SciNexuses

Journal Homepage: sciencesforce.com/scin

SciNexuses Vol. 1 (2024) 28–43

Paper Type: Original Article

Diabetes Prediction using Machine Learning and Explainable Artificial Intelligence Techniques

Khaled Elmenshawy [1,*], Nada Wael [2], Rana Ahmed [2] and Ahmed A. El-Douh [2]

[1] Computer Science, Midocean University; khaled@midocean.ae.

[2] Faculty of Information Systems and Computer Science, October 6th University, Giza, 12585, Egypt;
Emails: nadawael3539@gmail.com; ra3936299@gmail.com; ahmed.eldouh.csis@o6u.edu.eg.

Received: 18 Feb 2024 **Revised:** 02 Apr 2024 **Accepted:** 01 May 2024 **Published:** 03 May 2024

Abstract

Diabetes influences 537 million human beings globally and may result in diverse fitness issues, inclusive of coronary heart disease, kidney disease, nerve damage, and diabetic retinopathy. A new diabetes forecast framework has evolved with a non-public Bangladeshi dataset and diverse AI methods. The version makes use of a semi-controlled version with excessive inclination aid to expect insulin tires and makes use of algorithms like Decision Tree, SVM, Random Forest, logistic regression, KNN, and different organization methods. After getting ready and checking out all the older models, the proposed framework gave satisfactory results inside the XGBoost classifier with the ADASYN method with 80% accuracy,0.81 F1 coefficient, and an AUC of 0.84, with a 99.3% accuracy completed the use of a mixture of 3 classifiers (Stack). The version additionally makes use of area variant strategies to illustrate its flexibility. The source code is publicly accessible at https://github.com/diabetes_prediction.

Keywords: Diabetes Prediction, Machine Learning, XGBoost, Dataset Collection, Classification Algorithm.

1 | Introduction

Diabetes is a significant public health challenge as it is a chronic condition that affects tens of thousands of people worldwide. Prospective avenues for identifying individuals at high risk of developing diabetes are provided by predictive modeling, which is powered by advancements in machine learning and healthcare data analytics [1-3]. Enhancing the early detection and analysis of diabetes, a common metabolic infection that affects a significant portion of the global population is the primary objective of their research. In addition to Logistic Regression (LR), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Decision Tree (DT), the authors have employed several machine learning methods to classify individuals diagnosed with diabetes. Through the application of those sophisticated algorithms, the researchers were able to improve a prediction model that would be able to identify individuals who are at risk of developing diabetes based solely on relevant clinical data [4-5].

Due to diabetes's possible side effects, which include kidney damage, nerve damage, coronary heart disease, vision abnormalities, and other chronic illnesses, the observation emphasizes the significance of early diabetes detection [6]. The scientists have also emphasized how important it is to investigate and find more gadget-

Corresponding Author: khaled@midocean.ae

https://doi.org/10.61356/j.scin.2024.1306

learning algorithms to further improve the precision and efficacy of diabetes prediction models. The researcher aims to advance healthcare practices and early intervention in the management of diabetes by continuously improving the predicting capabilities of those algorithms [7]. Their paintings highlight the need for early detection in reducing the harmful effects of diabetes and aim to further the development of more reliable and environmentally friendly diagnostic equipment in the healthcare sector. The use of system mastering techniques allowed the article to forecast diabetes with an accuracy of 85%. [8] To develop an automated diabetes prediction machine using gadget learning and explainable AI strategies, the authors of the study "Diabetes prediction using machine learning and explainable AI techniques" conducted a thorough investigation.

The goal of the study was to improve the precision and efficacy of diabetes prediction by combining the well-known Pima Indian dataset with a private dataset from female employees of a Bangladeshi textile company [9].
The primary objective of the test adapted to handle the challenging circumstances surrounding the early detection of diabetes and improve the overall control of the illness with the use of cutting-edge technology methods.
The writers employed multiple techniques and methodologies to achieve their study aims. The data was divided into training and testing sets using the holdout validation technique, which made it easier to evaluate various devices and learn entirely based on category algorithms. One of the examination's major contributions was the development of a unique dataset of 203 diabetes mellitus samples from Bangladeshi women [10-12].

Overall, the research conducted by the authors represents a significant advancement in the field of diabetes prediction, merging real-time programs, explainable AI tactics, and gadget-mastering algorithms to improve the accessibility, efficiency, and accuracy of diabetes analysis and care. Their artwork provides invaluable perspectives and tools for medical professionals and individuals seeking to manage the challenging circumstances brought on by diabetes using advanced technological solutions. The XGBoost classifier finished with 81% accuracy in predicting diabetes, with an F1 rating of 0. To improve a diabetes prediction version of the employment of gadget-mastering strategies, the authors of [13] conducted a thorough observation. By using a new dataset and a pipeline version for diabetes prediction, the study aims to overcome the limitations of existing approaches by providing a unique approach that improves category accuracy. The authors have employed a variety of strategies and methodologies in their research to achieve this goal. Huge records analytics have been used by the authors to examine large datasets, find hidden patterns, and derive insightful information that will improve the accuracy of diabetes prediction. Additionally, the have-a-look observation involves the application of predictive assessment procedures, which blend statistical methods, device reading algorithms, and records mining techniques to infer future activities primarily relying on historical and current records. In conclusion, the authors have conducted a study that focuses on developing a diabetes prediction version through the application of big data analytics, system learning algorithms, and predictive assessment approaches.

Their objective is to improve diabetes typing accuracy by imposing a pipeline version for more accurate predictions and integrating external factors. The author hopes to advance healthcare practices in diabetes diagnosis and treatment by utilizing cutting-edge computational techniques. The accuracy of the [14] study is 96% when using Logistic Regression and 98% when using other algorithms. The sensing node gathers information on physiologically affected individuals, the combination and visualization node compiles and presents the data for medical specialists, and the prediction node employs system learning techniques to forecast the risk of diabetes.

A consistent cloud-based system aggregates data from several devices, such as blood pressure monitors, cellphones, smartwatches, and glucometers, to provide clinical specialists with a comprehensive picture of the affected person's lifestyle parameters. It enables clinical professionals to make informed decisions based on lifestyle insights and diabetes risk forecasts by enabling the real-time viewing of patient statistics on an online

portal. Their framework's overall goals are to provide a change technique for astute healthcare tracking and control for continuous diabetes by utilizing predictions from device mastery, ensuring interoperability, lowering costs, and providing patients with discreet tracking.

It is an all-in-one solution for early diabetes danger identification and control thanks to the integration of multiple medical devices and its user-friendly features. The study's support vector machine model for predicting diabetes risk has an accuracy of 83.20%. The authors of [15] conducted a thorough analysis of diabetes prediction using gadget-mastering techniques and have developed a mobile application for real-time diabetes prediction.
The authors hope to contribute new ideas and methods to the field of machine learning-based diabetes prediction to aid in the early detection and management of the disease, particularly in countries like Saudi Arabia where the disease is highly prevalent.

