Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 6688934, 13 pages
https://doi.org/10.1155/2024/6688934

WILEY | Hindawi

Research Article

A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App

Hosam El-Sofany,[1,2] **Samir A. El-Seoud**,[3] **Omar H. Karam**,[3] **Yasser M. Abd El-Latif**,[4] **and Islam A. T. F. Taj-Eddin**[5]

[1]*College of Computer Science, King Khalid University, Abha, Saudi Arabia*
[2]*Cairo Higher Institute for Engineering, Computer Science and Management, Cairo, Egypt*
[3]*British University in Egypt- BUE, Faculty of Informatics and Computer Science, Cairo, Egypt*
[4]*Faculty of Science, Ain Shams University, Cairo, Egypt*
[5]*Faculty of Computers and Information, Assiut University, Assiut, Egypt*

Correspondence should be addressed to Hosam El-Sofany; helsofany@kku.edu.sa

Received 8 August 2023; Revised 13 December 2023; Accepted 21 December 2023; Published 9 January 2024

Academic Editor: Gianni Costa

With the increasing prevalence of diabetes in Saudi Arabia, there is a critical need for early detection and prediction of the disease to prevent long-term health complications. This study addresses this need by using machine learning (ML) techniques applied to the Pima Indians dataset and private diabetes datasets through the implementation of a computerized system for predicting diabetes. In contrast to prior research, this study employs a semisupervised model combined with strong gradient boosting, effectively predicting diabetes-related features of the dataset. Additionally, the researchers employ the SMOTE technique to deal with the problem of imbalanced classes. Ten ML classification techniques, including logistic regression, random forest, KNN, decision tree, bagging, AdaBoost, XGBoost, voting, SVM, and Naive Bayes, are evaluated to determine the algorithm that produces the most accurate diabetes prediction. The proposed approach has achieved impressive performance. For the private dataset, the XGBoost algorithm with SMOTE achieved an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87. For the combined datasets, it achieved an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85. To understand how the model predicts the final results, an explainable AI technique using SHAP methods is implemented. Furthermore, the study demonstrates the adaptability of the proposed system by applying a domain adaptation method. To further enhance accessibility, a mobile app has been developed for instant diabetes prediction based on user-entered features. This study contributes novel insights and techniques to the field of ML-based diabetic prediction, potentially aiding in the early detection and management of diabetes in Saudi Arabia.

1. Introduction

Biotechnology and public healthcare infrastructure advancements have contributed to significant developments in sensitive and essential healthcare data processing. When intelligent methods of data analysis are applied, many significant characteristics may be used to detect and prevent various chronic illnesses in their early stages. *Diabetes* is a disease that is increasing at the fastest rate among individuals of all ages, including children and the elderly. Diabetes is a chronic condition that, if it becomes common, might cause a crisis in healthcare on a global scale. Diabetes is characterized by many *symptoms*, some of which include increased urine frequency, increased thirst, increased tiredness and drowsiness, decreased appetite, decreased body weight, impaired vision, mood changes, disorientation, difficulties in focusing, and recurrent infections [1]. Diabetes is very dangerous to a human's life, mainly because it raises the risk of contracting other fatal diseases, such as strokes, blindness, miscarriages, amputations, and kidney failure.

The International Diabetes Federation (IDF) reports that the number of people being diagnosed with diabetes is continuously increasing. The number of diabetics in the globe is expected to reach 642.8 million by the year 2030 [1]. In Saudi Arabia, the number of diabetes patients will reach approximately 5.61 million by 2030, according to 2023 statistics [2].

Research in ML is an emerging topic of artificial intelligence that examines how machines might acquire knowledge via their interactions with the world. In this work, an ML-based method for classifying, identifying, and predicting diabetes in its early stages has been presented [3]. Before we can comprehend diabetes and how it occurs, we need to have a good understanding of what goes on in the body when it is not affected by diabetes. As mentioned in [3], the meals that we consume, particularly those that are high in carbohydrates, are the sources of the sugar (glucose) that our bodies need. All humans, even diabetics, require carbohydrate-containing meals as their primary source of dietary energy. Foods like rice, bread, pasta, cereal, dairy, fruit, and vegetables all fall under the category of carbohydrates. The digestive process converts these foods into glucose. Glucose is transported all over the body via the circulatory system [4]. The brain receives a portion of the glucose to enhance its ability to think and perform tasks. The rest of the glucose is used immediately by our cells for energy and stored in the liver for later use. Insulin is needed for glucose to be used as an energy source in the body. Insulin is produced by pancreatic beta cells. Insulin functions similar to a door key. Insulin binds to the gates of cells and opens them, allowing glucose to enter the cell from the bloodstream. When insulin production is inadequate due to pancreatic dysfunction, or when insulin is produced but cannot be used by the body (a condition known as insulin resistance), this condition is known as *diabete*s. Then, glucose will build up in the bloodstream, leading to hyperglycemia and the development of diabetes. The chronic illness known as diabetes mellitus is distinguished by excessively high blood sugar and elevated urine sugar levels [5].

According to [1], there are three primary forms of diabetes, but the one that affects the most people is type 1, in which cells are unable to produce sufficient insulin and the immune system becomes compromised. There is no definitive research on type 1 diabetes causation or prevention. Type 2 diabetes is characterized by the body's cells not creating an appropriate amount of insulin or when the insulin that is produced by the body is not being used effectively. This is the kind of diabetes that affects most people, and as a result, it affects 90% of people who have been diagnosed with diabetes. Both the way people lead their lives and their genes have a role in causing this condition. Gestational diabetes happens when a woman who is pregnant suddenly gets high blood sugar, which can be bad for both her health and the health of her baby. Two-thirds of the time, it will show up again in later pregnancy. There is a high possibility that a woman with gestational diabetes will get *type 1* or *type 2* diabetes after childbirth. All types of diabetes should be treated instantly because they are dangerous. If one detects the early stages of these diseases, one can prevent the difficulties that are associated with them [6].

This article proposes a machine learning-based technique for diabetes prediction through a mobile app. The important contribution of this work includes the following:

(1) The main contribution of this work is to utilize a private dataset of diabetes mellitus. The dataset is comprised of 300 volunteer data samples that were obtained from specialty hospitals in both Saudi Arabia and Egypt during the 2022-2023 academic year. There are nine features that we have gathered from a total of 300 participants. These features include gender, age, glucose, insulin, blood pressure, pregnancy, skin thickness, body mass index, diabetes pedigree function, and the outcome of diabetes.

