Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–26
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241271867
journals.sagepub.com/home/dhj
Sage

Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications

Md. Alamin Talukder[1], Md. Manowarul Islam[2], Md Ashraf Uddin[3], Mohsin Kazi[4], Majdi Khalid[5], Arnisha Akhter[2] and Mohammad Ali Moni[6]

Abstract

Objective: Diabetes is a metabolic disorder that causes the risk of stroke, heart disease, kidney failure, and other long-term complications because diabetes generates excess sugar in the blood. Machine learning (ML) models can aid in diagnosing diabetes at the primary stage. So, we need an efficient ML model to diagnose diabetes accurately.

Methods: In this paper, an effective data preprocessing pipeline has been implemented to process the data and random oversampling to balance the data, handling the imbalance distributions of the observational data more sophisticatedly. We used four different diabetes datasets to conduct our experiments. Several ML algorithms were used to determine the best models to predict diabetes faultlessly.

Results: The performance analysis demonstrates that among all ML algorithms, random forest surpasses the current works with an accuracy rate of 86% and 98.48% for Dataset 1 and Dataset 2; extreme gradient boosting and decision tree surpass with an accuracy rate of 99.27% and 100% for Dataset 3 and Dataset 4, respectively. Our proposal can increase accuracy by 12.15% compared to the model without preprocessing.

Conclusions: This excellent research finding indicates that the proposed models might be employed to produce more accurate diabetes predictions to supplement current preventative interventions to reduce the incidence of diabetes and its associated costs.

Keywords

Diabetes, prediction, machine learning, efficient, preprocessing, diagnosing

Submission date: 13 February 2024; Acceptance date: 27 June 2024

Introduction

Diabetes mellitus (DM) is a chronic disorder that affects carbohydrate, protein, and fat metabolism, leading to abnormal blood glucose levels.[1] It is classified into two main types: type 1 and type 2 diabetes (T2D).[2] Type 1 diabetes typically occurs in children but can manifest in adults, particularly in their late 30s and early 40s. Patients with type 1 diabetes are usually not obese and often present with a life-threatening condition known as diabetic ketoacidosis.[3] The etiology of type 1 diabetes involves damage to pancreatic cells due to environmental or infectious agents, triggering an autoimmune response against β-cells. Autoimmunity is

[1]Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh
[3]School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, Australia
[4]Department of Pharmaceutics, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia
[5]Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia
[6]Artificial Intelligence & Data Science, Faculty of Health and Behavioural Sciences, The University of Queensland, Brisbane, Australia

Corresponding author:
Md. Alamin Talukder, Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh.
Email: alamin.cse@iubat.edu

considered the primary factor in the pathophysiology of type 1 diabetes. Type 1 diabetes is also associated with other autoimmune diseases.[4] On the other hand, T2D has a distinct pathophysiology and etiology, characterized by a combination of low insulin production and insulin resistance. Obesity, physical inactivity, poor diet, and urbanization contribute to the rising prevalence of T2D.[5] Dysfunction of β-cells plays a crucial role in progressing from prediabetes to diabetes. Despite their differing pathophysiology, both types of diabetes share similar complications, including macrovascular and microvascular complications.[6] It is a severe chronic disease,[7] associated with various consequences and increased mortality.[8]

Health regulations emphasize regular screenings for individuals with diabetes risk factors,[9] highlighting the importance of timely identification and intervention. Preventive measures are crucial alongside diabetes care.[10] Early diagnosis and lifestyle modifications, such as healthy eating and exercise, can reduce the progression from impaired glucose tolerance to prediabetes.[11] Technology, particularly machine learning (ML), has gained popularity for early detection and prevention in healthcare.[12–14] ML in diabetes management offers a promising avenue for predictive modeling. By analyzing vast datasets encompassing patient demographics, medical history, and lifestyle factors, ML algorithms can predict the likelihood of diabetes onset or progression with remarkable accuracy. These models not only assist in early detection but also empower healthcare providers to tailor personalized interventions, ultimately mitigating complications and improving patient outcomes.[15,16]

Several ML algorithms have been introduced for diabetes detection, offering benefits such as low computation costs, robustness, and high performance.[17] For instance, researchers have utilized classifiers such as Naive Bayes (NB), decision tree (DT), adaptive boosting (AdaBoost), and random forest (RF) for diabetes prediction,[18] while models such as generalized linear models with elastic net regularization (Glmnet), RF, extreme gradient boosting (XGBoost), and light gradient boosting (GB) machine have been explored for predicting type 2 diabetes.[19] According to recent projections, the prevalence of diabetes is expected to rise significantly, imposing a substantial burden on healthcare systems worldwide.[20] Early detection and effective management of diabetes are crucial for preventing complications and improving patient outcomes. ML algorithms have gained attention for their potential to enhance diabetes detection and prognosis by analyzing complex and non-linear medical data.[21] The following aims to provide a comprehensive overview of the ML approaches employed for diabetes detection and prognosis. By critically examining the existing research, we aim to identify the strengths and limitations of different techniques and highlight potential avenues for our proposal.

Ahmed et al.[22] developed an optimized ML-based classifier model for diagnosing diabetes using clinical data. Their approach included effective preprocessing techniques and achieved superior efficiency compared to existing methods, with an improvement in accuracy ranging from 2.71% to 13.13%. However, the generalizability of their model to different datasets and populations requires further investigation.

Hasan et al.[17] proposed a comprehensive architecture for diabetes prognosis, incorporating outlier exclusion, data normalization, and weighted ensembling of multiple ML models. Their suggested ensemble model achieved an area under curve (AUC) score of 95% on the Pima Indian dataset. However, the study's limitation is that it focused only on the performance of a single dataset, limiting the assessment of generalizability.

Howlader et al.[23] applied ML strategies to identify T2D patients. They performed extensive feature selection and analysis using various classification algorithms, with the generalized boosted regression model achieving the best accuracy rate of 90.91%. However, the study's scope was limited to the prediction of T2D and did not explore other types of diabetes or broader diabetes prognosis.

Deepajothi et al.[24] aimed to forecast diabetes in its initial phases by incorporating hereditary factors into a fuzzy classification model. Their suggested model achieved an accuracy rate of 83% for identifying T2D using the Pima Indian dataset. However, the study's limitation is that it did not compare the performance of their models with other existing diabetes prognosis methods.

Rajagopal et al.[25] developed a modified combined approach of artificial neural networks with genetic algorithms for diabetes detection. Their model achieved an accuracy rate of 80% on the Pima Indian dataset. However, the study did not explore the performance of other ML algorithms' performance or evaluate their approach's generalizability on different datasets.

Nuankaew et al.[26] proposed a unique predicting approach called average weighted objective distance (AWOD) for diabetes forecasting. Their technique achieved an accuracy rate of 93.22% on the Pima Indian dataset and 98.95% on the Mendeley dataset. However, the study did not compare the performance of AWOD with other existing diabetes prediction methods.

Wei et al.[27] developed a methodology to estimate the usefulness of ambient chemical exposure in diagnosing DM. Their ML model utilizing the least absolute shrinkage and selection operator regression achieved an AUC of 80% for diabetes detection. However, the study's limitation is that it focused only on the prediction of diabetes and did not consider other aspects such as prognosis or subtype classification.

Sivaranjani et al.[28] used support vector machine (SVM) and RF ML algorithms to predict the likelihood of developing diabetes-related disorders. The RF model achieved an accuracy of 83% after feature selection and principal component analysis dimensionality reduction. However, the study did not explore other classification algorithms' performance or evaluate their approach's generalizability on different datasets.

Ramesh et al.[29] introduced an end-to-end monitoring system for diabetes risk stratification and control. Their SVM model achieved an accuracy rate of 83.20% using the Pima Indian dataset. However, the study's limitation is that it focused on risk stratification and did not extensively evaluate the performance of their model on other aspects such as diagnosis or prognosis.

Ravaut et al.[30] constructed an ML-based GBDT model for estimating adverse outcomes related to diabetes. Their model achieved an AUC statistic of 77.7% for predicting the three-year risk of developing diabetes complications. However, the study's limitation is that it relied on organizational health data from a specific region, and the generalizability of their model to other populations requires further investigation.

Naz et al.[31] proposed a procedure for early diabetes estimation using various ML classifiers and the Pima dataset. Their deep learning (DL) approach achieved a success rate of 98.07% and retrieved viable properties. However, the study's limitation is that it focused solely on DL-based methods and did not explore the performance of other ML algorithms.

Hassan et al.[32] provided a diabetes prognosis model based on ML algorithms, achieving accuracy rates of 94.5%, 96.5%, and 97.5% for logistic regression (LR), SVM, and RF, respectively. However, the study's limitation is that it utilized a relatively small dataset of 250 variations, raising questions about the generalizability of its model to larger and more diverse datasets.

Gupta et al.[33] developed prognostic tools using DL and quantum ML (QML) approaches. Their DL classifier achieved an accuracy rate of 95%, while the QML classifier had 86% accuracy. However, the study's limitation is that it focused on comparing DL and QML methods without considering other traditional ML algorithms.

Gupta et al.[34] introduced the application of Moth-Flame optimization (MFO), a metaheuristic algorithm, for classifying diabetes data. They incorporated MFO to update feedforward neural network weights and conducted performance evaluations on the Wisconsin Hospital dataset, along with a comparative analysis of contemporary literature.

Majhi et al.[35] investigated and compared multiple ML approaches for early diabetes risk assessment and medical diagnosis enhancement. Their study employed two real-world datasets: a diabetic clinical dataset (DCA) from Assam, India, and the publicly available PIMA Indian diabetic dataset. Various classifiers were utilized, with LR yielding the most promising results on PIMA, achieving an accuracy of 79.22%.

In contrast to these studies, our current work aims to address the limitations mentioned above. We propose an optimized data preprocessing pipeline, tackle imbalanced datasets, prevent overfitting using k-fold cross-validation, and conduct extensive experimental validation on diverse datasets.

It is undoubtedly challenging to predict diabetes in its early stages due to the complex interdependencies between numerous factors. Creating a medical prediction model that aids medical professionals in the prediction procedure is necessary. An accurate diabetes prognosis is crucial to prevent premature death. Therefore, achieving a greater accuracy rate with a lower error rate is required to forecast diabetes better. In this paper, we have adopted several ML algorithms to predict diabetes and identify the best one based on clinical data related to diabetes to address these issues. To improve accuracy and achieve higher performance, it is necessary to preprocess the raw data to match the criteria of different classifiers. An efficient data preprocessing pipeline is provided to the learning algorithms for predictive modeling. To assess our proposal, an extensive experimental analysis has been performed on four diabetes datasets, each with different attributes. Experimental results show that our proposal surpasses state-of-the-art research in predicting diabetes, with an average accuracy of 95.5%.