In this paper, we can use a STACK method among 3 classifiers which gave us greater accuracy we can use the Pima Indian facts set merged with RTML non-public dataset with the usage of a merge method like SMOTE or ADYSN and we can see the result will boom or decrease. The significant contribution of this work is as follows:

- Dataset Introduction: Introduced a novel diabetes dataset, named 'RTML dataset', comprising 203 samples from female employees at Rownak Textile Mills Ltd, featuring six attributes: pregnancy, glucose, blood pressure, skin thickness, BMI, age, and diabetes outcome.
- Feature Alignment: Aligned the RTML dataset's features with those of the Pima Indian dataset and utilized a semi-supervised method to estimate the absent insulin data. Class Imbalance Resolution: Applied SMOTE and ADASYN techniques to address class imbalance and conducted hyperparameter optimization.
- Model Interpretability: Employed Explainable AI methods using SHAP and LIME to elucidate the model's decision-making process and identify the most influential predictive features.
- Accuracy Enhancement: Enhanced prediction accuracy to 99.3% by integrating three classifiers—bagging, logistic regression, and decision tree—using a stacking approach.

The following paragraph is a breakdown of the paper's structure. The introduction is in section one. The literature review is presented in section two with the table required. The proposed system has been discussed and illustrated in section three with suitable figures and charts. The result of the research is presented in section four with the five papers comparison table. Section five concludes the paper. Finally, Section Six contains the references.

2 | Literature Review

Jeevan, Y. [16] they have a look at exploring using category algorithms (along with Naïve Bayes, Decision Tree, and SVM) for predicting diabetes. Early detection and analysis of diabetes are vital because of its negative fitness outcomes and complications. Data technological know-how and gadget studying play a tremendous position in healthcare with the aid of enhancing diagnostic accuracy and exploring diseases. Various gadget-studying algorithms provide excessive accuracy in predicting results primarily based totally on entered information and statistical analysis. Previous research has hired regression-primarily based total strategies, help vector machines, grasping strategies, and classifiers like Decision Trees, Artificial Neural Networks, and Random Forests for diabetes prediction.

Tasin, I. [17] they have a look at creating an automatic diabetes prediction machine with the usage of gadget studying and explainable AI strategies. Researchers applied a personal dataset of lady sufferers in Bangladesh at the side of the Pima Indian diabetes dataset. They implemented characteristic choice algorithms, addressed magnificence imbalance in the usage of SMOTE and ADASYN strategies, and examined numerous category algorithms along with the choice tree, SVM, Random Forest, Logistic Regression, and KNN. The surest

overall performance became done with the XGBoost classifier and the usage of the ADASYN method, reaching eighty percent accuracies, a 0.81 F1 score, and an AUC of 0.84. Furthermore, the researchers integrated explainable AI strategies like LIME and SHAP to interpret the version`s predictions. The studies brought a personal dataset from a fabric enterprise in Bangladesh and harassed the importance of early and unique diabetes detection to mitigate complications.

Mujumdar, A. [18] they have a look at objectives to broaden a sturdy diabetes prediction version through the usage of gadget studying algorithms and massive information analytics. It addresses barriers of present strategies with the aid of incorporating each outside element (at the side of normal elements like Glucose, BMI, Age, and Insulin) for higher categories. The proposed version objectives to reinforce category accuracy as compared to preceding datasets. A pipeline version is carried out for diabetes prediction, in addition to enhancing category accuracy. By leveraging superior strategies and outside elements, they have a look at complementing the accuracy of the diabetes category, reaping benefits for healthcare experts in diagnosing and treating diabetes extra effectively.

Ramesh, J. [19] the study utilized a support vector machine (SVM) model to predict diabetes risk based on the Pima Indian Diabetes Database. They enhanced the dataset features, validated the model with cross-validation, and evaluated its performance using metrics such as accuracy, sensitivity, and specificity. The framework was designed to automate diabetes detection and notify healthcare providers promptly, integrating various healthcare and consumer devices to enhance diagnostic decision-making. El-Sofany, H. [20] the research article aims to develop a machine-learning system for predicting diabetes, specifically in Saudi Arabia where diabetes prevalence is on the rise. By applying ML techniques to datasets like the Pima Indians dataset, the study creates a digital tool for diabetes prediction. Noteworthy aspects include using a semi-supervised model with gradient boosting, effectively addressing imbalanced data using SMOTE, optimizing hyperparameters, assessing various ML algorithms, and designing a user-friendly mobile application for immediate diabetes risk assessment. The system seeks to enhance early detection and control of diabetes, potentially improving healthcare outcomes and preventing long-lasting complications linked to the condition.

Table 1. Comparison of related work.

Domain	strengths	weakness
Healthcare and Diabetes Management Data Science and Machine Learning Public Health and Disease Prevention	Comprehensive Dataset Advanced Techniques Model Evaluation Interdisciplinary Approach	limited external factors. dataset size. imputation of missing values. clustering technique. evaluation metrics. generalizability.
The domain of this research paper is focused on Healthcare and Machine Learning.	It enhances accessibility, promotes early detection of diabetes, and provides a convenient tool for monitoring diabetes risk	dataset availability and quality. imbalanced classes. algorithm selection. domain adaptation. mobile app usability and acceptance
Healthcare Technology Machine Learning, Predictive Analytics	Unique Dataset Feature Engineering Explainable AI Application Development	limited dataset missing data imputation class imbalance handling model interpretability evaluation metrics external validation future scope
The paper explores machine learning and data mining techniques for predicting and analyzing diabetes, highlighting their potential in healthcare and the intersection of	The paper presents a deep learning method for predicting diabetic blood glucose levels, utilizing machine learning algorithms and cloud-based encrypted data for	there remains an opportunity for improvement by exploring additional machine learning algorithms to enhance accuracy and results.
Healthcare Technology and Machine Learning in Diabetes Management.	Integration of multiple healthcare devices, implementation of machine learning for risk prediction, real-time alerts for medical professionals, addressing vendor interoperability, and user-friendly interface.	narrow dataset scope small-scale testing lack of longitudinal study interoperability concerns limited patient interaction