(2) The study addresses the critical need for early detection and prediction of diabetes in Saudi Arabia, where the prevalence of the disease is increasing. By applying ML techniques to Pima Indians and private datasets, the researchers propose a computerized system for diabetes prediction.

(3) One key contribution of this work is the use of a semisupervised model combined with strong gradient boosting. This approach effectively predicts diabetes-related features of the dataset, which is a novel technique compared to prior research. By employing this method, the study achieves impressive performance in predicting diabetes, with an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87 for the private dataset and an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85 for the combined datasets. In this work, we have implemented ML and explainable AI techniques to predict diabetes.

(4) Another contribution of this work is the effective handling of imbalanced classes using the Synthetic Minority Oversampling Technique (SMOTE). This addresses a common problem in diabetic prediction, where the number of positive cases is significantly lower than the number of negative cases. By employing SMOTE, the study ensures that the prediction model is trained on a balanced dataset, improving the accuracy of the predictions.

(5) This work also included the implementation of hyperparameter optimization on the employed ML classifiers. For each of the ML frameworks, hyperparameter tuning has been done. The proposed technique with improved hyperparameters reached a maximum accuracy of 97.4% on the private dataset and an accuracy of 83.1% on the combined dataset.

(6) Furthermore, the study evaluates ten different ML classification techniques to determine the algorithm that produces the most accurate diabetes prediction. By comparing the performance of various algorithms including logistic regression, random forest, KNN, decision tree, bagging, AdaBoost, Extreme Gradient Boosting (XGBoost), voting, SVM, and Naive Bayes,

the researchers can identify the XGBoost algorithm as the most effective for predicting diabetes. The adaptability of the proposed system is demonstrated through the application of a domain adaptation method. This showcases the potential of the system to be applied in different settings and populations, beyond the initial dataset used in the study. To enhance accessibility and promote early detection of diabetes, researchers have developed a mobile app for instant diabetes prediction based on user-entered features. This mobile app further contributes to the field by providing a user-friendly and easily accessible tool for individuals to monitor their diabetes risk.

(7) To understand the model's diabetes prediction process, an explainable AI approach using SHAP libraries (under the Python environment platform) is implemented. Using this method, we may better understand the features that are most important for making accurate predictions.

(8) Overall, this study contributes novel insights and techniques to the field of ML-based diabetic prediction. The use of a semisupervised model, the application of the SMOTE technique to handle imbalanced classes, the evaluation of multiple ML algorithms, and the development of a mobile app are all significant contributions that can aid in the early detection and management of diabetes in Saudi Arabia. This work has the potential to improve healthcare outcomes and prevent long-term health complications associated with diabetes.

2. Literature Review

Diabetes affects a sizable percentage of the adult population. Many research works were proposed for the prediction of diabetes symptoms. A wide variety of approaches, including ML, neural networks (NNs), data mining, and genetic algorithms, are discussed in these studies. In recent years, ML has gained popularity as a model-building technique and received a lot of attention from the medical community. ML has proven to have strong prediction powers as well as the capacity to analyze many variables in parallel. Moreover, ML has developed methods for variable screening that can recognize and comprehend intricate correlations between variables. Previous research has proven that ML may be a useful technique for predicting diabetes. Some closely related works using ML algorithms are discussed in this section.

In [7], the authors developed five different models using different ML algorithms. Some of them include linear SVM, multifactor dimensionality reduction, radial basis function, kernel SVM, KNN, and artificial neural network. This study used a Boruta wrapper, which completely selects important features from a dataset. According to the experimental findings, all the models appeared to have achieved good results. However, KNN and linear SVM are the two models that performed the best at identifying whether a patient has diabetes or not.

In [8], type 2 diabetes could be diagnosed using a hybrid model. Researchers proposed the T2ML model, which included a suggested list of steps. Cleaning the data to ensure homogeneity was the first step, followed by the selection of a subset of features based on XGB classifiers and RF. After that, K-means clustering was used to exclude data that had been erroneously classified, and ultimately, a logistic regression classifier and clustering were coupled to categorize the data that were left out.

In addition, Maniruzzaman et al. [9] selected four classifiers for the prediction of diabetes patients including RF, NB, AB, and DT. Risk factors for diabetic sickness were calculated with LR using the p value and odds ratio. These strategies relied on three distinct techniques for partitioning, referred to as K2, K5, and K10. These classifiers were evaluated using accuracy and AUC as performance metrics. The researchers found that age, blood pressure, diastolic blood pressure, total cholesterol, and body mass index are all significant risk factors for diabetes. Additionally, the performance of LR and RF-based classifiers combined improved, which can help make predicting diabetes individuals much easier.

Moreover, to make an accurate prediction about type 2 diabetes mellitus, the authors of [10] compared many commonly used regression models, including Glmnet, RF, XGB, and LightGBM. Initially, the dataset consisted of 111 variables, which have been reduced to 58 variables after a data preprocessing step. This study compared the performance, calibration, and interpretability of multivariable regression models and ML-based prediction models. Their findings demonstrate that updating prediction models with new data not only enhances the performance of the prediction but also maintains variable importance ranking, but not uniformly across all ML models.

Among other ML algorithms, the researchers in [11] employed DT, RF, KNN, AB, XGB, and NB. They used the ensemble model that was suggested for the dataset PID, where preprocessing is essential for a reliable and accurate prediction. The results showed that the optimal setup for predicting diabetes is the combination of two boosting type classifiers (XGB and AB). When the suggested preprocessing is used, the dataset utilizing the optimal combination (AB + XGB) can predict diabetes with a better degree of accuracy. In [12], the authors attempted to make use of ML techniques to find effective models to predict diabetes. They used many ML algorithms for the training of datasets, such as LR, DT, NB, gradient boosting (GB), RF, KNN, and SVM. To improve the prediction models' accuracy rates, preprocessing procedures were used which include label encoding and normalization. SVM outperformed the alternative methods, according to the authors. The suggested model used effective preprocessing step approaches, such as label encoding and normalization, to increase the models' predictive power. The authors further discovered and ranked several risk indicators using a variety of feature selection approaches.

In [13], the authors used ML techniques to look for diabetic patients. They first utilized a bootstrapping resampling strategy in the PIMA dataset to increase accuracy

before using KNN, NB, and DT. Results proved that applying the preprocessing step to the data increases relied on the accuracy of almost all classifiers, but the decision trees led over others. For an accurate diagnosis of five main illnesses, including diabetes, ML approaches were proposed in [14]. In this study, the authors trained LR and RF models using the BRFSS dataset. A chatbot was used to gather user input, anticipate the prevalence of chronic diseases, and model the data using interactive data visualization techniques to offer risk-reduction recommendations. They tried several factors and concluded that RF could identify diabetes with good accuracy.