The paper makes the following contributions:

- **Optimized data preprocessing pipeline:** We develop a robust pipeline for preprocessing diabetes-related datasets. This includes handling missing values, outliers, label encoding, and normalization. Our preprocessing techniques improve dataset quality, leading to enhanced classifier performance.
- **Addressing imbalanced datasets:** We tackle the challenge of imbalanced data by implementing random oversampling techniques. This creates a balanced dataset, improving the performance of diabetes detection and prognosis models.
- **Overfitting prevention:** To prevent overfitting, we employ k-fold cross-validation during model training. This ensures that the models generalize well to unseen data, enhancing their reliability for diabetes prediction.
- **Extensive experimental validation:** Through extensive experiments on diverse diabetes datasets, our approach consistently outperforms existing methods. We demonstrate superior accuracy, precision, recall, and F1-score, validating its effectiveness for diabetes detection and prognosis.

These contributions advance the field of diabetes research by providing an optimized preprocessing pipeline, addressing dataset imbalance, preventing overfitting, and demonstrating superior performance through extensive experimentation.

The hypothesis of this paper is described as follows:

Hypothesis

This study proposes an optimized data preprocessing pipeline and addresses the challenge of imbalanced datasets in the context of diabetes detection and prognosis. By implementing robust preprocessing techniques, including handling missing values, outliers, label encoding, and normalization, and employing random oversampling, we hypothesize that the dataset quality will be enhanced, leading to improved classifier performance. Furthermore, by employing k-fold cross-validation to prevent overfitting and conducting extensive experimental validation, we expect to demonstrate our approach's superiority in accuracy, precision, recall, and F1-score compared to existing methods. These findings will validate the effectiveness of our proposed methodology for diabetes detection and prognosis, making a significant contribution to the field.

The remainder of this paper is structured as follows: The "Proposed Methodology" section outlines our proposed approach and the various machine learning algorithms employed for diabetes prediction. In the "Results" section, we present the experimental findings. The "Discussion" section provides a comparative analysis of our model against existing methods. Finally, the "Conclusion" section summarizes our findings and discusses potential future enhancements.

Proposed methodology

In this section, we have described our proposed methodology along with the different machine algorithms adopted in the system. Firstly, we explain the working principle of the proposal. Then, we briefly describe the ML models.

Figure 1 shows the basic workflow of the proposed approach. The proposal has five significant parts: data collection, data preprocessing, handling imbalanced class problems, splitting datasets using k-fold cross-validation and applying ML algorithms to train and test the models, and evaluating the performance.

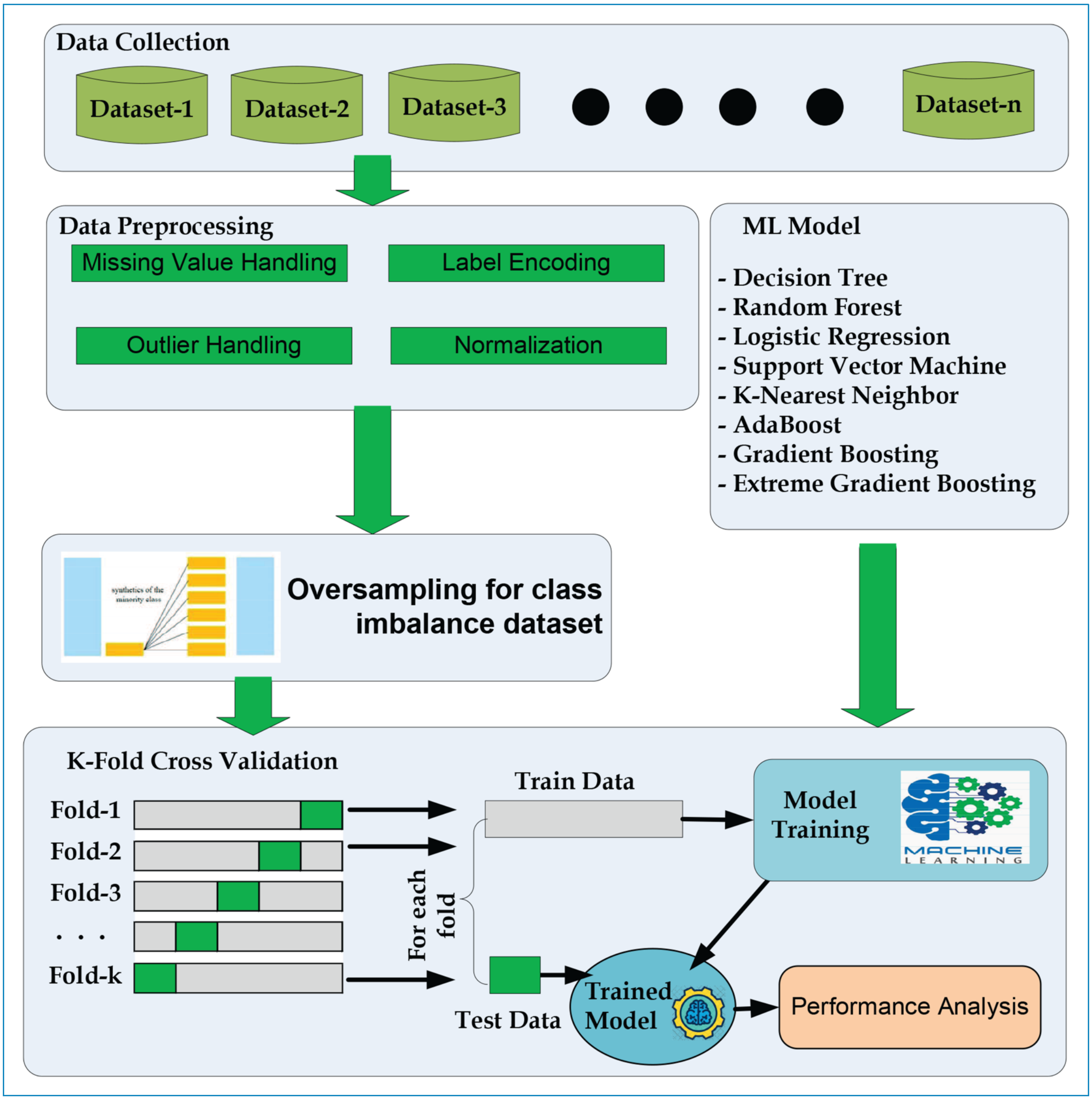


Figure 1. The workflow diagram for diabetes prediction.

The steps performed in our study are as follows:

- Data collection: We gathered the necessary data from publicly available sources.
- Data preprocessing: After data collection, we conducted data preprocessing to prepare the dataset for model training and evaluation. This includes cleaning the data, handling missing values, and performing outlier handling, label encoding, and data normalization.
- Handling imbalanced class problems: To address any class imbalance issues in the dataset, we employed techniques such as oversampling, undersampling, or synthetic data generation to balance the classes. This step aims to prevent biases in the model due to imbalanced data.
- Splitting datasets using k-fold cross-validation: We utilized k-fold cross-validation to split the dataset into training and testing folds. The preprocessing steps, including data balancing, were applied within each fold. This ensures that the data preprocessing steps are performed independently for each fold and prevents any information leakage between the training and testing sets.
- ML model building and evaluation: Within each fold, we built ML models using the training data and evaluated their performance on the corresponding testing data. This process was repeated for each fold, resulting in multiple model evaluations.

By following this pipeline, we aimed to ensure a robust evaluation of our proposed method for diabetes prediction. In our proposed methodology shown in Figure 1, the "Data Preprocessing" box is placed outside the k-fold cross-validation loop in this proposal for the following reasons:

- Preventing data leakage: By performing data preprocessing separately for each fold, outside the cross-validation loop, information leakage from the test set to the training set is avoided. This ensures unbiased evaluation and a more accurate assessment of the model's generalization capabilities.
- Realistic evaluation: Applying data preprocessing techniques outside the cross-validation loop mimics real-world scenarios where models encounter unseen, preprocessed data during deployment. This approach provides a realistic evaluation of the model's effectiveness and its performance on truly unseen data.
- Efficiency: Placing data preprocessing outside the cross-validation loop improves computational efficiency. Preprocessing is performed once on the entire dataset before cross-validation, reducing redundant computations within each fold and speeding up the overall evaluation process.

It is important to consider the specific requirements of the study and the nature of the preprocessing techniques used when deciding the placement of data preprocessing and cross-validation to ensure unbiased and reliable model evaluation.

Data collection

To ensure the robustness of our model, we gathered data from four distinct datasets, each containing different variables related to diabetes. These datasets were collected from various sources, including demographic data, diabetes statistics, and health characteristics obtained from individuals across different countries and healthcare institutions. The first dataset used in our study is the Pima Indian Diabetes dataset,[36] which is widely recognized as a valuable resource for evaluating ML algorithms in predicting diabetes within the general population. Dataset 2,[37] referred to as the Austin public health diabetes self-management education participant demographics 2015–2017, comprises demographic information collected from participants in the diabetes self-management education program conducted by Austin Public Health. For Dataset 3,[38] a survey was conducted to gather data from 950 records, including 19 attributes that have been identified as having a measurable influence on diabetes. Lastly, Dataset 4[39] was collected from the Iraqi society, as well as from the laboratory of Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. By utilizing these diverse datasets, we aim to enhance the generalizability and applicability of our proposed model for diabetes prediction and prognosis.

Data analysis and data preprocessing

Data preprocessing is the process of preparing original data for ML. It is indeed the most important step in the process of developing an ML model. It is a necessary step for ML algorithms to improve the model's accuracy and efficiency.

Data preprocessing covers data preparation, which includes data integration, cleansing, normalization, and transformation, as well as data reduction activities such as feature selection, instance selection, and discretization.

We have conducted some data analysis after collecting the data. Then, we have done preprocessing tasks including outliers removal and dealing with missing values, data normalization, encoding, and so on.