Work Year	Authors	methodology	Result	Phase
Aishwarya, 2019, [12]	Aishwarya Mujumdar Dr. Vaidehi	Data Collection. Feature Selection. Model Development. Evaluation. Comparison. Implementation of Pipeline Model. Utilization of Big Data Analytics.	Machine learning algorithms were used to develop a diabetes prediction model using the pre-processed data and clustering outcomes. The model's accuracy and effectiveness in predicting diabetes were evaluated using specific metrics.	Data Collection and Pre-processing Clustering Analysis Model Development Evaluation and Validation
Hosam El-Sofany [13] "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App" 2023	Hossam El-Sofany Samir A. El-should Omar H. Karam Yasser M. Abd El-Latif Islam A. T. F. Taj-Eddin	Data Gathering and Pre-processing Dataset Partitioning ML Algorithm Selection Dataset Components	The research developed a mobile-based model for predicting diabetes risk using machine learning, achieving 97.4% accuracy on private and combined datasets, thereby enhancing early diabetes detection and prediction. Select keywords: 64 words.	Development of a mobile app for instant diabetes prediction
Isfafuzzaman Tasin [14] "Diabetes prediction using machine learning and explainable AI techniques" 2023	Isfafuzzaman Tasin' Tansin Ullah Nabil Sanjida Islam Riasat Khan	Dataset Collection and Preprocessing. Model Development and Evaluation. Deployment of Prediction System. Performance Evaluation and Comparison. Website and Smartphone Application	The study found that the XGBoost classifier with the ADASYN approach was the best, achieving 81% accuracy and an AUC of 0.84. The bagging classifier had the best overall performance, while the decision tree approach had the lowest accuracy and F1 score.	The research paper focuses on creating an automatic diabetes prediction system using machine learning and AI techniques, involving dataset collection, preprocessing, model training, evaluation, and deployment.
D. Indira [15] "Prediction of Diabetes using Machine Learning"2019	Dr. Y. Jeevan Nagendra Kumar Ms. Nistala Kameswari Shalini	KNN SVM Logistic Regression Random Forest Decision Tree Gradient Boosting	The study reveals that the KNN algorithm outperforms other classification algorithms in accurately diagnosing individuals with diabetes, achieving an 85% accuracy rate.	analyzing the performance of different machine learning algorithms on a diabetes dataset. The goal was to determine the most effective algorithm for accurately classifying people diagnosed with diabetes.
Jayroop Ramesh [16] "A remote healthcare monitoring framework for diabetes prediction using machine learning" 2021	Jayroop Ramesh. Raafat Aburukba. Assim Sagahyroon.	SVM for diabetes prediction. Enhanced dataset features. Validated with cross-validation. Evaluated using performance metrics.	The SVM-RBF model demonstrated impressive accuracy, sensitivity, and specificity in diabetes prediction, reducing patient interaction and paving the way for future enhancements in healthcare devices.	Developing a Remote Monitoring Framework for Diabetes Prediction.

3 | Proposed System

This section describes the design of the proposed autonomous diabetes prediction system, including the application and operation of various machine-learning techniques. Figure 1 shows the several stages of this research effort. The dataset was first collected and preprocessed to remove pertinent discrepancies, like addressing imbalanced class problems and replacing null occurrences with mean values. Following that, the dataset's training and test sets were the holdout validation method. To determine which classification algorithm will work best for this dataset, many classification algorithms were then used. Ultimately, the optimal prediction model is implemented within the suggested website and mobile application architecture.

3.1 | Dataset

The Pima Indian dataset is an open-source dataset [21] that is publicly available for machine learning classification, which has been used in this work along with a private dataset. It contains 768 patients' data, and 268 of them have developed diabetes. Figure 2 shows the ratio of people having diabetes in the Pima Indian dataset. Table 1 demonstrates the eight features of the open-source Piman Indian dataset. RTML private dataset: A significant contribution of this work is to present a private dataset from Rownak Textile Mills Ltd, Dhaka, Bangladesh, referred to as RTML, to the scientific community.

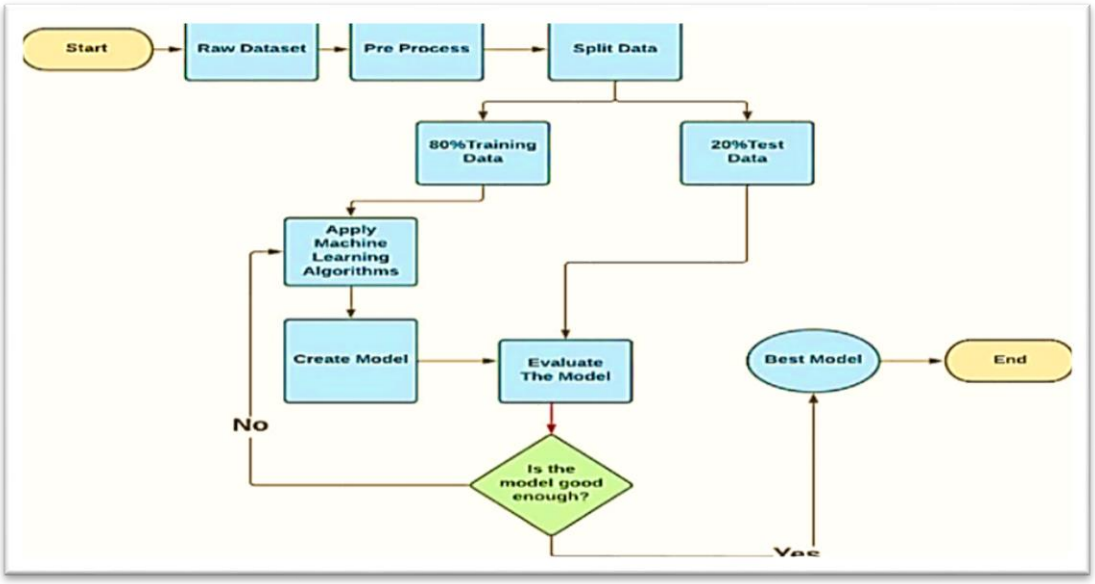

Figure 1. Working sequences of the proposed Diabetes prediction.

Table 2. Features of the Pima Indian Dataset.

Pregnancies	Skin thickness	Diabetes pedigree function
Glucose	Insulin	Age
Blood pressure	BMI	

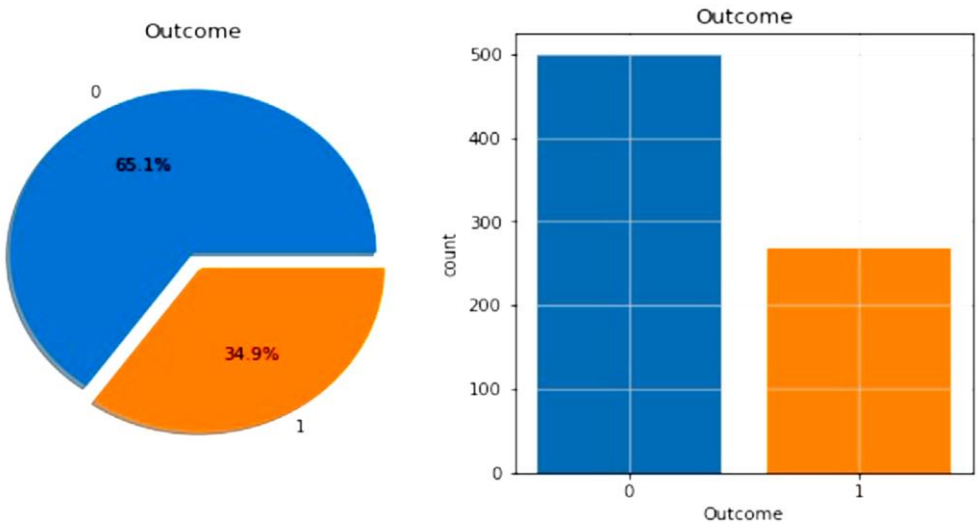

Figure 2. Percentage of people Having Diabetes in the Pima Indian.

Table 3. Features of the RTML Private Dataset.