Similarly, Khanam and Foo [15] used the PIDD dataset with its different attributes, and seven distinct ML models were trained. In this method, two features were discarded as part of a feature selection process. They found that the model with SVM and LR performed well in predicting diabetes. An NN model was trained using the same dataset with many hidden layers with different epochs. In comparison to previous methods, the authors indicate that an NN with two hidden layers achieves better performance.

Additionally, a meta-analysis of ML's ability to diagnose diabetes was carried out [16]. It was discovered that the ML algorithms in use today are strong enough to assist physicians in predicting whether a patient would eventually acquire type 2 diabetes.

In [17], ML was used to conduct a meta-analysis of diabetes prediction methods. With the use of the innovative PROBAST (Prediction Model Risk of Bias Assessment Tool), the potential for bias in the ML models was examined. To conduct the meta-analysis and assess heterogeneity, the Meta-DiSc software package was used. It was discovered that ML models outperformed traditional screening techniques in terms of predicting diabetes.

The basic algorithm in [18] was LR although a few additional ML approaches, including DT, NB, SVM, and KNN, were utilized in ensemble techniques to evaluate performance improvement. Two datasets were primarily used in the experiment, and two alternative strategies for feature selection were used. The Pima Indians dataset, which includes nine different features, was chosen as the initial dataset. The Vanderbilt dataset, which has 16 features, was the second dataset utilized. This study's findings indicated that the LR algorithm is one of the most successful ones that may be used to construct prediction models. In addition to the method of procedure, the researchers demonstrated that a number of additional parameters also affect the model's accuracy.

Furthermore, to anticipate diabetes more accurately and to classify diabetes properly, the authors in [19] offered an efficient model for doing so. The researchers used a variety of ML algorithms, including LR, RF, SVM, DT, KNN, AB, Gaussian Naive Bayes (GNB), and Gaussian process classifier (GPC). These models' performances were evaluated according to their respective precision, accuracy, F-measure, recall, and error metrics.

The development and evaluation of semisupervised learning models for insulin prediction is the main aim of the research study [20]. The researchers employ a private dataset that includes patient information such as demographics, clinical features, and historical insulin records. Traditional supervised models often face limitations due to the scarcity of labeled data, significantly impacting their performance. Semisupervised learning algorithms use unlabeled data to address this issue. The authors begin by preprocessing the dataset, addressing missing values, scaling features, and splitting it into training and testing sets. They then propose and compare various semisupervised models, including self-training, cotraining, and label propagation algorithms. Through extensive experimentation and evaluation, the authors evaluate each model using accuracy, precision, recall, and F1 score. The findings of the article demonstrate the potential benefits of semisupervised learning models for insulin prediction. The experiments reveal that the inclusion of unlabeled data during model training enhances prediction performance, particularly when labeled data are limited. Notably, the self-training model exhibits the highest accuracy and F1 score, suggesting its efficacy in leveraging unlabeled data. However, the authors recognize that their study's restricted dataset and lack of comparison with advanced supervised models limit the generalizability of their findings. One of the strengths of the article lies in its exploration of semisupervised learning techniques for insulin prediction, an area that has received limited attention in existing literature. The authors provide comprehensive details on dataset preprocessing, model architecture, and evaluation metrics. Furthermore, they emphasize the importance of incorporating unlabeled data and demonstrate the potential of such models using a private dataset. To conclude, the article successfully investigates the application of semisupervised learning models for predicting insulin levels using a restricted dataset. The findings suggest that incorporating unlabeled data holds promise for enhancing prediction performance, especially when labeled data are scarce.

The existing research on diabetes prediction using ML techniques has shown progress in identifying diabetes-related features and developing accurate prediction models. However, there are still some gaps in the literature that need to be addressed. The following are some gaps, along with an explanation of how the proposed article contributes to filling them:

(i) Many studies have focused on using a single ML algorithm, such as logistic regression, SVM, or random forest, for diabetes prediction. While these algorithms have demonstrated promising results individually, there is a lack of comprehensive comparison and evaluation of multiple ML algorithms. This limits generalizability and the choice of the most reliable and accurate diabetes prediction approach. The proposed article directly addresses this gap. It conducts a thorough comparison and evaluation of multiple ML techniques, including logistic regression, random forest, KNN, decision tree, bagging, AdaBoost, XGBoost, voting, SVM, and Naive Bayes. By comparing their performance metrics like accuracy, F1 coefficient, and AUC on

diabetes prediction, the article identifies the most accurate algorithm for this task.

(ii) Imbalanced classes in diabetes prediction datasets make it difficult for the minority class (diabetes-positive cases) to receive accurate predictions. While some studies have attempted to address this issue using techniques like oversampling or undersampling, there is a need for a comprehensive evaluation of different techniques and their impact on prediction accuracy. The proposed article also fills this gap by addressing the issue of imbalanced classes. It applies SMOTE to balance the dataset. This technique has been proven effective in addressing imbalanced classes and improving the accuracy of minority class prediction. This study evaluates how SMOTE affects the efficiency of various ML algorithms and how well it improves the accuracy of diabetes predictions.

(iii) The development and evaluation of practical applications for self-diagnosis and monitoring of diabetes is another gap in the literature. While some studies have proposed mobile apps or other tools, their effectiveness, usability, and adaptability to different datasets and populations require further investigation. To address this gap, the proposed article aims to develop a practical mobile app that allows users to enter their diabetes-related features and obtain instant predictions. The app will be designed to be user-friendly, easily accessible, and adaptable to different datasets and populations. The proposed system's adaptability will be evaluated using a domain adaptation method to ensure its effectiveness in different real-world scenarios.

By filling these gaps, the article hopes to advance the field of diabetes research and enhance early detection and prevention of diabetes, particularly in countries like Saudi Arabia with a high prevalence of the disease.

3. The Proposed Diabetes Prediction System

In this part, we will show the methodology that was used and the ML algorithms that were applied throughout the process of developing the proposed ML system for diabetes prediction. The sequences of the proposed system for predicting diabetes are illustrated in Figure 1. First, the dataset needs to be gathered and preprocessed to eliminate the necessary inconsistencies within it. For example, null occurrences were replaced with average values, and problems with unbalanced class sizes were addressed, among other things. The dataset was partitioned into two separate groups using the *holdout* process: the *test set* and the *training set*. After that, a number of various classification techniques were implemented to determine the one that performed the best in terms of accuracy regarding this dataset. Finally, the prediction model that has the highest performance is integrated into the structure of the mobile application that has been proposed.

