.

[10] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *International Journal of Information Technology*, vol. 12, no. 3, pp. 731–739, 2020.

[11] D. S. W. Ting, L. R. Pasquale, L. Peng et al., "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.

[12] J. Kalita and V. Emilia, *Advances in Intelligent Systems and Computing 740 Recent Developments in Machine Learning and Data Analytics*, 2018.

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2010–1981, 2010.

[14] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artificial Intelligence in Medicine*, vol. 104, no. February, Article ID 101815, 2020.

[15] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[16] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, no. 2019, pp. 706–716, 2020.

[17] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, no. April, Article ID 100179, 2019.

[18] T. Mahboob Alam, M. A. Iqbal, Y. Ali et al., "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, no. January, Article ID 100204, 2019.

[19] A. Cahn, A. Shoshan, T. Sagiv et al., "Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model," *Diabetes*, vol. 36, no. 2, pp. 1–8, 2020.

[20] J. Balcázar, Y. Dai, and O. Watanabe, "A random sampling technique for training support vector machines," *Lecture Notes in Computer Science*, pp. 119–134, 2001.

[21] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[22] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–15, 2019.

[23] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90–100, 2022.

[24] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI) 2018*, p. 447, 4 June 2018.

[25] A. Choudhury and D. Gupta, *A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques*, Vol. 740, Springer, Singapore, 2019.

[26] V. Rodriguez-Romero, R. F. Bergstrom, B. S. Decker, G. Lahu, M. Vakilynejad, and R. R. Bies, "Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques," *Clinical and Translational Science*, vol. 12, no. 5, pp. 519–528, 2019.

[27] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance analysis of machine learning techniques to predict diabetes mellitus," in *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) 2019*, pp. 1–4, 2019.

[28] S. Islam Ayon, M. Milon Islam, and M. Milon Islam, "Diabetes prediction: a deep learning approach," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 2, pp. 21–27, 2019.

[29] H. Liu, J. Li, J. Leng et al., "Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China," *Diabetes*, vol. 37, no. 5, 2020.

[30] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods," *Machines*, vol. 7, no. 4, pp. 74–11, 2019.

[31] M. Makino, R. Yoshimoto, M. Ono et al., "Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[32] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Prognostic modeling and prevention of diabetes using machine learning technique," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[33] M. W. Segar, M. Vaduganathan, K. V. Patel et al., "Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score," *Diabetes Care*, vol. 42, no. 12, pp. 2298–2306, 2019.

[34] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, 2019.

[35] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019*, no. Iccmc, pp. 367–371, IEEE, Erode, India, 27 March 2019.

[36] N. Yuvaraj and K. R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,” *Cluster Computing*, vol. 22, no. S1, pp. 1–9, 2019.

[37] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P. K. Abhilash, K. Sandeep, and D. Indira, “Prediction of diabetes using machine learning,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7, pp. 2547–2551, 2019.

[38] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan, and S. R. Vikash, “Machine learning tree classifiers in predicting diabetes mellitus,” in *Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp. 84–87, IEEE, Coimbatore, India, 15 March 2019.

[39] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” *Health Information Science and Systems*, vol. 8, no. 1, pp. 1–14, 2020.

[40] P. Kashyap and P. Kashyap, “Industrial applications of machine learning,” in *Machine Learning for Decision Makers*, pp. 189–233, Apress, New York, US, 2017.

[41] F. Cabitza, A. Locoro, and G. Banfi, “Machine learning in orthopedics: a literature review,” *Frontiers in Bioengineering and Biotechnology*, June, vol. 6, , 2018.

[42] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264, IGI Global, Hershey PA, 2009.

[43] P. Andreasson, J. Johansson, S. Liljestrand, and M. Granath, “Quantum error correction for the toric code using deep reinforcement learning,” *Quantum*, vol. 3, p. 183, 2019.

[44] F. Ma, T. Sun, L. Liu, and H. Jing, “Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network,” *Future Generation Computer Systems*, vol. 111, pp. 17–26, Oct. 2020.

[45] A. Unnikrishnan, F. Ajesh, and R. S. Nair, “Detection of abnormal visual events using HOFO and KNN,” vol. 2, no. 9, pp. 3196–3210, 2015.

[46] M. Jethanandani, T. Perumal, and A. Sharma, “Random k-Labelsets method for human activity recognition with multi-sensor data in smart home,” in *Proceedings of the 2019 IEEE 16th India Council International Conference (INDICON)*, 13-15 Dec. 2019.

[47] “Pima Indians diabetes database | kaggle,” 2021, https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[48] C. Kruse, P. Eiken, P. Vestergaard, C. Kruse, P. Eiken, and P. Vestergaard, “Machine learning principles can improve hip fracture prediction,” *Calcified Tissue International*, vol. 100, no. 4, pp. 348–360, 2017.

[49] N. Golestani and M. Moghaddam, *Magnetic Induction-Based Human Activity Recognition*, pp. 17-18, MI-HAR), no. Mi, 2019.