Features	Minimum	Maximum	Average
Pregnancies	0	8	1.61
Glucose(mg/dL)	52.2	274	109.39
Blood pressure (mm Hg)	5.9	115	71.09
Skin thickness (mm)	2.9	23.3	10.78
BMI (kg/m2)	2.61	41.62	22.69
Age (years)	17	77	27.02

Female volunteers gave their free consent to take part in the research. Six characteristics are included in this dataset: age, BMI, skin thickness, pregnancy, blood pressure, glucose, and the result of diabetes in 203 female people between the ages of 18 and 77. The Glu-co-leader Enhance blood sugar meter was used in this investigation to monitor blood glucose. The participant's blood pressure and skin thickness were measured using an OMRON HEM-7156T blood pressure monitor and a digital LCD body fat caliper, respectively. Table 2 presents unique attributes of the private RTML dataset together with their average, maximum, and minimum values.

3.2 | Dataset Preprocessing

We identified a few unusual zero values in the combined dataset. For instance, a zero Body Mass Index (BMI) and skin thickness are not possible. Its matching mean value has taken the place of the zero value. Using the holdout validation technique, the training and test datasets were divided, with 80% of the data being the training and 20% being the test. Mutual Information: Mutual information looks for ways to quantify how dependent variables are on one another. Higher values suggest greater dependency, and it generates information gain [22]. The significance of each feature in this dataset is displayed in Figure 3 together with the mutual information between the different features.

For instance, based on this mutual information technique, the diabetes pedigree function appears to be less significant, as seen in this image. Semi-supervised learning: A combined dataset has been used in this work by incorporating the open-source Pima Indian and private RTML datasets. According to Table 2, the RTML dataset does not contain the insulin feature, which is predicted using a semi-supervised approach. Before merging the collected dataset with the Pima Indian dataset, a model was created using the extreme gradient boosting technique (XGB regressor). Various regression and ensemble learning techniques have been successfully extensively employed to forecast missing values in numerous papers [23, 24]. The best-performing regressor strategy to predict the insulin feature of the RTML dataset from the Pima Indian dataset was selected after thorough research. The Pima Indian dataset was first utilized to choose the optimal regression model since the RTML dataset did not contain the real insulin value. To predict the chosen outcome, which is the insulin of the validation samples of the Pima Indian dataset, three supervised regression models were used: extreme gradient boosting technique (XGB), support vector regression (SVR), and Gaussian process regression (GPR). First, the Pima Indian dataset was divided into an 8:2 ratio.

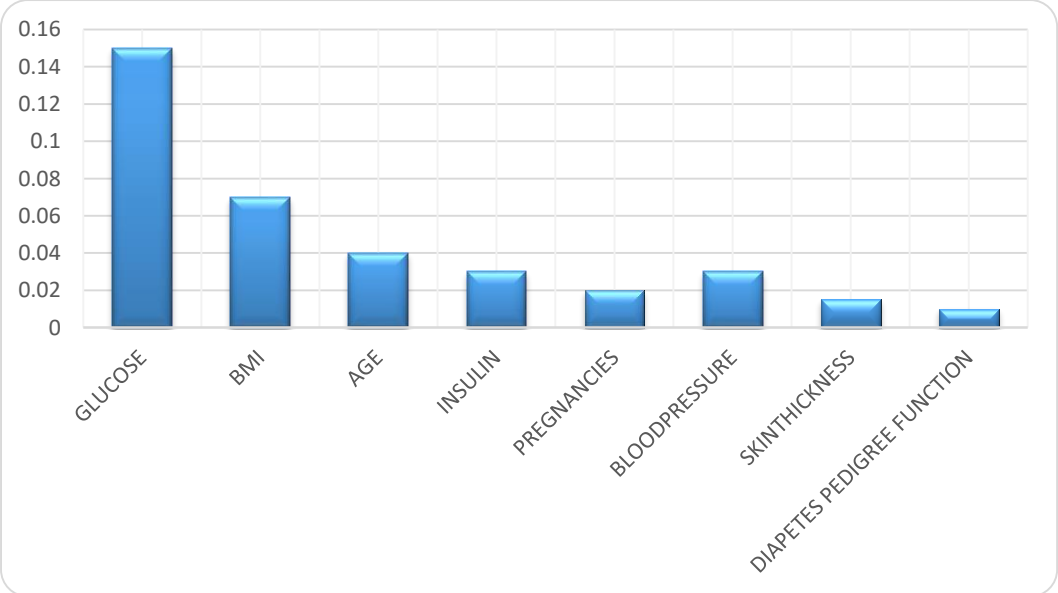

Figure 3. Feature importance hierarchy.

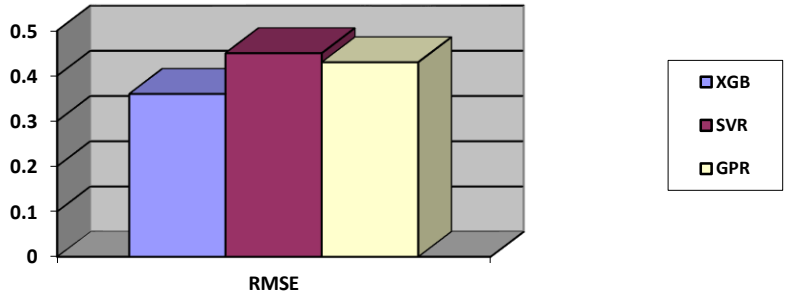

Figure 4. RMSE of various regression models on the Pima Indian Dataset.

Next, we calculated each regression's root mean square error (RMSE) formulated as:

$$\text{RMSE (1)} = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \quad (1)$$

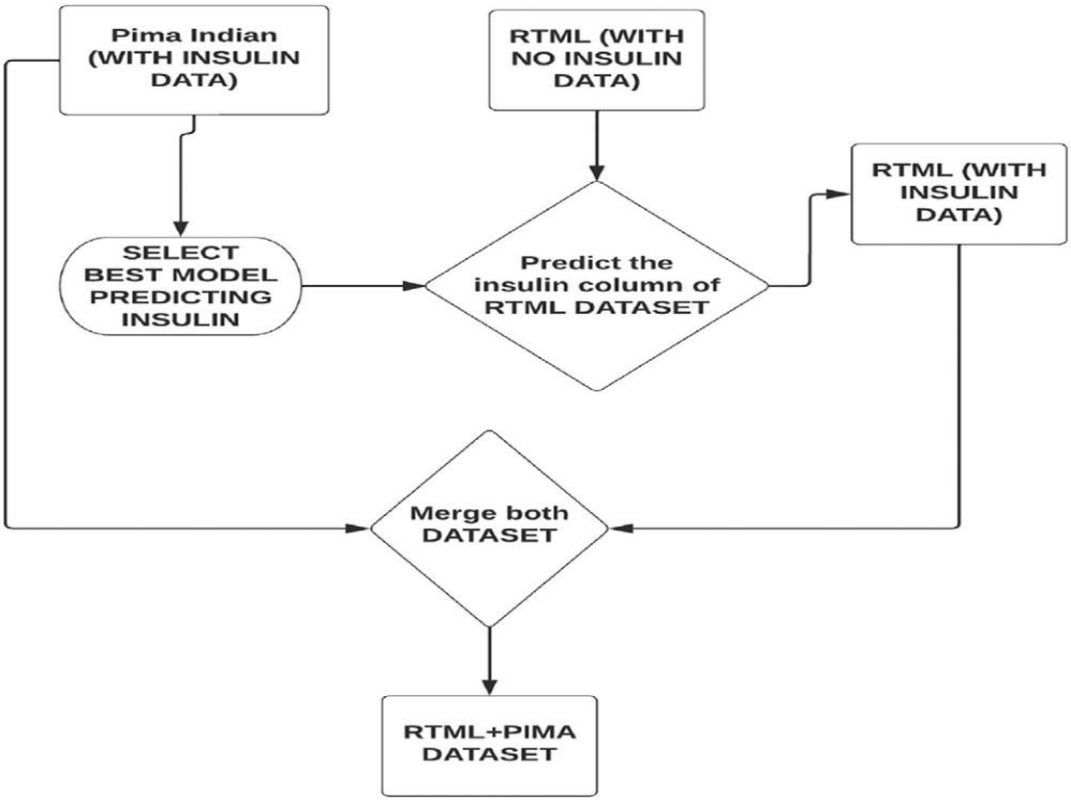

Figure 5. Working steps of predicting insulin of the RTML dataset.

$$X_{scaled} = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}} \quad (2)$$

where Xmax and Xmin denote maximum and minimum values in the individual feature column, respectively.