3.1. Dataset Components. In this study, both private and Pima Indians datasets were used for ML classification (Pima Indians dataset is an open-source diabetes dataset that was initially gathered by the *National Institute of Diabetes and Digestive and Kidney Diseases*) [21]. The private dataset consists of 300 observations, while the Pima Indians dataset has 768 observations. Both the private dataset and the NI_DDKD dataset have eight features, as follows:

(1) Age: in years

(2) Glucose: the amount of glucose that is still present in the blood after two hours, typically referred to as the "2-hour postprandial blood sugar level"

(3) Insulin: the insulin test result that measures the level of insulin in a person's blood (μU/ml)

(4) Blood pressure: the BP test measures blood pressure against artery walls as it passes through the body (mm Hg)

(5) Pregnancies: the aggregate number of times that the woman has carried a pregnancy

(6) SkinThickness: the thickness of the triceps fold of skin (mm)

(7) BMI: (weight in kg)/(height in m)^2

(8) Diabetes pedigree function refers to a function that assigns a score to the chance of developing diabetes depending on the individual's family medical history

The *target variable* refers to "*outcome*" and presents a class variable that takes the value of 0 or 1 to indicate whether or not diabetes is present in the patient

Figure 2 shows the percentage of diabetes among Pima Indians participants. There are 768 records, and 268 of those individuals have been diagnosed with diabetes. The private dataset includes 300 participants (195 female and 105 male) aged 15–77. Table 1 shows the eight features of the Pima Indians and private datasets.

3.2. Dataset Preparation and Processing. The dataset used for this research is a collection of widely available Pima Indians and selected private datasets. We detected a few unexpected zero values in the combined dataset that we analyzed. For instance, both the BMI and thickness of the skin cannot be equal to 0. The mean value that corresponds to the zero value has been substituted in its place. With the use of the *holdout validation* method, the two datasets have been partitioned so that the training dataset consists of 75% of the data and the test dataset consists of 25% of the data.

The study uses the concept of *mutual information* which refers to any attempt to quantify the interdependence of different variables. It increases information acquisition, and larger numbers suggest a stronger dependence. Figure 3 shows a visual representation of the importance of each feature of a utilized dataset, which represents the mutual information of the many qualities that make up the dataset. For instance, based on the information shown in this figure, the *SkinThickness* for diabetes is less important than previously thought, when using this mutual information

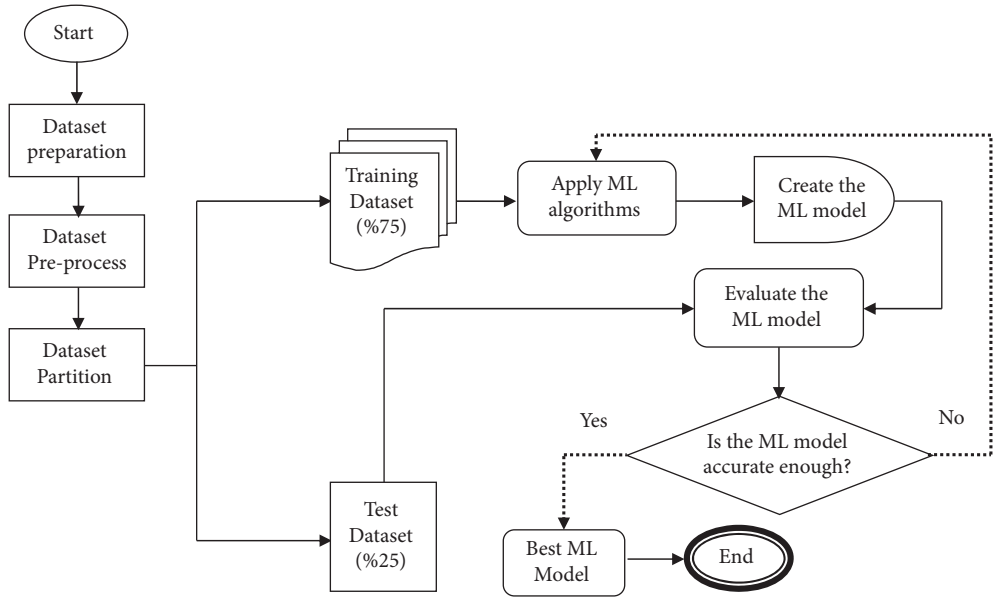

Figure 1: Sequences of the proposed approach for predicting diabetes.

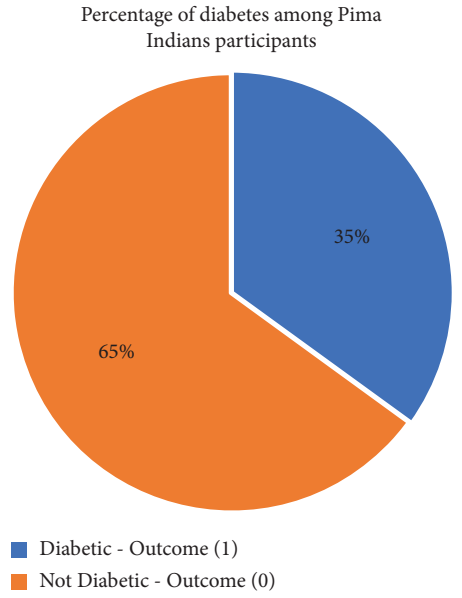

Figure 2: Percentage of diabetes among Pima Indians participants.

Table 1: The features of both Pima Indians and private datasets.

Features	Pima Indians dataset			Private dataset		
	Max	Min	Average	Max	Min	Average
Pregnancies	17	0	3.85	8	0	1.45
Glucose (mg/dl)	199	44	120.89	205	35	110.15
Blood pressure	122	40	69.11	165	55	69.11
SkinThickness	99	10	20.54	89	11	18.56
Insulin (μU/ml)	846	15	79.81	540	20	89.86
Body mass index (BMI)	67.1	19.1	31.99	77.3	44.5	40.45
Diabetes pedigree function	2.42	0.08	0.47	3.45	1.05	0.92
Age (year)	81	21	33.24	77	15	25.41

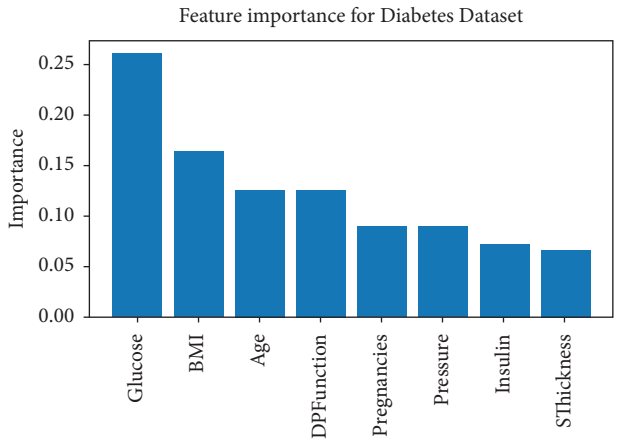

Figure 3: The importance of the features of the diabetes dataset.

approach. The comparison between the Pima Indians and a private dataset, including maximum, minimum, and average values, is shown in Table 1.