1. **Outliers handling**: A dataset may contain extreme values that are beyond the acceptable limits and dissimilar to the rest of the data. This kind of data may reduce the performance of the ML algorithm. Any value $\leq Q_1 - 1.5 \times$ interquartile range (IQR) or $\geq Q_3 - 1.5 \times$ IQR is considered as outlier. Identifying and handling outliers can be expressed as the following Algorithm 1 using first quartile Q_1, third quartile Q_3, and IQR:
2. **Missing value handling**: One of the most difficult tasks for analysts is dealing with missing values, because making the proper decision on how to deal with them leads to more robust data models. We can handle missing values in various ways, such as ignoring the

row, imputation of the missing values with data means, median or mode of the observation, imputation with different ML algorithms or prediction using regression, so on. To increase the model performance, the mean value of the respective attribute has been imputed to manage the missing values, which are crucial for diabetes prediction and can be calculated in equation (1) as follows:

$$x = \begin{cases} \text{mean}(X) & \text{if x is missing} \\ x & \text{otherwise} \end{cases} \tag{1}$$

here, the value of instance x is imputed with the attribute mean as denoted as mean(X).

3. **Label encoding**: All the data entered in the ML algorithm must be numerical. However, in most cases, the dataset may contain categorical data rather than numerical values. So, if the dataset contains any categorical attribute, it must be needed to convert to numeric values before fitting and evaluating an ML model. Label encoding is the process of transforming labels of text/categorical values into a numerical format which is understandable by the ML algorithms. For example, we can convert the categorical values of Gender status “Male” to “1” and “female” to “0.”

Algorithm 1 Handling outlier problem

```
Require:
    Q1: first quartile (25%) of the data
    Q3: third quartile (25%) of the data
    U: upper limit; L :lower limit
    IQR: interquartile range; N: number of features
Ensure: Outlier removed dataset
    IQR ← Q3 − Q1
    U ← Q1 + 1.5 ∗ IQR
    L ← Q3 − 1.5 ∗ rmIQR
    for i = 0 to S do
        if $XU$ then
            $X \leftarrow U$
        else if $XL$ then
            $X \leftarrow L$
        end if
end for
```

4. **Standardization**: When features of an input dataset have considerable variations between their ranges or when they are collected or measured in different measurement units, standardization becomes necessary. Differences hamper the performance results for ML models in the range of initial features. So normalization or standardization can solve this issue and improve the prediction quality. We have used standardization to rascal the values of any attribute for better accuracy of the classification model using the following equation (2):

$$x_s = \frac{x - \text{mean(X)}}{\sigma} \tag{2}$$

here, x_s, mean(X) and σ represent the standardization value, mean value, and standard deviation of the attribute X.

Class balancing using oversampling

There is a class imbalance when observation in one class exceeds observation in other classes. This may cause poor performance as the ML algorithm may ignore the minority class. To deal with this problem, we apply the random oversampling method to balance the dataset. The benefit of oversampling is that no information from the original training set is lost because all members of the minority and majority classes are kept, and it also significantly increases the size of the training set.

ML algorithms

We have used seven different ML classifiers to train the model, including DT, NB, K-nearest neighbor (KNN), LR, extreme gradient boosting (XGBoost), and SVM, and predict diabetes. Depending on several performance metrics, the performance of each classifier has been analyzed. The following subsection will describe the ML algorithms used for the predictive model.

- **Naive Bayes (NB):**
  - NB is based on Bayes’ theorem.[40] It assumes that the value of one feature in a class is independent of the presence of any other feature. Despite its simplicity, NB often outperforms more sophisticated classification methods.[22]
- **Decision tree (DT):**
  - DT is a supervised classifier commonly used for solving classification problems.[41] It effectively captures decision-making information from the dataset. Internal nodes represent features, and leaf nodes represent outcomes.
- **Random forest (RF):**
  - RF is an ensemble classifier that trains multiple DTs.[42] The final classification is based on the majority vote of all trees. Increasing the number of trees generally improves performance.

- **Logistic regression (LR):**
  - LR is a supervised classifier used for binary classification tasks based on event probabilities.[16] It assumes linear separability of data.
- **K-nearest neighbor (KNN):**
  - KNN is a supervised learning method for classification.[43] It assigns the new case to the class nearest to it in the training dataset.
- **Gradient boosting (GB):**
  - GB predicts continuous or categorical target variables by iteratively improving models.[44] It corrects errors from previous models, enhancing overall performance.
- **Extreme gradient boosting (XGBoost):**
  - XGBoost is a gradient-boosted DT implementation known for its speed and efficiency.[42]
- **Adaptive boosting (AdaBoost):**
  - AdaBoost is an ensemble learning technique that improves the performance of ML algorithms, often using DTs with only one split.[45]
- **Support vector machine (SVM):**
  - SVM efficiently separates datasets by finding optimal decision boundaries or hyperplanes in n-dimensional space, maximizing the margin between support vectors.[46]

K-fold cross-validation

Cross-validation is a resampling technique for evaluating ML models. It guarantees that every observation in the dataset can be selected in the training and test data. It is one of the best tactics if we have limited raw data. A badly chosen value for k could result in an inaccurate illustration of the model's skill, such as accuracy with a high variance. A $k = 10$ means the k-fold process divides the whole dataset five or 10 times for evaluating the model. For our experiments, we use $k = 10$ fold validation and take the performance average for the result analysis.

Results

We have conducted our proposal using various preprocessing techniques as well as utilizing several ML algorithms to analyze and find the best model to use in the prediction of diabetes for clinical purposes. We have tested our proposal with four different datasets and each of them contains different types and numbers of attributes to prove that our preprocessing techniques are more efficient than the normal preprocessing process as well as other existing research works.

Experimental setup

All the experiments were conducted on a computer having an Intel Core i7 processor with a 4 GB graphics card, 16 GB RAM, and a 64-bit Windows operating system running at 1.80 GHz using the Python programming language. The dataset is shuffled and divided into 10 folds at random, with one fold used for the testing and the others used for training each ML model. Then, the resultant average value of any performance results has been taken to assess them.

Table 1. Confusion matrix.

Predicted results	Actual positive	Actual negative
Yes	TP	FP
No	FN	TN

TP: true positive; FP: false positive; FN: false negative; TN: true negative.

Performance metrics

The prediction of any ML algorithm could have four distinct results depending on the confusion matrix as indicated in Table 1: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Then, we consider the following metrics to analyze the proposal:

- Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of the number of correct predictions (TPs and TNs) to the total number of predictions made (equation (3)).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

- Precision quantifies the ability of the model to avoid FPs. It is calculated as the ratio of TPs to the sum of TPs and FPs (equation (4)).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

- Recall measures the model's ability to capture all positive instances. It is calculated as the ratio of TPs to the sum of TPs and FNs (equation (5)).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

- The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated as twice the product of precision and recall divided by the sum of precision and recall (equation (6)).

$$\text{F1\_score} = 2 * \frac{(\text{precision} * \mathit{recall})}{(\text{precision} + \mathit{recall})} \tag{6}$$

- Mean absolute error (MAE) measures the average absolute difference between predicted values and actual values. It is calculated as the sum of absolute differences divided by the total number of instances (equation (7)).

$$\text{MAE} = \frac{\sum_{i=1}^{n} \text{predict}(i) - \text{actual}(i)}{n} \tag{7}$$

- Mean-squared error (MSE) measures the average of the squares of the errors or deviations. It is calculated as the sum of squared differences divided by the total number of instances (equation (8)).

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (\mathrm{predict}(i) - \mathrm{actual}(i))^2}{n} \tag{8}$$

- Root MSE (RMSE) is the square root of the MSE and represents the average magnitude of the error. It provides a measure of how spread out the errors are (equation (9)).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{predict}(i) - \mathrm{actual}(i))^2}{n}} \tag{9}$$

- The area under receiver operative curve (AUROC) quantifies a binary classifier's ability to distinguish between classes. It ranges from 0.5 to 1, where higher values signify better performance. In medical settings, AUROC assesses a test's accuracy in identifying patients with a condition while minimizing misclassifications. It offers a concise measure of the model's overall discriminative power, crucial for evaluating diagnostic tools in clinical practice.

Hyperparameters

Hyperparameter selection is an important aspect of ML model development, and we acknowledge that our previous submission lacked information on the hyperparameter optimization process. Here, we provided an overview of the hyperparameter values used for each ML model in Table 2.

In the case of the DT and RF models, we used the default hyperparameter values as provided by the respective implementations. For LR, we set the maximum number of iterations (max_iter) to 1,200,000. For SVMs, we used a gamma value of "scale" and enabled probability estimation by setting the probability parameter to True. The KNNs model utilized a value of 10 for the number of neighbors (n_neighbors). For AdaBoost, we set the number of estimators (n_estimators) to 100. In the case of GB, we used 100 estimators, a learning rate of 1.0, and a maximum depth of 1. Lastly, the XGBoost model employed 100 estimators, a learning rate of 1.0, and a maximum depth of 30.

Table 2. Hyperparameter tuning values for ML models.

Sl.No.	ML model	Hyperparameter	Values
1	Decision tree	Default	Default
2	Random forest	Default	Default
3	Logistic regression	Max iterations (max_iter)	1,200,000
4	Support vector machines	Gamma (gamma)	"Scale"
		Probability (probability)	True
5	K-nearest neighbors	n_neighbors	10
6	AdaBoost	n_estimators	100
7	Gradient boosting	n_estimators	100
		Learning rate (learning_rate)	1.0
		Max depth (max_depth)	1
8	XGBoost	n_estimators	100
		Learning rate (learning_rate)	1.0
		Max depth (max_depth)	30

AdaBoost: adaptive boosting; ML: machine learning; XGBoost: extreme gradient boosting.

Results of Dataset 1

The Pima Indian Diabetes Dataset is one of the most useful datasets for testing ML algorithms for predicting diabetes in the general population.[36] This dataset was provided by the National Institute of Diabetes and Digestive and Kidney Diseases and is used to determine whether a patient has diabetes based on diagnostic measures such as pregnancy, glucose level, blood pressure, skin thickness, diabetes pedigree function, insulin, body mass index (BMI), and age. The attributes of the Pima Indian Diabetes Dataset are listed below:

1. Pregnancies: Number of occurrences of pregnancy
2. Glucose: In a glucose tolerance measure, the plasma glucose concentration after 2 h
3. Skin thickness: The thickness of the skin folds on the triceps (mm)
4. Insulin: Serum insulin (μU/ml) after 2 h
5. BMI: Body mass index
6. Age: Age of the person in years
7. Outcome: Class variable as a result (0 or 1)

The boxplot in Figure 2(a) shows that the dataset contains outliers, whereas Figure 2(b) shows clean data after applying the preprocessing algorithm. In the boxplot, different features have multiple outliers data which are indicated by multiple diamond signs beside each feature and after handling the outlier the boxplot looks like no diamond signs on each feature which proves no outlier existed.