Figure 4 shows that the XGB technique shows the lowest RMSE of insulin on the Pima Indian dataset, where N is the total number of validation samples of the Pima Indian dataset. As a result, the missing insulin column in the RTML dataset that was gathered from the Pima Indian dataset was predicted using this model. Figure 5 shows the procedures involved in predicting insulin using the RTML dataset. Combined dataset: Following the semi-supervised method, we combined the RTML and Pima Indian datasets and predicted the insulin feature. Except for the diabetic pedigree function, which was determined to be the least significant feature based on mutual information, the combined dataset included 877 records with all the features. ADASYN and SMOTE for class imbalance The merged dataset used in this work comprises the imbalance problem with

samples with and without diabetes, 302 and 669, respectively. To address this issue, the training dataset has been subjected to the SMOTE and ADASYN procedures, but the testing data remains unchanged. ADASYN, or Adaptive Synthetic Sampling, is a synthetic data generation technique that produces more data for examples that are "harder to learn" and avoids duplicating minority samples [25]. Consequently, the minority class will receive the same level of sampling as the majority class. Normalization of min–max: We applied the min-max normalization strategy in this study. The following equation has been used to scale the data to the same range.

3.3 | Machine Learning Classifiers

This work uses a variety of ensemble and machine-learning techniques to construct an automated diabetes prediction system that will be briefly covered below. To avoid overfitting, the GridSearchCV framework was used in this study to determine the ideal values for each hyperparameter across all machine learning models. Decision tree: An arrangement of rules provides a learning function that is represented by a decision tree. The discrete-valued target function approximation approach is carried out via the decision tree learning technique. Information gain is calculated using either entropy or Gini [26], and each node is selected according to these coefficients, which are written as:

$$Gini_i = 1 - \sum_{k=1}^{n} (Pi, k)^2 \quad (3)$$

$$Entropy = \sum_{i=1}^{n} -\mathcal{P}\mathrm{i} \log_2 \mathcal{P}\mathrm{i} \quad (4)$$

In Eqs. (3) and (4), n represents the number of distinct class values. We observed that max depth = 2, minimum samples leaf = 50, and 'Gini' impurity metrics work well in the employed dataset in this work using the GridSearchCV hyperparameter tuning.

3.3.1 | KNN Classifier

A discrete-valued function can be approximated by the K number of nearest classifiers [27]. To classify, it makes a plane with accessible preparing focuses and calculates the distance between the query and trained points. It determines the K number of neighbors (depending on the dataset) and classifies them using majority voting. In our research, we used K = 5 for the binary classification.

3.3.2 | Random Forest

Random forest is a machine learning system that averages the predictions of several decision trees. As a result, the random forest can be considered an ensemble learning model [28]. In this research, we have applied random forest with estimators = 400, minimum samples leaf = 5, and 'Gini' impurity metrics utilizing hyperparameter tuning.

Support vector machine: SVM performs supervised classification by choosing the best hyperplane [29]. In this study, we experimented with various SVM kernels in the training set. Finally, we discovered the SVM with a linear kernel, parameters C = 10 and gamma = 1, produces the best results in this dataset.

3.3.3 | Logistic Regression

Logistic regression can be used to predict a binary class. To predict the outcome, it fits an 'S' shaped function [30]. The hyperparameter optimization technique obtained the maximum number of iterations for the convergence of the logistic regression model to be 150.

3.3.4 | AdaBoost

AdaBoost is an ensemble technique. This classifier initially works on the original dataset and then fits repeated copies of the classifier to the same dataset. This framework adjusts the weights of improperly classified instances so that successive classifiers focus more on difficult circumstances. We have applied AdaBoost with estimator = 50 and learning rate = 0.10 in this work.

3.3.5 | XGBoost

XGBoost is an ensemble machine-learning technique based on decision trees that employ a gradient-boosting approach [31]. The parameters used for the proposed XGBoost classifier are as follows: estimators' maximum depth = 4 and 'binary logistic' objective function.

3.3.6 | Voting Classifier

It is an ensemble technique to improve the classification by voting [32]. This paper implements a voting classifier that selects the majority class predicted by each classifier with a 'soft' voting hyperparameter.

3.3.7 | Bagging

Bagging classifiers are ensemble classifiers that fit base classifiers to random subsets of the original dataset and then aggregate their predictions voting to.

3.4 | Deployment of the Prediction System

To operate instantly on actual data, the suggested machine learning-based diabetes prediction system has been integrated into a framework for a website and smartphone application. Web application: The frontend portion of the suggested website is built using HTML and CSS. Following that, we decided to use ADASYN in conjunction with XGBoost as our machine-learning model since it offered the greatest results. Spyder, an Anaconda-compatible Python environment platform, was used to deploy the model.

Table 4. Features of the RTML private Dataset.

	Precision	Recall	F1 Score	Accuracy	AUC
Logistic regression	0.78	0.77	0.77	77%	0.88
KNN	0.78	0.76	0.76	76%	0.85
Random forest	0.78	0.78	0.78	78%	0.87
Decision tree	0.75	0.73	0.73	73%	0.75
Bagging	0.80	0.79	0.79	79%	0.87
Adaboost	0.79	0.78	0.78	78%	0.85
XGboost	0.78	0.78	0.78	78%	0.84
Voting	0.79	0.79	0.79	79%	0.86
SVM	0.78	0.75	0.76	75%	0.87

4 | Results and Discussion

This section presents the results and discussion of the proposed automatic diabetes prediction system. First, the performance of various machine learning techniques is discussed. We used precision, recall, F1 score, AUC, and classification accuracy to evaluate various ML models. Equations of these metrics are expressed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$\text{F1 score} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}} \tag{7}$$

where TP indicates that both the model's prediction and the outcome are positive. FP denotes the model's positive prediction, while the outcome is negative. TN indicates that both the model's prediction and the actual result are negative. FN denotes a positive outcome despite the model's prediction of a negative one. All of the machine learning models in this work have been trained using a stratified 8:2 train-test split using the holdout validation approach.