The proposed research uses the *Extreme Gradient Boosting* approach (XGB). It is considered a gradient-boosted decision tree ML library that is scalable and distributed and used for classification, regression, and ranking problems. Before the acquired dataset was combined with the Pima Indians dataset, the XGB regressor model was developed. In several publications, the prediction of missing values has been achieved using a variety of regression and ensemble learning methods [22].

Comprehensive research was conducted to identify the optimal method for predicting the insulin characteristics of the mentioned dataset. Three supervised ML-based methods, including support vector regression (SVR), Gaussian process regression (GPR), and XGB, were implemented in the mentioned datasets and used for predicting the results of interest (insulin levels in the tested samples). After that, we used the formula in (1) to calculate the RMSE of selected regression models (RMSE refers to *root mean square error* and is considered the standard deviation of the residuals (i.e., errors of prediction)):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Pi - Aj)^2}{n}}, \tag{1}$$

where n is the dataset validation sample count, P_i represents the predicted values, and A_j represents the actual values. As shown in Table 2, the GPR method has the smallest RMSE value of insulin on the dataset. As a result, predictions have been made on the insulin level of the mentioned dataset using the proposed model.

Due to specific characteristics and constraints of the data, a semisupervised learning model was used to predict insulin in the private dataset. Some possible reasons include the following:

Table 2: RMSE for different regression models applied to the diabetes dataset.

Regression model	Root mean square error (RMSE)
GPR	0.413
XGB	0.447
SVR	0.478

(1) Limited labeled data: Semisupervised learning works well with few labeled samples. Labeled data are used to train the model with known outcome variables, such as insulin levels.

(2) Cost-effectiveness: Labeling huge volumes of data, especially medical datasets, is expensive. Semi-supervised learning makes use of an abundance of unlabeled data, which are cheaper to annotate. Because of its low cost, it is attractive for private datasets.

(3) Use of unlabeled data: Unlabeled data often contain valuable information that can be used to improve model performance. By incorporating both labeled and unlabeled data, semisupervised models can leverage the underlying patterns and structures present in the unlabeled data to enhance their predictive capabilities. This is particularly relevant when working with complex medical datasets, where the relationship between features and the target variable may not be well understood.

(4) Privacy concerns: Private datasets may have severe privacy or confidentiality restrictions that limit labeled data access. Semisupervised learning allows models to generate accurate predictions while protecting sensitive data. Semisupervised models employ unlabeled data, lowering the danger of disclosing sensitive information.

Article [20] also provides some potential benefits for employing semisupervised learning in the insulin prediction domain. After using a semisupervised method to predict the features of insulin, we combined the two datasets mentioned to create a merged dataset. The combined dataset had 1068 records with all characteristics except the *SkinThickness* which was determined to be less important by *mutual information*. The combined dataset utilized in this research has 498 (268 + 230) diabetes samples and 570 (500 + 70) non-diabetic samples, both of which contribute to the problem of imbalance. The *Synthetic Minority Oversampling Technique* (SMOTE) has been used for training the dataset to solve the problem of imbalanced classes, but the testing dataset has been left unchanged. The min/max normalization method was also utilized in this study. Using the following equation, the data were transformed so that they fall within the same range:

$$X_{\mathrm{scaled}} = \frac{X - X_{\mathrm{min}}}{X_{\mathrm{max}} - X_{\mathrm{min}}}, \tag{2}$$

where $X_{\max}$ and $X_{\min}$ represent the highest and lowest possible scores in the feature column.

3.3. Machine Learning Algorithms. To implement the mobile diabetes prediction app, this study used 10 different ML and ensemble techniques, which are mentioned in Section 3.3. To prevent overfitting, the GridSearchCV framework has been utilized to determine the optimal values of various hyperparameters for all the ML techniques.

(i) A DT is a diagrammatic representation of the rule-based learning function. The DT learning method is an approach to the approximation of target functions with discrete values. Each node is selected using coefficients based on the *Gini* or *entropy* measures of information gain, which are written as follows:

$$\text{Gini} = 1 - \sum_{k=1}^{n} \left(\mathrm{P}_{i,k}\right)^2, \tag{3}$$

$$\text{Entropy} = \sum_{i=1}^{n} -\left(P_i \log_2 P_i\right). \tag{4}$$

In both (3) and (4), the value of n refers to the total number of unique class values. Using the *GridSearchCV* hyperparameter tuning, we were able to determine that the parameter's *maximum* depth is equal to 2 and the *minimum* sample leaf is equal to 50, and "Gini" impurity metrics perform effectively with the dataset that is being used in this study [23].

(ii) KNN is a supervised learning, nonparametric classifier in which an approximation function with discrete values can be achieved by using K numbers of the closest ML models. To categorize the data, it first generates a plane that contains all of the training points, then it measures the distance between the query and the plane, and finally, it produces a classification. This method identifies a certain number, K, of neighbors (determined by the dataset) and groups them according to the results of the majority vote. During our investigation, we utilized the $K = 5$ binary categorization [24].

(iii) RF is an ML technique that takes the predictions of many decision trees and creates an average using that information. As a consequence of this, RF has characteristics that make it suitable for use as an ensemble learning model. In this study, we utilized RF with estimators equal to 400, a minimum sample size of five for each leaf, and "Gini" impurity measures with hyperparameter adjustment [23].

(iv) *SVM* is a statistical method that does supervised classification by selecting the optimal hyperplane. In this investigation, we tried out a few different SVM kernels on the training dataset and compared their performance. The SVM algorithm with parameters *gamma* = 1, $C = 10$, *test_size* = 0.1, and *random_state* = 39 performed the best results [25].

(v) LR is a statistical technique that may be utilized to make predictions regarding binary classes. It best matches an *S*-shaped function, which may be used to forecast the result. The hyperparameter optimization approach was used to determine that the logistic regression model needed just 150 iterations to converge. This number was found to be sufficient [24].