Figure 2(d) depicts the balanced dataset distribution of the original imbalance dataset in Figure 2(c), where label 0 represents no diabetes and 1 represents diabetes. In the pie chart, we can see that it contains more portion of 1 than 0 and after random oversampling, we can see that we have the same portion of labels 0 and 1, ensuring data is balanced now.

Figure 3(a) and (b) presents the accuracy and MSE of our experiments for Dataset 1. The accuracy comparison before and after applying the proposal. The accuracy results of DT, RF, LR, SVM, KNN, AdaBoost, GB and XGBoost are 77.27%, 85.53%, 78.95%, 78.95%, 80.26%, 80.26%, 81.58%, and 83%, respectively. We found that, depending on the ML algorithms, accuracy performance increases from 4.95% to 12.15%. On the other hand, the MSE values of the algorithm reduced significantly, 5.27%, 12.15%, 5.95%, 4.95%, 8.26%, 6.23%, 8.85%, and 10.27% for DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost, respectively.

Table 3 summarizes the other performance metrics. We found that the proposal can improve the precision values from 3.03% to 15.98% for any ML approach. An efficient data preprocessing and data balancing can improve the data quality; hence ML algorithms can accurately classify the test data. We found similar results for recall; the values increased from 0.41% to 11.94%. The F1-score also improved as expected, from 1.1% to 12.25%. On the other hand, the table also indicates that due to the high performance of the proposal, the values of MAE and RMSE are reduced significantly. It is observed that MAE values reduced from 4.95% for SVM to 12.15% for RF. Similarly, RMSE reduced greatly from 5.11% to 13.56%.

The extensive analysis of the experiments shows that RF outperforms others as it is an ensemble classifier that works by training a large number of DTs, leading to higher accuracy with good reliable predictions than other single algorithms. So, we consider this model as our proposed model.

Figure 4(a) depicts the confusion matrix and Figure 4(b) shows the receiver operative curve (ROC) curve for our proposed model. In the confusion matrix, the large number of TP and TN than FP and FN is a very crucial point to be a better prediction model for ML. In Figure 4(a), we can see that the TP and FN rates are high and FN and FP rates are very low which gives a better sign of diabetes prediction. The TP, TN, FP, and FN rates are 45%, 41%, 8%, and 6%, respectively. On the other hand, in the ROC, the value of AUC which is near 1 is the best model for ML. In Figure 4(b), we can see that the AUC value is 0.93 (93.07%), which means the positive and negative labels are mostly segregated, and the model is effective.

After analyzing multiple performance indicators, we can find that among all the ML algorithms RF outperforms others with an accuracy rate of 86%, 14% MSE rate, and 8% FP, as well as 6% FN rates.

Results of Dataset 2

Dataset 2[37] namely, Austin public health diabetes self-management education participant demographics 2015—2017, contains demographic information collected from Austin Public Health, Austin. Among the 1688 rows and 25 columns; various important attributes are considered for our experiment including diabetes status, health indicators, health behaviors including race/ethnicity diabetes status, heart disease, high blood pressure, tobacco use, previous diabetes education, diabetes knowledge, fruits and vegetable consumption, sugar-sweetened beverage consumption, and so on. The attributes of the Austin Public Health dataset are listed below:

1. Age: Age in year
2. Gender: Gender of the patient
3. Race/ethnicity: Race/ethnicity of participant
4. Heart disease: Heart disease diagnosis (yes/no)
5. High blood pressure: High blood pressure diagnosis (yes/no)
6. Tobacco use: Tobacco user (yes/no)
7. Previous diabetes education: Previous diabetes education reported by participant (yes/no)
8. Diabetes knowledge: Self-reported knowledge of diabetes (poor/fair/good)
9. Fruits and vegetable consumption: Fruits and/or vegetables eaten each week
10. Sugar-sweetened beverage consumption: Sugar-sweetened beverages consumed each week
11. Food measurement: Number of times food was measured each week
12. Carbohydrate counting: Number of times carbohydrates were counted each week
13. Exercise: Number of days participant exercised each week
14. Diabetes status: Diabetes status (yes/no) of participant

The boxplot in Figure 5(a) shows that the dataset contains outliers, whereas Figure 5(b) shows clean data after applying the preprocessing algorithm. In the boxplot, different features have multiple outliers data which is indicated by multiple diamond signs beside each feature and after handling the outlier the boxplot looks like no diamond signs on each feature which proves no outlier existed.

Figure 5(d) depicts the balanced dataset distribution of the original imbalance dataset Figure 5(c), where the label “No diabetes” represents no diabetes and “Diabetes”

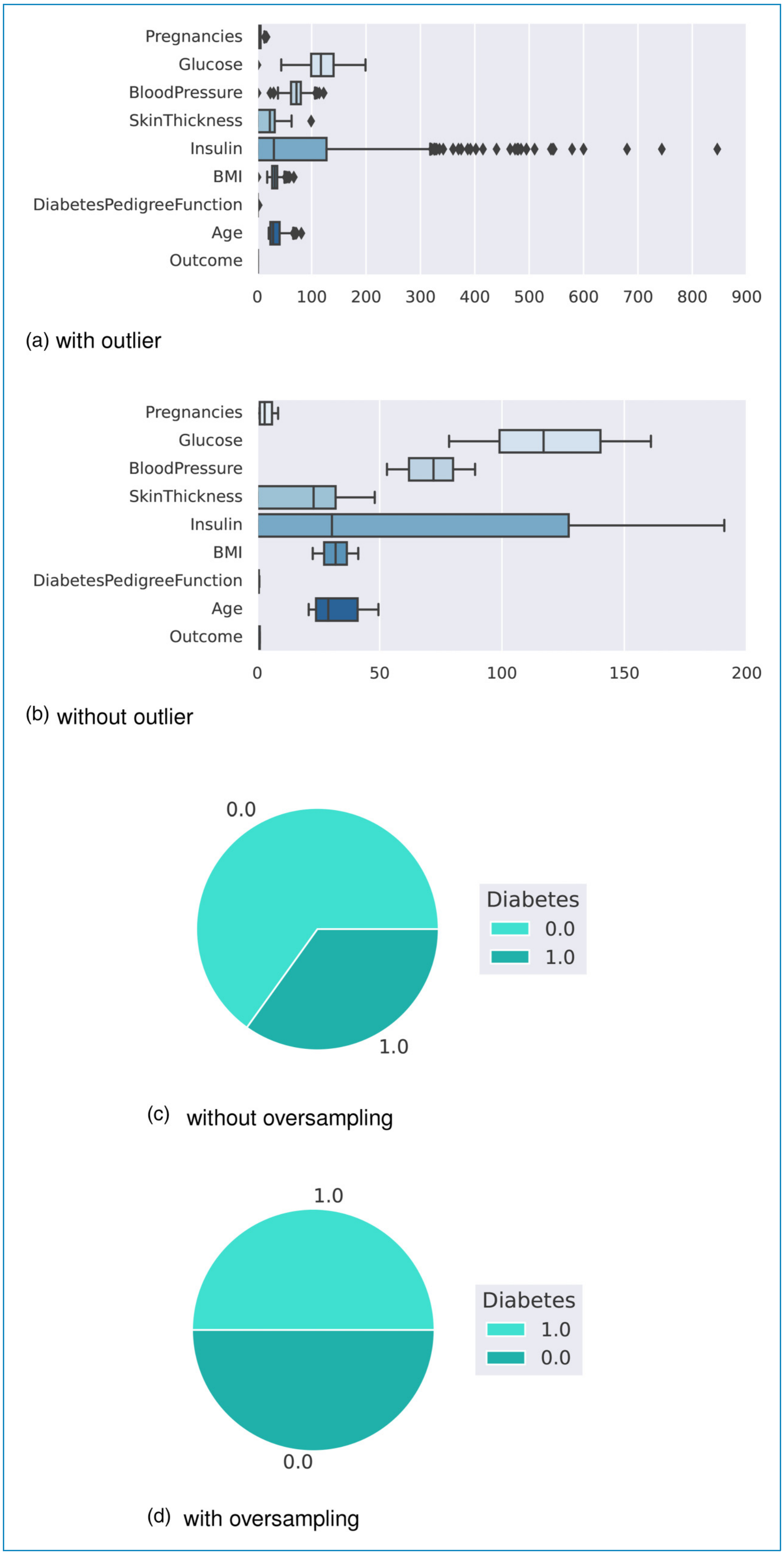
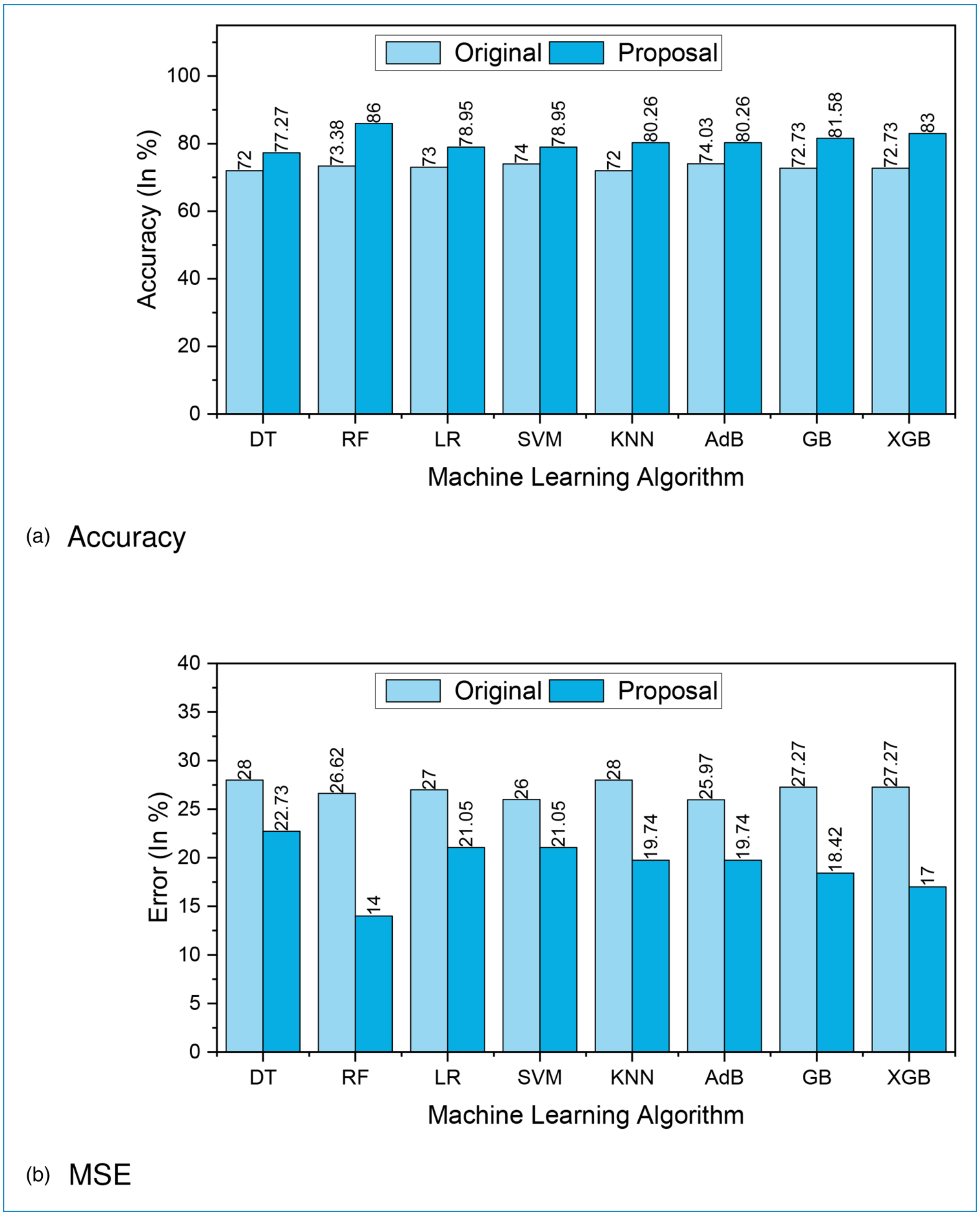