Table 4 compares different performance metrics of various classifiers for the merged dataset with the SMOTE synthetic oversampling technique. According to this table, the bagging classifier achieved the best overall performance with 79% accuracy and 0.79 and 0.87 F1 scores and AUC, respectively. Table 5 shows various performance metrics of all the classifiers using the ADASYN approach in the merged datasets. According to Table 4, the XGBoost framework performed better than other classifiers with 81% accuracy and 0.84 AUC. Conversely, the decision tree approach achieved the lowest accuracy and F1 score. The machine learning model is then trained and assessed on distinct samples, or source and target datasets, respectively, using the domain adaptation approach. Initially, the larger-sized, open-source Pima Indian dataset is used in this work to train the autonomous diabetes prediction algorithm. Lastly, the private RTML dataset is used to assess the model.

Table 6 demonstrates the performance metrics for the private dataset. It is interesting to note that the XGBoost with ADASYN framework has been applied in the training dataset in this case.

Figure 6 depicts the confusion matrix for XGBoost with ADASYN. According to this figure, the XGBoost technique correctly classified 141 instances with *TP* = 43 and *TN* = 98. The ROC (receiver operating characteristics) curve of the XGBoost with the ADASYN approach has been illustrated in Figure 7. This figure shows the AUC value of XGBoost is 0.84. Next, explainable AI techniques with SHAP and LIME frameworks are implemented to understand how the model predicts the decision. Figure 8 shows the XGBoost with ADASYN feature importance with the help of explainable AI, SHAP library.

Table 5. Performance metrics of various classifiers using Adasyn in the merged dataset.

Classifier	Precision	Recall	F1 Score	Accuracy	AUC
Logistic regression	0.76	0.75	0.75	75%	**0.84**
KNN	0.76	0.73	0.73	73%	**0.82**
Random forest	0.76	0.76	0.76	76%	**0.84**
Decision tree	0.81	0.72	0.72	72%	**0.78**
Bagging	0.80	0.79	0.79	79%	**0.84**
AdaBoost	0.75	0.76	0.76	76%	**0.84**
XGBoost	0.81	0.81	0.81	81%	**0.84**
Voting	0.77	0.77	0.77	77%	**0.84**
SVM	**0.78**	**0.78**	**0.77**	**78%**	**0.83**

Table 6. Performance metrics for the private dataset (domain adaptation technique).

Precision-Recall	Accuracy	Percentage
0.95	0.96	96%

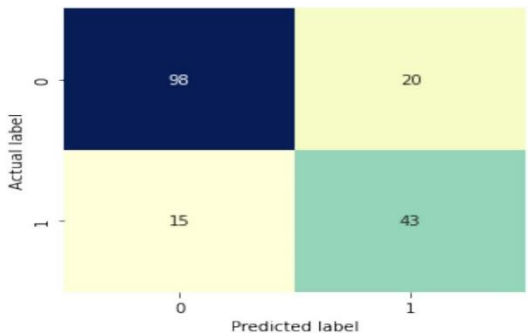

Figure 6. Confusion matrix for XGBoost with ADASYN technique.

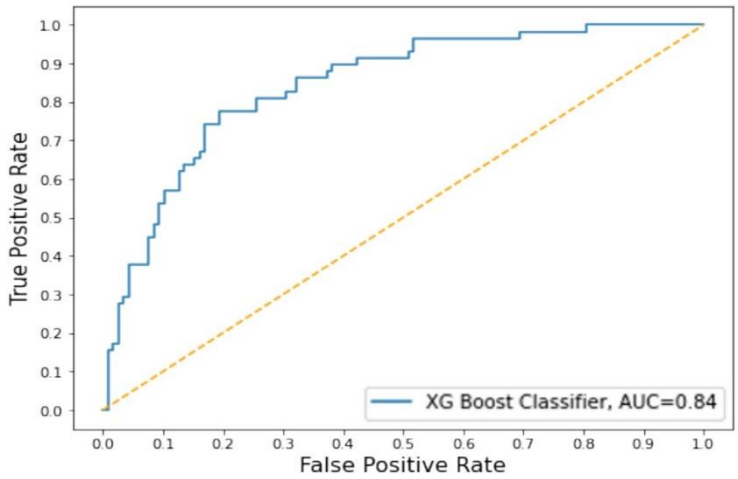

Figure 7. ROC curve and AUC value for the XGBoost with ADASYN.

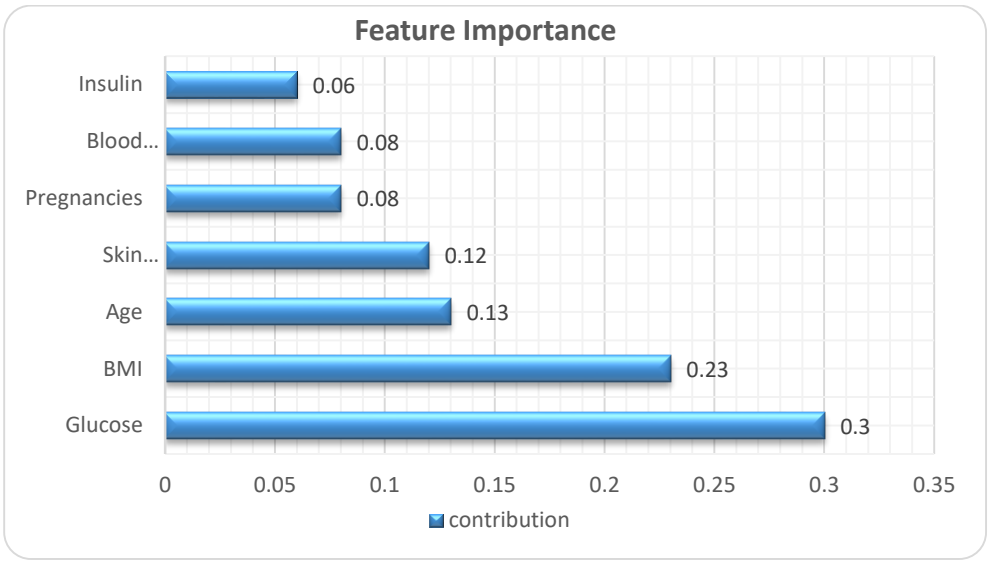

Figure 8. Explainable AI interpretation of feature importance of XGBoost with ADASYN.

Here is the breakdown of the contribution of each feature mentioned:

i). Glucose: This feature has a contribution of 0.3, indicating it has the highest importance among the listed factors.

ii). BMI: The contribution of BMI is 0.23, suggesting it is the second most important factor in the analysis.

iii). Age: With a contribution of 0.13, Age is considered moderately important in determining the outcome.

iv). Skin Thickness: This feature has a contribution of 0.12, indicating its relevance in the analysis.

v). Pregnancies: The contribution of Pregnancies is 0.08, suggesting its impact on the outcome.

vi). Blood Pressure: Similar to Pregnancies, Blood Pressure also has a contribution of 0.08.

vii). Insulin: This feature has a contribution of 0.06, indicating its relatively lower importance compared to other factors.