(vi) AB is an example of a method for ensembles. This algorithm first operates on the main dataset, and then it adjusts subsequent copies of itself to the same dataset to achieve optimal performance. This framework modifies the weights of cases that were incorrectly categorized to direct the attention of succeeding classifiers more toward challenging situations. In this study, the AB algorithm was used with the *estimator* set to 50 and the *rate* set of learning to 0.10.

(vii) XGB is a gradient boosting-based ensemble ML algorithm that uses decision. The following are the settings that were utilized for the proposed XGB classifier: estimators' *maximum depth* = 4, the objective function was "*binary logistic*", test_size = 0.1, and random_state = 52 when applying the SMOTE to the training set [26].

(viii) A voting classifier is an ensemble approach that was developed to improve classification via the use of voting. This article presents the implementation of a voting classifier, which uses a voting hyperparameter referred to as "*soft*" to select the majority category predicted by each algorithm [23].

(ix) Bagging classifier is a type of ensemble classifier that works by first using the initial dataset to train a basic classifier on a sample of the data and then combining the individual predictions of the base classifiers to get a final classification based on the results of the voting process. As examples of different hyperparameters, the implemented bagging classifier makes use of base *estimators* equal to 500, a *maximum* number of samples equal to 120, and an *out-of-bag* value equal to "*True*" [27].

(x) NB is an ML algorithm used for classification. It uses the Bayes theorem as its foundation, with the assumption that features are conditionally independent after the class label has been known. Due to the assumption's simplicity, the method is quick and can be extended to high-dimensional data. When it comes to classification tasks, particularly the classification of texts and spam filtering, NB is a straightforward and reliable method. It is robust to irrelevant features and can handle missing data. The following are the settings that were utilized for the NB classifier:

TestSize = 0.2, RandomState = 38 for testing data, and TestSize = 0.2, RandomState = 53 for implementing the SMOTE on the dataset used for training.

3.4. Deployment of the Proposed System. ML algorithms (classifiers) served as the foundation for the proposed system that has been implemented into a mobile application framework so that it can function instantly on actual data for predicting diabetes. We developed the interface of the requested application using Android Studio, J2ME, PHP, MySQL, HTML, XML, and CSS. This study selected the XGB ML model with the SMOTE as a final choice because it offered the highest level of performance and accuracy (see Table 3). The model has been deployed using many integrated development environments (IDEs), such as the Python environment platform and Spyder. We also developed a mobile app to test the functionality of the prediction system, which allowed us to show the system's capabilities in real time. The user interface of this application is built with Android Studio. We used Java as the primary language for programming. A subsequent step involved integrating the pickle package into Android Studio to actualize the model. We used Heroku as the hosting server for the proposed system when creating the application programming interface (API). Figure 4 shows the flowchart outlining the process for designing the proposed ML-based diabetes prediction app. The proposed app has been deployed into two subsystems, namely, the website app (on the left) and the mobile app (on the right).

4. Results and Discussion

In this section, the findings of the proposed diabetes prediction app are presented, along with an explanation of the system. Firstly, the effectiveness of a wide variety of ML techniques was evaluated. After that, a demonstration of the website framework that has been created as well as an Android Mobile application follows. When evaluating the various ML models, we looked at their classification accuracy. These measures can be described by the following formulas:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1\ \text{score} &= \frac{2 \times \text{Recall} \times \text{Percision}}{\text{Recall} + \text{Percision}}, \end{aligned} \tag{5}$$

where TP means that the system is correctly predicting a positive value and that the value itself is likewise positive. FP means that the model made a positive prediction, but the actual outcome was negative. TN means that the system is making a false prediction, and the outcome also confirms this prediction. FN means that the model is predicting a negative value, whereas the actual value is positive. All ML models were validated using the *holdout validation* method with a categorized 7 : 3 train-test partition.

Table 3: Metrics for the performance of 10 ML algorithms using the SMOTE.

Ser.	Classifier	Accuracy (%)	Precision	Recall	F1 score
1	Logistic regression	77.0	0.63	0.70	0.66
2	Random forest	75.0	0.60	0.66	0.63
3	KNN	64.0	0.49	0.73	0.59
4	Decision tree	72.0	0.79	0.77	0.78
5	Bagging	75.0	0.81	0.81	0.81
6	AdaBoost	73.0	0.80	0.77	0.78
7	XGBoost	83.1	0.70	0.84	0.76
8	Voting	75.0	0.83	0.76	0.79
9	SVM	77.0	0.87	0.77	0.82
10	Naive Bayes	81.2	0.73	0.71	0.72

Figure 5 shows the confusion matrix for XGBoost with the SMOTE technique. According to this figure, the XGBoost algorithm correctly classified 300 instances with TP = 39 and TN = 261.

Finally, to better understand the model's decision-making process, an explainable AI approach using SHAP libraries is implemented. Figure 6 illustrates the importance of the XGBoost with SMOTE feature using the SHAP library.

Table 3 presents a comparison of the various performance measures for ten ML classifiers when applied to the combined datasets using the SMOTE. Table 3 shows that the XGB classifier had the greatest overall performance, as seen by its accuracy of 83.1% as well as its F1 score = 0.76 and AUC = 85.31% (see Figure 7). The KNN classifier, on the other hand, obtained the min accuracy and F1 score.

Next, the domain adaptation technique was utilized for the testing and training of the ML model on source and target datasets. In this study, the proposed system for the prediction of diabetes is initially trained using the open-source dataset containing a higher number of Pima Indians. After that, the system is tested on a private dataset that has a much reduced dimension. The key performance metrics for the private dataset are displayed in Table 4. In this instance, it is worth noting that XGB with the SMOTE was used for the training dataset. The proposed approach has a 97.4% accuracy rate, an F1 coefficient of 0.95, and an AUC of 0.87.

Finally, the proposed system was implemented as a mobile app utilizing XGB and SMOTE. Figure 8 is a representation of an immediate diagnosis of diabetes provided by the mobile app that was developed with the assistance of real data, which was developed with the help of the most effective classification (i.e., with XGB).