Figure 2. Before and after outlier removal and oversampling results: (a) with outlier; (b) without outlier; (c) without oversampling; and (d) with oversampling.



(a) Accuracy

(b) MSE

Figure 3. The performance results of Dataset 1: (a) accuracy and (b) mean-squared error (MSE).

represents diabetes. In the pie chart, we can see that it contains more portion of "No diabetes" than "Diabetes" and after random oversampling, we can see that we have an equal portion of the labels "No diabetes" and "Diabetes" which ensures data is balanced now.

Figure 6(a) and (b) presents the accuracy and MSE of our experiments for Dataset 2. The accuracy comparison before and after applying the proposal. The accuracy results of DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost are 95.45%, 98.48%, 83.33%, 93.94%, 87.88%, 96.97%, 93.94%, and 93.94%, respectively. We found that, depending on the ML algorithms, accuracy performance increases from 0% to 8.74%. On the other hand, the MSE values of the algorithm reduced significantly, 5.71%, 8.74%, 0%, 5.48%, 1.52%, 5.94%, 4.2%, and 1.63% for DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost, respectively.

The other performance metrics are summarized in Table 4. We found that, for any ML approach, the proposal can improve the precision values from 0% to 14.41%. An efficient data preprocessing and data balancing can improve the data quality, hence ML algorithms can accurately classify the test data. We found similar results for recall, the values increase from 8.27% to 20.57%. The F1-score also improved as expected from 0.47% to 14.19%. On the other hand, the table also indicates that

Table 3. Performance result of Dataset 1 (Pima Indian dataset).

	Precision		Recall		F1-score		MAE		RMSE	
ML	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal
DT	72.32	75.35	72.11	76.26	71.96	75.72	28	22.73	52.92	47.67
RF	70.98	85.99	70.81	85.99	70.89	85.99	26.62	14	51.6	37.42
LR	73	78.95	73.01	73.7	73	75.06	27	21.05	51.96	45.88
SVM	74.34	78.95	73.7	74.11	73.96	75.06	26	21.05	50.99	45.88
KNN	71.99	82.92	71.99	73.89	71.99	75.68	28	19.74	52.92	44.43
AdaBoost	71.68	78.85	71.31	77.21	71.48	77.87	25.97	19.74	50.96	44.43
GB	70.39	79.89	70.71	79.89	70.54	79.89	27.27	18.42	52.22	42.92
XGBoost	70.51	83.09	71.11	83.05	70.75	83	27.27	17	52.22	41.23

ML: machine learning; MAE: mean absolute error; RMSE: root mean-squared error; DT: decision tree; RF: random forest; LR: logistic regression; SVM: support vector machine; KNN: K-nearest neighbor; AdaBoost: adaptive boosting; GB: gradient boosting; XGBoost: extreme gradient boosting.

due to the high performance of the proposal, the values of MAE and RMSE are reduced significantly. It is observed that MAE values reduced by 8.74% for RF to 0% for LR. Similarly, RMSE reduced greatly from 19.72% to 0%.

The extensive analysis of the experiments shows that RF outperforms others so we consider this model as our proposed model. Figure 7(a) depicts the confusion matrix and Figure 7(b) shows the ROC curve for our proposed model. In the confusion matrix, the large number of TP and TN than FP and FN is a very crucial point to be a better prediction model for ML. In Figure 7(a), we can see that the TP and FN ratess are high and FN and FP rates are very low, which gives a better sign of diabetes prediction. The TP, TN, FP and FN rates are 56.06%, 42.42%, 1.52%, and 0%, respectively. On the other hand, in the ROC curve, the value of AUC which is near 1 is the best model for ML. In Figure 7(b), we can see that the AUC value is 0.99 (99.35%) which means the positive and negative labels are mostly segregated, and the model is effective.

After analyzing multiple performance indicators, we can find that among all the ML algorithms RF outperforms others with an accuracy rate of 98.48%, 0% MSE rate, and 1.52% FP as well as 0% FN rates.

Results of Dataset 3

Diabetes 3[38] conducted a survey and collected a dataset containing 950 records and 19 attributes that have a measurable influence on diabetes such as family diabetes history, blood pressure, exercise, BMI, smoking level, alcohol consumption, sleeping hours, food habits, pregnancy, urination frequency, stress level, and so on. The attributes of the survey dataset are listed below:

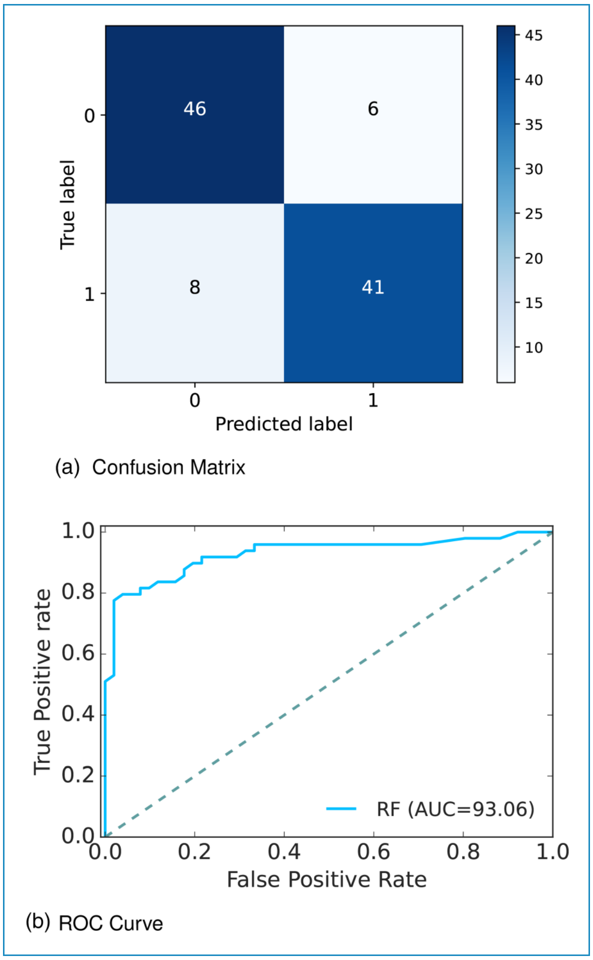
1. Age: Age in year
2. Gender: Gender of the participant
3. Family_Diabetes: Family history with diabetes
4. highBP: Diagnosed with high blood pressure
5. PhysicallyActive: Walk/run/physically active
6. BMI: Body mass index
7. Smoking: Smoking
8. Alcohol: Alcohol consumption
9. Sleep: Hours of sleep
10. SoundSleep: Hours of sound sleep
11. RegularMedicine: Regular intake of medicine
12. JunkFood: Junk food consumption
13. Stress: Not at all, sometimes, often, always
14. BPLevel: Blood pressure level
15. Pregancies: Number of pregnancies
16. Pdiabetes: Gestation diabetes
17. UriationFreq: Frequency of urination
18. Diabetic: Yes or no

We found a total of 48 missing values in the original dataset, including four missing values for BMI, 42 missing values for Pregnancies, one missing value for Pdiabetes, and one missing value for Diabetic.

The boxplot in Figure 8(a) shows that the dataset contains outliers, whereas Figure 8(b) shows clean data after applying the preprocessing algorithm. In the boxplot, different features have multiple outliers data which is indicated by multiple diamond signs beside each feature and after handling the outlier, the boxplot looks like no diamond signs on each feature, which proves no outlier existed.

Figure 8(d) depicts the balanced dataset distribution of the original imbalance dataset 8(c). where the label "No" represents no diabetes and "Yes" represents diabetes. In the pie chart, we can see that it contains more portion of "No" than "Yes," and after random oversampling, we can see that we have an equal portion of the labels "No" and "Yes," which ensures data is balanced now.