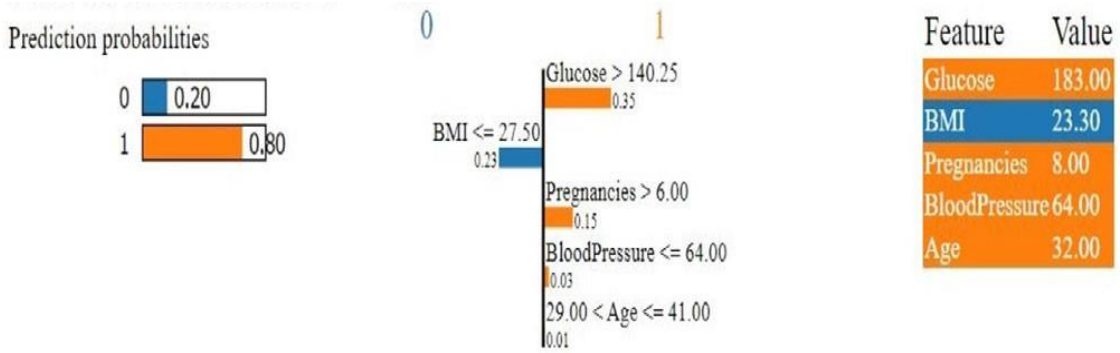

Figure 9. LIME explainable AI prediction interpretation.

Figure 9 illustrates an interpretation of the XGBoost model implemented by the LIME explainable AI method. According to this figure, the model predicts diabetes correctly for this specific person with 80% confidence. The ML model predicts this class as the person has a glucose level of more than 140.25 and involves pregnancies of more than six.

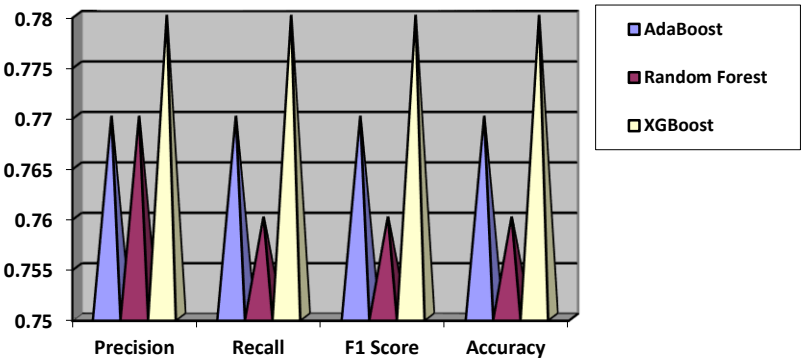

Figure 10. Performance metrics of classifiers in the merged dataset (RTML insulin obtained from Pima Indian mean).

Finally, a survey was conducted in which users rated the application's various features. The participants rated each feature on a scale of 1 to 10, and their average was calculated. According to Figure 10, the diabetes prediction and daily diet chart features of the application achieved the highest ratings of 8.40 and 8, respectively. It is worth mentioning that the RTML dataset's insulin feature has been predicted from the Pima Indian dataset by applying the XGB regression technique for all of the results discussed above. However, alternative investigations have been conducted to obtain the insulin feature of the RTML dataset, that is the mean and median imputation of various patients' insulin of the Pima Indian dataset. Tables 7 and 8 demonstrate various performance metrics of the machine learning models with the ADASYN technique when the RTML dataset's missing insulin features are obtained from the mean and median values of the Pima Indian dataset.

Table 7. Performance metrics of classifiers in the merged dataset (insulin removed from Pima Indian).

Classifier	Precision	Recall	F1 Score	Accuracy
AdaBoost	0.73	0.71	0.72	72%
Random Forest	0.72	0.70	0.71	71%
XGBoost	0.74	0.73	0.73	74%

Table 8. Comparison of the proposed system with similar diabetes prediction works.

Reference	Classifier	F1 score	Recall	Accuracy	Precision:
[32]	Deep belief network model	0.81	1.0	N/A	**0.68**
[33]	SVM with RBFkernel			82%	
[34]	SVM	0.73	0.75	75%	**0.72**
[35]	Ensemble (XGBoost)	0.81	0.79	88.8%	**0.84**
[36]	Soft voting	0.72	0.72	79.1%	**0.73**
Proposed work	**XGBoost withADASYN**	**0.81**	**0.80**	**88.5%**	**0.82**

Finally, the study's paper examines a revolutionary technique for predicting diabetes the use of gadgets to gain knowledge of and explainable synthetic intelligence. It starts with the aid of highlighting diabetes as an enormous disorder impacting tens of thousands and thousands globally. Early and unique analysis is emphasized to lessen diabetes complications. They have a look at introducing a brand-new dataset of Bangladeshi ladies in conjunction with the typically used Pima Indian dataset. Various gadgets gaining knowledge of strategies and characteristic choice algorithms are carried out to enhance version overall performance. Notably, mutual statistics courses' characteristic choice, even as semi-supervised gaining knowledge of severe gradient boosting predicts lacking insulin data. A key power is the technique to cope with magnificence imbalance through SMOTE and ADASYN. By balancing magnificence distribution, predictions end up more dependable and accurate. Explainable AI strategies like SHAP and LIME additionally improve interpretability, permitting a deeper perception of prediction factors. Numerous classifiers are evaluated, together with selection trees, logistic regression, KNN, and bagging are used together. Beyond version development, the device is deployed in an internet and cellular app for real-time use. The user-pleasant interface demonstrates usability for immediate forecasting as shown in Table 9.

Table 9. Results before optimization.

Research paper title	Accuracy achieved
"Prediction of Diabetes using Machine Learning"	85% using KNN
"Diabetes prediction using machine learning and explainable AI techniques"	96% using Logistic Regression, 98.8% using AdaBoost Classifier
"A remote healthcare monitoring framework for diabetes prediction using machine learning"2022	81% with XGBoost, 97.4% for the private dataset, and 83.1% for combined datasets
"A remote healthcare monitoring framework for diabetes prediction using machine learning"2021	97.4% for private datasets, 83.1% for combined datasets

As we noticed the scientists (from the previous studies) reached 97.4% accuracy only, they used a lot of classifiers and the best one was the XGBoost algorithm with SMOTE, so we needed to optimize this accuracy to reach more than 97.4%, so we used Stack technique between 3 classifiers (bagging, logistic regression, and decision tree). Finally, the accuracy reached 99.3%.