This research aims to use ML classifiers for predicting diabetes disease. Based on the experimental results, XGB regression yielded the most accurate predictions of insulin with the smallest RMSE. Based on the mutual information-based feature selection method, the most important characteristics for diabetes prediction are glucose level, BMI, age, and insulin. Methods for optimizing hyperparameters and oversampling with synthetic data, such as the SMOTE, have

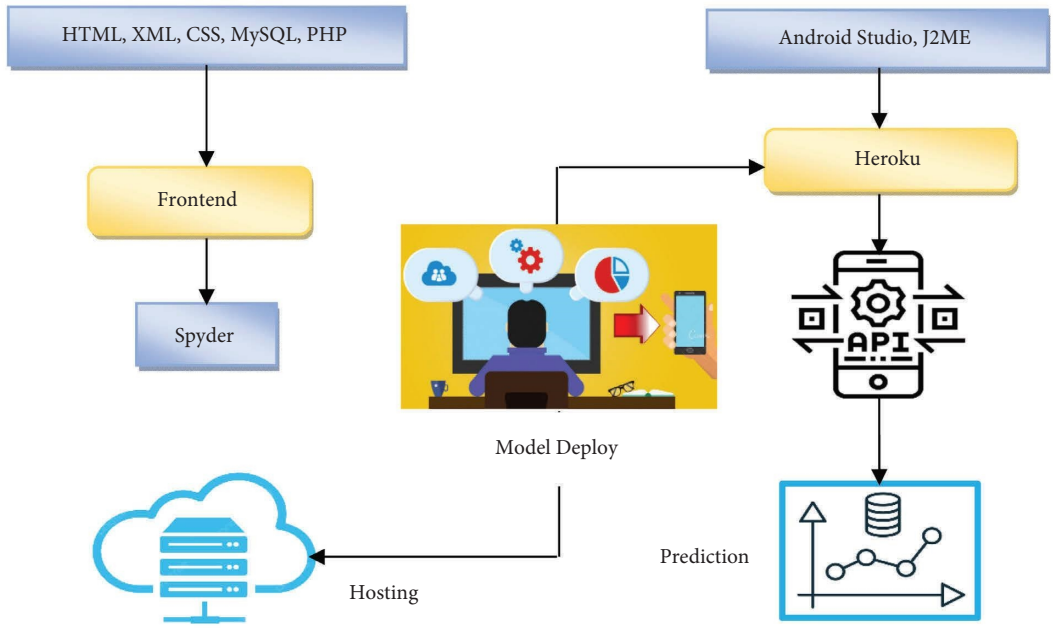

Figure 4: The process for designing the proposed ML-based diabetes prediction app.

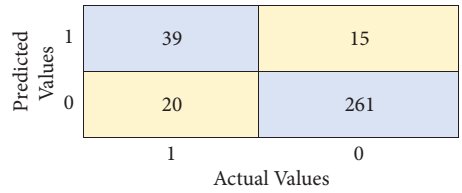

Figure 5: Confusion matrix for XGBoost with SMOTE technique.

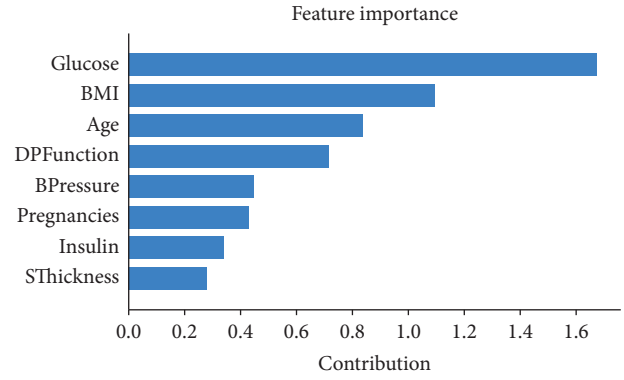

Figure 6: Interpretation of explainable AI of feature importance of XGBoost with SMOTE.

been implemented. The best results were obtained using the XGB method with the SMOTE. The average variance for classification accuracy in the insulin prediction was achieved by applying the XGB regression model and generated by taking the average and median values of the mentioned datasets, which were around 1.33% and 2.33%, respectively.

5. Limitations

Despite the promising results and potential of the proposed technique for predicting diabetes disease through a mobile app using machine learning, several limitations need to be acknowledged:

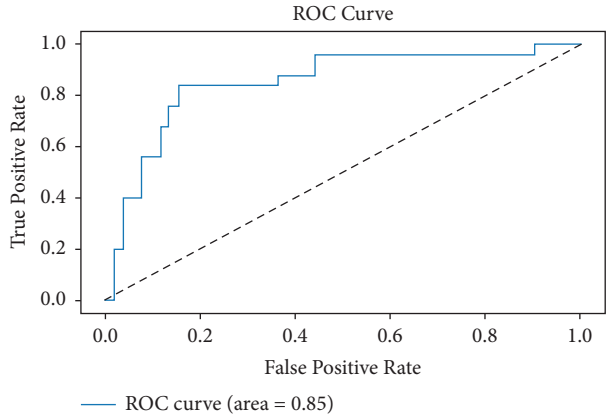

Figure 7: ROC curve and AUC for XGB using the SMOTE.

Table 4: The best performance metrics of the private dataset using the SMOTE.

Accuracy (%)	Precision	F1 score	Recall
97.4	0.96	0.95	0.97

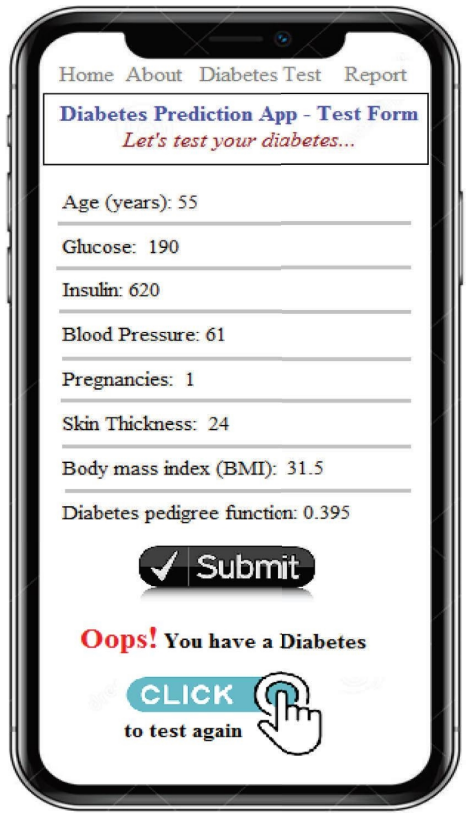

Figure 8: Immediate prediction of diabetes by the proposed mobile app.

(1) *Dataset availability and quality*: The availability and quality of the dataset employed for training and testing have a significant impact on the accuracy and performance of any ML model. In this study, the researchers used the Pima Indians dataset and private diabetes datasets. However, the availability and representativeness of these datasets may be limited, and the quality of the data could vary. This limitation could affect the generalizability of the proposed technique to a larger population and different data sources.