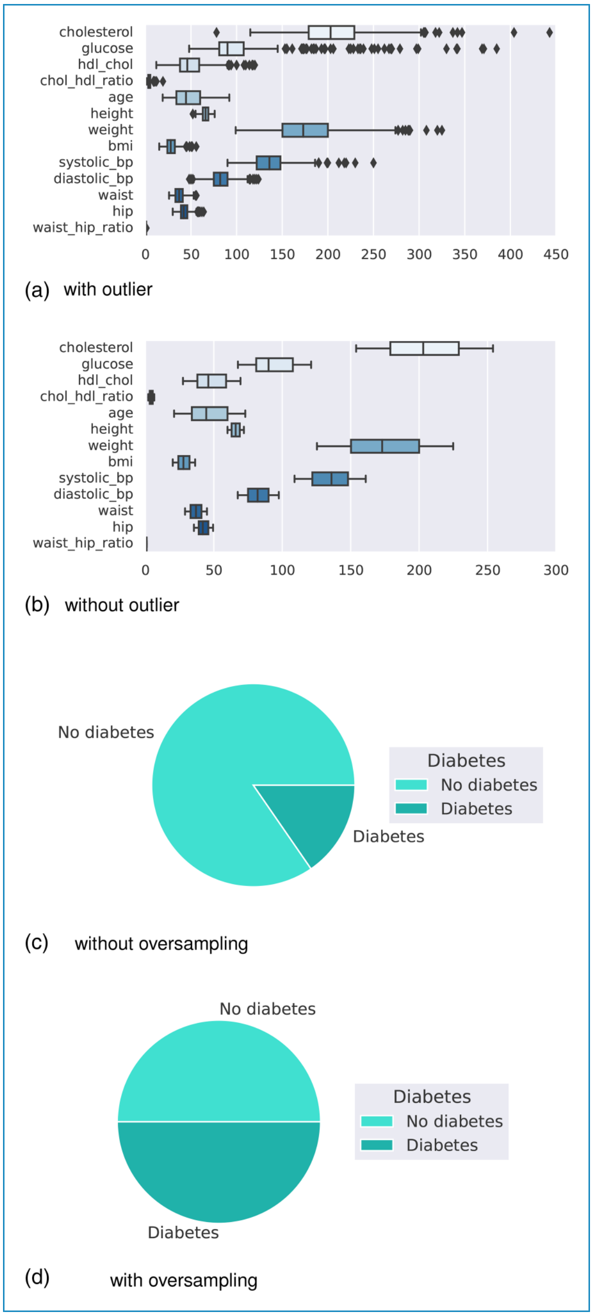
Figure 9(a) and (b) presents the accuracy and MSE of our experiments for Dataset 3. The accuracy comparison before and after applying the proposal. The accuracy results of DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost are 98.54%, 97.81%, 91.11%, 91.97%, 88.32%, 91.58%, 92.63%, and 99.27%, respectively. We found that, depending on the ML algorithms, accuracy performance increases from 2.25% to 10.75%. On the other hand, the MSE values of the algorithm reduced significantly, 2.98%, 2.25%, 8.63%, 10.75%, 7.66%, 3.26%, 4.31%, and 3.71% for DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost, respectively.



(a) Confusion Matrix

(b) ROC Curve

Figure 4. Confusion matrix and receiver operative curve (ROC) curve of Dataset 1: (a) confusion matrix and (b) ROC curve.



(a) with outlier

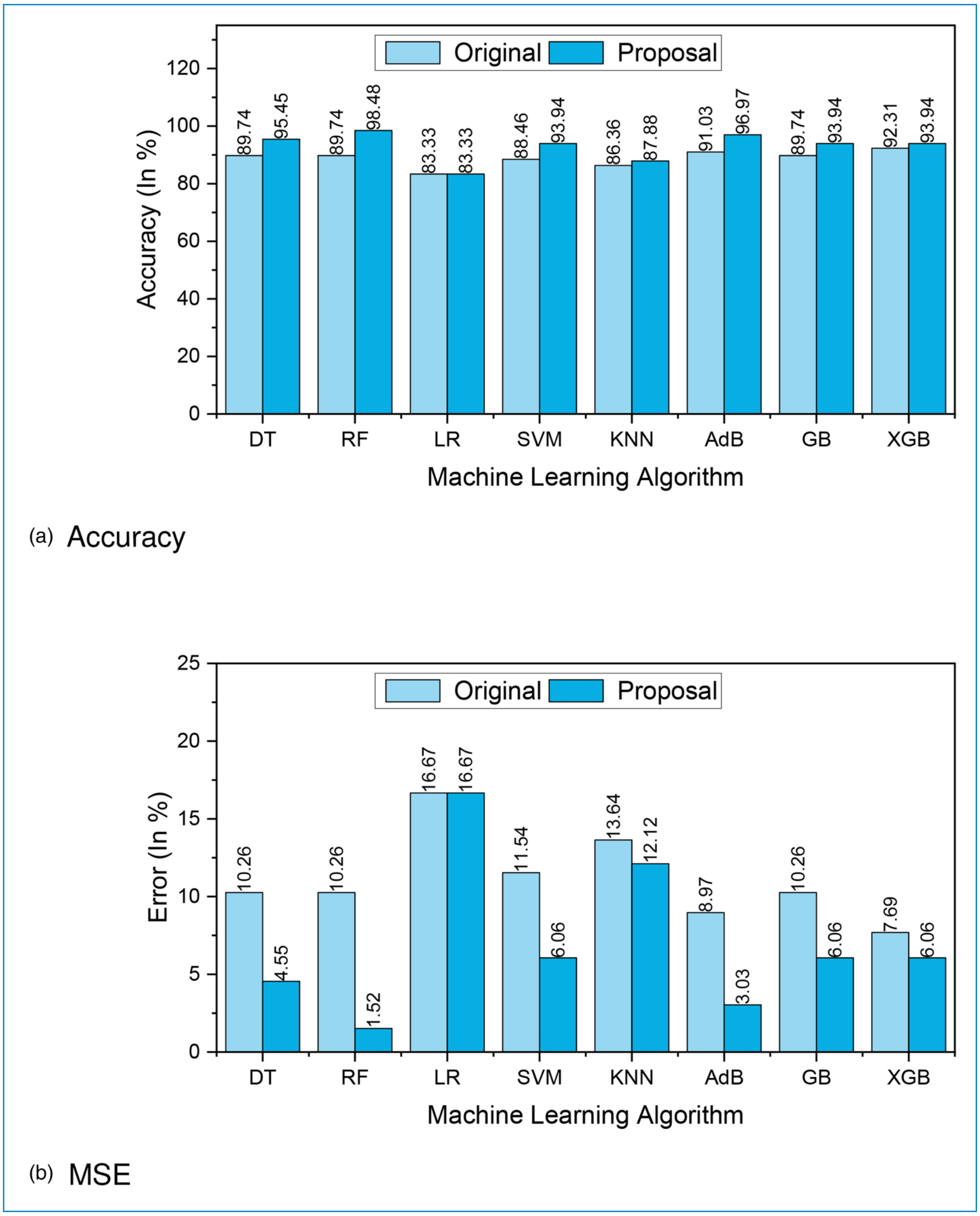
(b) without outlier

(c) without oversampling

(d) with oversampling

Figure 5. Before and after outlier removal and oversampling results: (a) with outlier; (b) without outlier; (c) without oversampling; and (d) with oversampling.

The other performance metrics are summarized in Table 5. We found that, for any ML approach, the proposal can improve the precision values from 1.77% to 9.20%. An efficient data preprocessing and data balancing can improve the data quality, hence ML algorithms can accurately classify the test data. We found similar results for recall, the values increase from 2.06% to 25.34%. The F1-score also improved as expected from 1.99% to 22.74%. On the other hand, the table also indicates that due to the high



(a) Accuracy

(b) MSE

Figure 6. The performance results of Dataset 2: (a) accuracy and (b) mean-squared error (MSE).

performance of the proposal, the values of MAE and RMSE are reduced significantly. It is observed that MAE values reduced from 2.25% for RF to 10.75% for SVM. Similarly, RMSE reduced greatly from 5.15% to 15%.

The extensive analysis of the experiments shows that XGBoost outperforms others as it is a gradient-boosted DT solution that uses L1 and L2 regularization which leads to getting a better accuracy rate than others. So, we consider this model as our proposed model.

Figure 10(a) depicts the confusion matrix and Figure 10(b) shows the ROC curve for our proposed model. In the confusion matrix, the large number of TP and TN than FP and FN is a very crucial point to be a better prediction model for ML. In graph 10(a), we can see that the TP and FN rates are high and the FN and FP rates are very low, which gives a better sign of diabetes prediction. The TP, TN, FP, and FN rates are 49.64%, 44.53%, 3.65%, and 2.19%, respectively. On the other hand, in the ROC curve, the value of AUC which is near 1 is the best model for ML. In Figure 10(b), we can see that the AUC value is 0.99 (99.36%) which means the positive and negative labels are almost segregated, and the model is effective.

After analyzing multiple performance indicators, we can find that among all the ML algorithms XGBoost outperforms others with an accuracy rate of 99.27%, 0.73% MSE rate, and 3.65% FP as well as 2.19% FN rates.

Table 4. Performance result of Dataset 2 (Austin dataset).

	Precision		Recall		F1-score		MAE		RMSE	
ML	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal
DT	84.27	95.28	75	95.57	82.67	95.4	10.26	4.55	32.03	21.32
RF	84.27	98.68	84.27	98.28	84.27	98.46	10.26	1.52	32.03	12.31
LR	83.06	83.06	75	83.27	82.67	83.14	16.67	16.67	40.82	40.82
SVM	83.85	94.3	75	93.48	80.94	93.8	11.54	6.06	33.97	24.62
KNN	86.11	87.65	75	88.07	82.67	87.78	13.64	12.12	36.93	34.82
AdaBoost	85.78	96.92	87.4	96.92	86.55	96.92	8.97	3.03	29.96	17.41
GB	84.27	93.73	84.27	94.22	84.27	93.89	10.26	6.06	32.03	24.62
XGBoost	87.22	93.73	75	94.22	82.67	93.89	7.69	6.06	27.74	24.62

ML: machine learning; MAE: mean absolute error; RMSE: root mean-squared error; DT: decision tree; RF: random forest; LR: logistic regression; SVM: support vector machine; KNN: K-nearest neighbor; AdaBoost: adaptive boosting; GB: gradient boosting; XGBoost: extreme gradient boosting.