5 | Conclusions

The advanced studies provide a critical summary of the application of system learning algorithms in the diagnosis and prognosis of diabetes, highlighting the critical role that early intervention plays in improving fitness outcomes for those at risk for the condition. Through a series of experiments, the study demonstrates the effective application of predictive styles, showing how algorithms such as XGBoost, SVM-RBF, and AdaBoost can find viable diabetic examples with remarkably high accuracy rates. Additionally, a flexible and advanced framework for automated diabetes diagnosis has been introduced, utilizing cloud concepts to seamlessly combine several client and healthcare devices, marking a significant advancement in the delivery of customized treatment The paper also highlights promising directions for future research, including the use of large patient cohorts in longitudinal studies, the integration of additional devices and patient data, and the optimization of a set of rule selections to improve accuracy and usability in actual healthcare settings. The research emphasizes the potential to reduce diabetes-related risks and outcomes and stresses the significance

of applying artificial intelligence and system analysis techniques for early diabetes prediction. Long-term effects on impacted individuals and first-class lifestyles will result from this. We observed that the scientists only attained 82% accuracy despite using numerous classifiers, with the most satisfactory one evolving into the xgbooster. Consequently, we wanted to maximize this accuracy to achieve more than 82%, so we employed the stack among three classifiers (bagging, logistic regression, and choice tree). so, the accuracy reached 99.3%.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

"Conceptualization, K.E. and A.E.; Methodology, R.A.; Software, N.W.; Validation, K.E., A.E. and N.W.; formal analysis, N.W.; investigation, R.A.; resources, K.E.; data maintenance, R.A.; writing-creating the initial design, A.E.; writing-reviewing and editing, K.E.; visualization, N.W.; monitoring, A.E.; project management, K.E.; funding procurement, K.E. All authors have read and agreed to the published version of the manuscript.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Atlas, G.: Diabetes. International Diabetes Federation. 10th ed., IDF Diabetes Atlas.
[2] Akhtar, S., et al.: Prevalence of diabetes and pre-diabetes in Bangladesh: A systematic review and meta-analysis. BMJ Open 10, e036086 (2020)
[3] Prabhu, P., Selvabharathi, S.: Deep belief neural network model for prediction of diabetes mellitus. In: International Conference on Imaging, Signal Processing and Communication, pp. 138–142 (2019)
[4] VijiyaKumar, K., Lavanya, B., Nirmala, I., Caroline, S.S.: Random Forest algorithm for the prediction of diabetes. In: International Conference on System, Computation, Automation, and Networking, pp. 1–5 (2019)
[5] Mohan, N., Jain, V.: Performance analysis of support vector machine in diabetes prediction. In: International Conference on Electronics, Communication and Aerospace Technology, pp. 1–3 (2020)
[6] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Annual Symposium on Computer Applications in Medical Care pp. 261–265 (1998)
[7] Aurélien, G.: Hands-On Machine Learning with Scikit-Learn and Ten- TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., Sebastopol, CA
[8] Mitchell, T.M.: Machine Learning. McGraw-Hill, Inc., New York
[9] Chatrati, S.P., Hossain, G., Goyal, A., et al.: Smart home health monitoring system for predicting type 2 diabetes and hypertension. J. King Saud Univ. Compute. Inf. Sci. 34(3), 862–870 (2020)
[10] Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes pre-diction using ensembling of different machine learning classifiers. IEEE Access 8, 76516–76531, (2020)

[11] Cervantes, J., García-Lamont, F., Rodríguez, L., Lopez-Chau, A.: A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408, 189–215 (2020)
[12] Pranto, B., et al.: Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information 11, 1–20 (2020)
[13] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, pp. 1322–1328 (2008)
[14] Deberneh, H.M., Kim, I.: Prediction of type 2 diabetes based on a machine learning algorithm. Int. J. Environ. Res. Public Health 18, 1–14 (2021)
[15] Olisah, C.C., Smith, L., Smith, M.: Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Compute. Methods Programs Biomed. 220, 1–12 (2022)
[16] Ahmed, N., et al.: Machine learning based diabetes prediction and development of smart web application. Int. J. Cogn. Compute. Eng. 2, 229–241 (2021)
[17] Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y.: AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J. Supercomput. 77, 5198–5219 (2021)
[18] Ramesh, J., Aburukba, R., Sagahyroon, A.: A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technol. Lett. 8, 45–57 (2021)
[19] Mounika, V., Neeli, D.S., Sree, G.S., Mourya, P., Babu, M.A.: Prediction of type-2 diabetes using machine learning algorithms. In: International Conference on Artificial Intelligence and Smart Systems, pp. 127–131 (2021)
[20] Malasinghe, L.P., Ramzan, N., Dahal, K.: Remote patient monitoring: a comprehensive study. J. Ambient Intell. Hum. Compute. 10(1), 57–76 (2019)
[21] Singh, A.S., Masuku, M.B.: 'Sampling techniques and determination of sample size in applied statistics research: an overview', Int. J. Economics, Commerce Manage., 2014, 2, (11), pp. 1–22.
[22] Karegowda, A.G., Jayaram, M.A.: 'Cascading GA & CFS for feature subset selection in medical data mining'. IEEE Int. Advance Computing Conf., 2009, vol. 5, pp. 1–4
[23] Martis R., Crowther C. A., Shepherd E., Alsweiler J., Downie M., Brown J.: Treatments for women with gestational diabetes mellitus: an overview of Cochrane systematic reviews. Cochrane Database of Systematic Reviews. 8(8), (2018).
[24] Caesarean Birth NICE Guideline [NG192], National Institute for Health and Care Excellence, 2021
[25] Hirst, J.E., et al.: GDm-Health: A pilot study examining the acceptability of mobile phone assisted remote blood glucose monitoring for women with gestational diabetes mellitus. Reprod. Sci. 21(3), 252a–253a (2014)
[26] Weinert, L.S.: International Association of Diabetes and Pregnancy Study Groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy: Comment to the International Association of Diabetes and Pregnancy Study Groups Consensus Panel. Diabetes Care 33(7), e97 (2010). https://doi.org/10.2337/dc10-0544
[27] Akhtar, S., Nasir, J.A., Sarwar, A., Nasr, N., Javed, A., Majeed, R., Salam, M.A. & Billah, B. (2020) Prevalence of diabetes and diabetes in Bangladesh: A systematic review and meta-analysis. BMJ Open, 10, no. 9, Article ID e036086. DOI: 10.1136/bmjopen-2019-036086, PubMed: 32907898.
[28] Aljumah, A.A., GulamAhamad, M. & Siddiqui, M.K. Application of Data Mining: Diabetes Health Care in.
[29] Atlas, G. Diabetes. International, 10th, IDF Diabetes Atlas. Diabetes Federation.
[30] El-Sofany, H., El-Seoud, S.A., Karam, O.H., Abd El-Latif, Y.M. & Taj-Eddin, I.A.T.F. (2024) A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. International Journal of Intelligent Systems, 2024, 1–13. DOI: 10.1155/2024/6688934.
[31] Jeevan, Y. & Kumar, N. (2019) Prediction of diabetes using machine learning. In: Article in International Journal of Innovative Technology and Exploring Engineering. https://www.researchgate.net/publication/371178003.
[32] Kalyankar, G.D., Poojara, S.R. & Dharwadkar, N.V.," Predictive Analysis of Diabetic Patient Data Using Chine Learning and Hadoop", International Conference On I-S C978-1-5090-3243-32017.
[33] Malasinghe, L.P., Ramzan, N. & Dahal, K. (2019) Remote patient monitoring: A comprehensive study. Journal of Ambient Intelligence and Humanized Computing, 10, 57–76. DOI: 10.1007/s12652-017-0598-x.
[34] Mujumdar, A. & Vaidehi, V. (2019) Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292–299. DOI: 10.1016/j.procs.2020.01.047.
[35] Ramesh, J., Aburukba, R. & Sagahyroon, A. (2021) A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technology Letters, 8, 45–57. DOI: 10.1049/htl2.12010, PubMed: 34035925.
[36] Tasin, I., Nabil, T.U., Islam, S. & Khan, R. (2023) Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10, 1–10. DOI: 10.1049/htl2.12039, PubMed: 37077883.