(2) *Imbalanced classes*: The presence of imbalanced classes, where one class is significantly more prevalent than the other, can affect the performance of ML algorithms. In this study, the researchers addressed this issue by using the SMOTE technique. While SMOTE helps in balancing the classes, it is not a perfect solution and may present synthetic samples that do not precisely represent the minority class. This limitation could lead to biased predictions and lower accuracy in real-world scenarios.

(3) *Algorithm selection*: The researchers applied several ML classification techniques to determine the best algorithm for diabetes prediction. However, the choice of algorithms is subjective and could impact the results. There might be other algorithms not considered in this study that could potentially achieve better accuracy or different trade-offs. Therefore, the selection of ML algorithms should be carefully considered and evaluated in future research.

(4) *Domain adaptation*: The proposed system's adaptability was demonstrated through the application of domain adaptation methods. However, the generalization of the proposed technique to different populations or settings may still be limited. The effectiveness of the technique in diverse populations with varying demographics, lifestyles, and healthcare systems needs to be further investigated. Additionally, the potential challenges and limitations associated with domain adaptation should be thoroughly addressed.

(5) *Mobile app usability and acceptance*: The development of a mobile app for users to enter features and predict diabetes instantly is a significant contribution of this study. The success of the proposed technique ultimately relies on user engagement and adoption of the mobile app. So, it is essential in future work to evaluate key factors such as user experience, privacy concerns, and accessibility to ensure the app's effectiveness in a real-world setting.

6. Conclusions

Both life expectancy and quality may be decreased by diabetes. Early detection of this chronic ailment has the potential to lessen the severity of numerous diseases and their repercussions. In this research, we have introduced the implementation of a mobile-based model that can automatically predict a person's risk of acquiring diabetes using a range of ML techniques. This app was developed as part of this research and applied to the available Pima Indians and a selected private dataset. In this study, multiple ML and ensemble algorithms were evaluated based on their accuracy. The XGB algorithm had the highest level of performance using the SMOTE, with an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87 for the private dataset and an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85 for the combined datasets. To show the proposed system's adaptability, the domain adaptation method was applied. Following that, the approach of domain adaptation was used to illustrate the flexibility of the proposed prediction app. Finally, a mobile app has been developed to allow users to enter features and predict diabetes instantly. In conclusion, the best performing XGB method has been implemented into a mobile app to predict diabetes. There are several potential extensions to the scope of our work, such as our suggestion that more confidential data be collected from a greater number of patients to get more accurate results. While the proposed technique using ML for the prediction of diabetes disease through a mobile app shows promise, it is important to acknowledge the limitations outlined in the previous section. Future research should aim to address these limitations to improve the accuracy, generalizability, and usability of the proposed technique in real-world healthcare settings.

Data Availability

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

Ethical Approval

This study received ethical approval on March 16, 2023, from the Deanship of Scientific Research at King Khalid University. This study was conducted under the ethical guidelines and regulations of KKU.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hosam El-Sofany was responsible for conceptualization, methodology, design, implementation, writing, reviewing, editing, proofreading, and checking against plagiarism using the iThenticate program provided by King Khalid University. Samir A. El-Seoud and Omar H. Karam were responsible for methodology, writing, reviewing, and proofreading. Yasser M. Abd El-Latif was responsible for methodology, reviewing, and editing. Islam A. T. F. Taj-Eddin was responsible for the revision.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through a large group research project under grant no. RGP2/299/44.

References

[1] G. Atlas, "Diabetes. International diabetes federation," *IDF Diabetes Atlas*, International Diabetes Federation, Brussels, Belgium, 10th edition, 2021.

[2] S. Akhtar, J. A. Nasir, A. Sarwar et al., "Prevalence of diabetes and pre-diabetes in Bangladesh: a systematic review and meta-analysis," *BMJ Open*, vol. 10, no. 9, Article ID e036086, 2020.
[3] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki et al., "A novel diabetes healthcare disease prediction framework using machine learning techniques," *Journal of Healthcare Engineering*, vol. 2022, Article ID 1684017, 10 pages, 2022.
[4] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, "Detection and prediction of diabetes using data mining: a comprehensive review," *IEEE Access*, vol. 9, pp. 43711–43735, 2021.
[5] K. J. Rani, "Diabetes prediction using machine learning," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 6, pp. 294–305, 2020.
[6] 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.
[7] 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, 2020.
[8] S. Albahli, "Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 5, pp. 1069–1075, 2020.
[9] 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, pp. 7–14, 2020.
[10] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, no. 1, Article ID 11981, 2020.
[11] 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.
[12] N. Ahmed, R. Ahammed, M. M. Islam et al., "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.
[13] A. Aada and S. Tiwari, "Predicting diabetes in medical datasets using machine learning techniques," *International Journal of Scientific Research and Engineering Trends*, vol. 5, no. 2, pp. 257–267, 2019.
[14] G. Bhola, A. Garg, and M. Kumari, "Comparative study of machine learning techniques for chronic disease prognosis," *Computer Networks and Inventive Communication Technologies*, vol. 58, pp. 131–144, 2021.
[15] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.
[16] S. Kodama, K. Fujihara, C. Horikawa et al., "Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: a meta-analysis," *Journal of Diabetes Investigation*, vol. 13, no. 5, pp. 900–908, 2022.
[17] Z. Q. Zhang, L. Q. Yang, W. T. Han et al., "Machine learning prediction models for gestational diabetes mellitus: meta-analysis," *Journal of Medical Internet Research*, vol. 24, no. 3, Article ID e26634, 2022.
[18] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, Article ID 100032, 2021.
[19] P. Palimkar, R. N Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random forest Classifier Approach," *Advanced Computing and Intelligent Technologies*, Springer, Singaporepp. 219–224, 2022.
[20] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, pp. 1–10, 2023.
[21] G. Atlas, "Diabetes. International diabetes federation," 2017, http://www.diabetesatlas.org.
[22] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, no. 12, 2022.
[23] G. Aurélien, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Inc, Sebastopol, CA, USA, 2021.
[24] T. M. Mitchell, *Machine Learning*, McGraw-Hill, Inc, New York, NY, USA, 2021.
[25] S. P. Chatrati, G. Hossain, and A. Goyal, "Smart home health monitoring system for predicting type 2 diabetes and hypertension," *Journal of King Saud University*, vol. 34, no. 3, pp. 862–870, 2020.
[26] 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.
[27] J. Cervantes, F. García-Lamont, L. Rodríguez, and A. Lopez-Chau, "A comprehensive survey on support vector machine classification: applications, challenges, and trends," *Neurocomputing*, vol. 408, pp. 189–221, 2020.