Results of Dataset 4

Finally, the information was collected from Iraqi society, as well as the Medical City Hospital's laboratory and Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital in Dataset 4.[39] The data consist of medical information as well as laboratory analysis including age, gender, creatinine ratio (Cr), BMI, urea, cholesterol, low-density lipoprotein (LDL), very low-density lipoprotein (VLDL), triglycerides (TG) and high-density lipoprotein (HDL) cholesterol, hemoglobin A1c (HBA1c), and so on. The attributes of the Iraqi Medical City dataset are listed below:

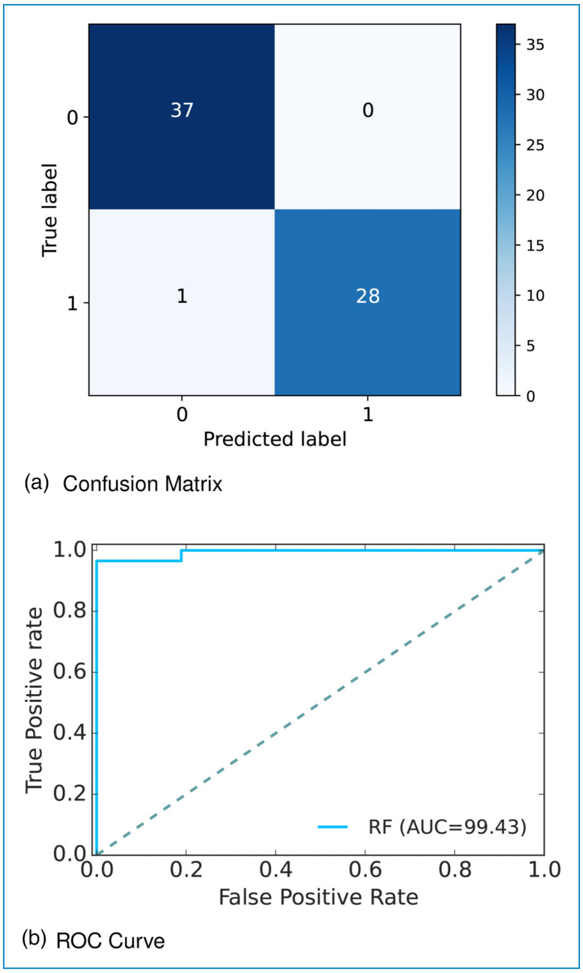
1. Age: Age of the patient
2. Gender: Gender of the participant
3. Sugar level blood: Sugar level of the patient
4. Cr: Creatinine ratio
5. BMI: Body mass index
6. Urea: blood urea level
7. Chol: Cholesterol
8. TG: triglycerides level
9. HDL: High-density lipoprotein cholesterol level
10. LDL: Low-density lipoprotein level
11. VLDL: very low-density lipoprotein level
12. HBA1c: Average blood glucose (sugar)–hemoglobin A1C
13. Class: Diabetic, non-diabetic, or pre-diabetic

The boxplot in Figure 11(a) shows that the dataset contains outliers, whereas Figure 11(b) shows clean data after applying the preprocessing algorithm. In the boxplot, different features have multiple outliers data, which is indicated by multiple diamond signs beside each feature and after handling the outlier the boxplot looks like no diamond signs on each feature which proves no outlier existed.

Figure 11(d) depicts the balanced dataset distribution of the original imbalance dataset in Figure 11(c). where the label "Y" represents diabetic, "N" represents non-diabetic and "P" represents pre-diabetic. In the pie chart, we can see that it contains more portion of "Y" than "P" and "N," and after random oversampling, we can see that we have an equal portion of all the labels which ensures data is balanced now.

The research specifically aims to develop an ML model for classifying diabetes, focusing on the task of assigning diabetes labels (diabetes or no diabetes) using various diabetes datasets. While it may seem odd to include HbA1c as an input variable in this dataset (Dataset 4), we considered it relevant for our classification task. The inclusion of HbA1c as a feature helps the model learn patterns and relationships between other variables and the presence of diabetes. By including this feature, we aim to capture additional information that may contribute to the accurate classification of diabetes. We understand that the close-to-perfect performance of the model might raise suspicions. However, we assure you that our research was conducted rigorously, following standard practices and using appropriate evaluation metrics.

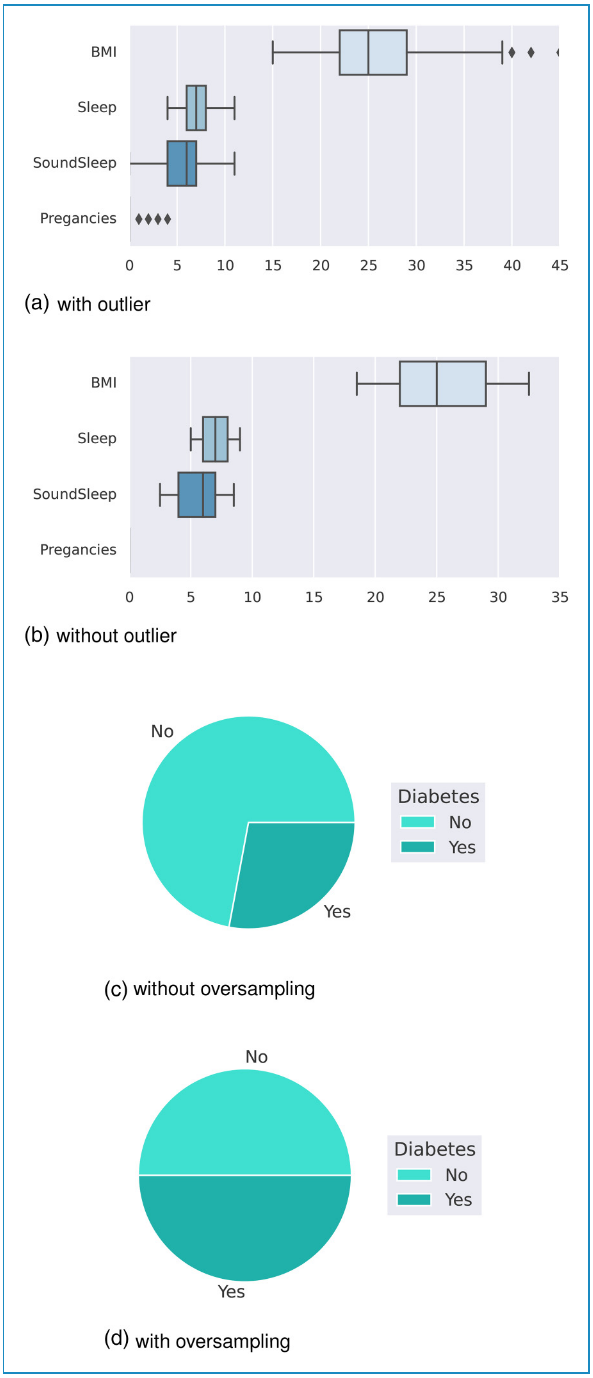
Figure 12(a) and (b) presents the accuracy and MSE of our experiments for Dataset 4. The accuracy comparison before and after applying the proposal. The accuracy results of DT, RF, LR, SVM, KNN, AdaBoost, GB, and



(a) Confusion Matrix

(b) ROC Curve

Figure 7. Confusion matrix and receiver operative curve (ROC) curve of Dataset 2: (a) confusion matrix and (b) ROC curve.



(a) with outlier

(b) without outlier

(c) without oversampling

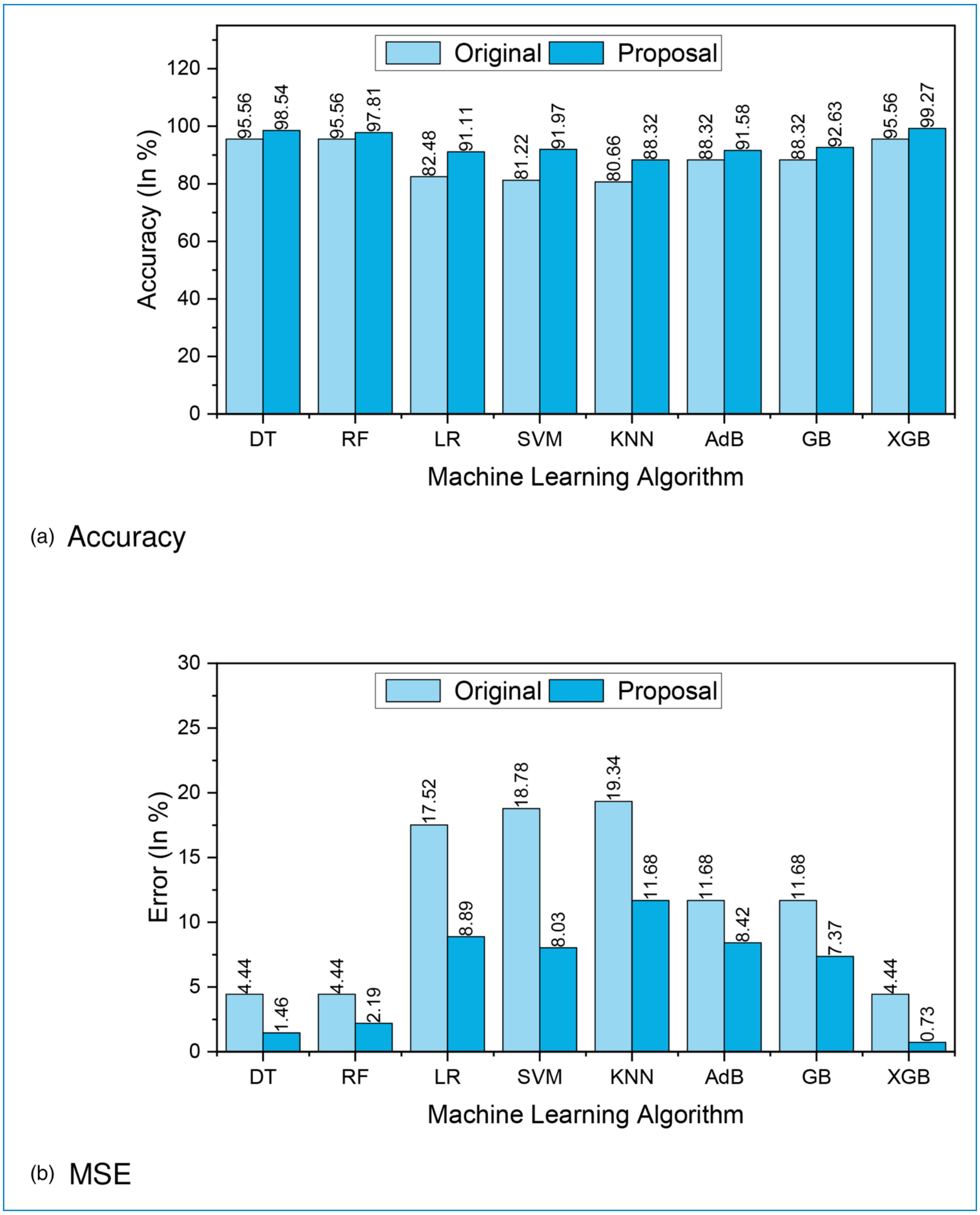
(d) with oversampling

Figure 8. Before and after outlier removal and oversampling results: (a) with outlier; (b) without outlier; (c) without oversampling; and (d) with oversampling.

XGBoost are 100%, 99.60%, 95.65%, 96.84%, 95.65%, 99.21%, 99.60%, and 99%, respectively. We found that, depending on the ML algorithms, accuracy performance increases from 0.19% to 9.84%. On the other hand, the MSE values of the algorithm reduced significantly, 2%, 1.6%, 11.09%, 21.84%, 18.28%, 3.71%, 7.6%, and 0.19% for DT, RF, LR, SVM, KNN, AdaBoost, GB, and XGBoost, respectively.

The other performance metrics are summarized in Table 6. We found that, for any ML approach, the proposal can improve the precision values of 0.67% to 46.88%. An efficient data preprocessing and data balancing can improve the data quality, hence ML algorithms can accurately classify the test data. We found similar results for recall, the values increased from 0.83% to 47.59%. The F1-score also improved as expected from 0.42% to 46.52%. On the other hand, the table also indicates that due to the high performance of the proposal, the values of MAE and RMSE are reduced significantly. It is observed that MAE values reduced from 0.19% for XGBoost to 13.84% for SVM. Similarly, RMSE reduced greatly from 0.89% to 32.22%.

The extensive analysis of the experiments shows that DT outperforms others since the ability to capture relevant decision-making information from the available dataset is the most important feature of the DT which leads to higher accuracy. So, we consider this model as our proposed model.



(a) Accuracy

(b) MSE

Figure 9. The performance results of Dataset 3: (a) accuracy and (b) mean-squared error (MSE).

Figure 13(a) depicts the confusion matrix and Figure 13(b) shows the ROC curve for our proposed model. In the confusion matrix, the large number of TP and TN than FP and FN is a very crucial point to be a better prediction model for ML. In Figure 13(a), we can see that the TP and FN rates are high and FN and FP rates are very low which gives a better sign of diabetes prediction. The TP, TN, FP and FN rates are 35.97%, 64.03%, 0.0%, and 0.0%; 29.25%, 70.75%, 0.0%, and 0.0%; 34.78%, 65.22%, 0.0%, and 0.0% for N, Y, and P class, respectively. On the other hand, in the ROC, the value of AUC which is near 1 is the best model for ML. In the Figure 13(b), we can see that the AUC value is 1 (100%), which means the positive and negative labels are completely segregated, and the model is as effective as it can be.

After analyzing multiple performance indicators, we can find that among all the ML algorithms DT outperforms others with an accuracy rate of 100%, 0% MSE rate, and 0% FP as well as FN rates.

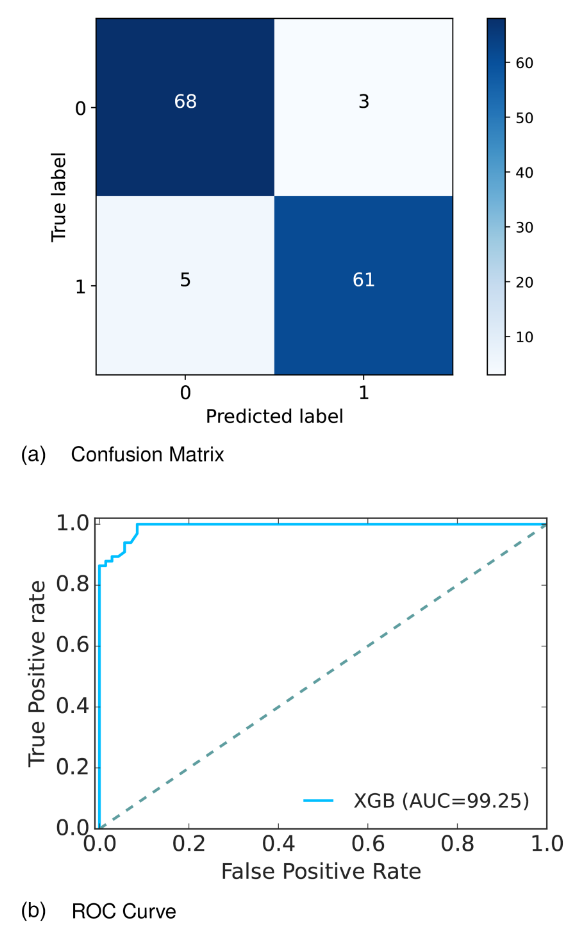
Discussion

In this study, we conducted a comprehensive analysis of ML models for diabetes detection using four distinct datasets: Pima Indian, Austin Public, Tigga, and Mendeley. The performance metrics were evaluated to assess the effectiveness

Table 5. Performance result of Dataset 3 (Tigga dataset).

	Precision		Recall		F1-score		MAE		RMSE	
ML	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal
DT	95.52	98.54	92.99	98.54	94.16	98.54	4.44	1.46	21.08	12.08
RF	94.32	97.79	94.32	97.83	94.32	97.81	4.44	2.19	21.08	14.8
LR	82.6	87.5	82.35	92.61	82.41	89.42	17.52	8.89	41.85	29.81
SVM	84.86	91.99	66.59	91.93	69.22	91.96	18.78	8.03	43.34	28.34
KNN	79.11	88.31	68.14	88.31	70.57	88.31	19.34	11.68	43.97	34.17
AdaBoost	88.49	90.26	88.2	90.26	88.27	90.26	11.68	8.42	34.17	29.02
GB	88.37	91.2	88.25	91.92	88.29	91.55	11.68	7.37	34.17	27.14
XGBoost	95.52	99.31	92.99	99.24	94.16	99.27	4.44	0.73	21.08	8.54

ML: machine learning; MAE: mean absolute error; RMSE: root mean-squared error; DT: decision tree; RF: random forest; LR: logistic regression; SVM: support vector machine; KNN: K-nearest neighbor; AdaBoost: adaptive boosting; GB: gradient boosting; XGBoost: extreme gradient boosting.



(a) Confusion Matrix

(b) ROC Curve

Figure 10. Confusion matrix and receiver operative curve (ROC) curve of Dataset 3: (a) confusion matrix and (b) ROC curve.

of the models in accurately identifying individuals at risk of diabetes as shown in Table 7.

On the Pima Indian dataset, the RF model achieved an accuracy of 85.53%. The precision and recall were 86.96% and 81.29%, respectively, resulting in an F1-score of 83.06%. The MAE and MSE were 14.47, and the RMSE was 38.04. For the Austin Public dataset, the RF model exhibited exceptional performance with an accuracy of 98.48%. The precision, recall, and F1-score values were 98.68%, 98.28%, and 98.46%, respectively. The model demonstrated low errors with an MAE of 1.52, MSE of 1.52, and RMSE of 12.31. The Tigga dataset showed outstanding results with the XGBoost model achieving an accuracy of 99.27%. The precision, recall, and F1-score values were 99.31%, 99.24%, and 99.27%, respectively. The model's errors were minimal with an MAE of 0.73, MSE of 0.73, and RMSE of 8.54. Remarkably, the DT model achieved perfect performance on the Mendeley dataset, attaining an accuracy, precision, recall, and F1-score of 100%. Additionally, the model exhibited zero errors with an MAE, MSE, and RMSE of 0.

Interestingly, the DT model achieves perfect performance on the Mendeley dataset, with accuracy, precision, recall, and F1-score all reaching 100%. This remarkable result suggests that the features within the Mendeley dataset may be well-suited for DT-based classification, possibly due to the dataset's inherent structure or the nature of the variables involved. It is worth noting that while the DT model demonstrates flawless performance on this particular dataset, its generalization to other datasets may vary, warranting further investigation into its robustness across different data domains. Moreover, the low error metrics

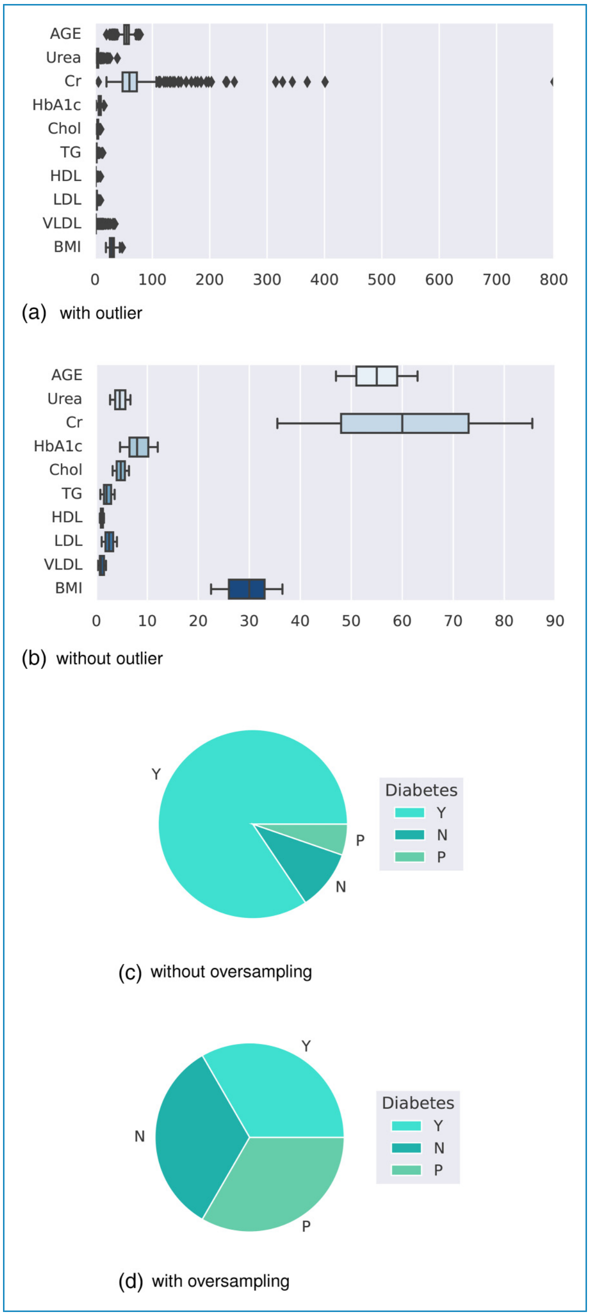


Figure 11. Before and after outlier removal and oversampling results: (a) with outlier; (b) without outlier; (c) without oversampling; and (d) with oversampling.

(MAE, MSE, and RMSE) observed across all models and datasets indicate the models' capability to make accurate predictions with minimal deviation from the actual values. These findings emphasize the reliability of ML algorithms in diabetes detection tasks and underscore their potential utility in clinical settings for early risk assessment and intervention.

Furthermore, the results of this study highlight the efficacy of ML models in diabetes detection across diverse datasets. While certain models excel in specific contexts, the overall performance underscores the promise of ML techniques in augmenting traditional diagnostic approaches and improving patient outcomes in diabetes management. Further research is warranted to explore the generalizability of these models across larger and more diverse populations, as well as their integration into clinical practice for personalized healthcare delivery.

The results indicate that the ML models can effectively detect diabetes. The RF model showed good performance on the Pima Indian dataset, while both the RF model on the Austin Public dataset and the XGBoost model on the Tigga dataset demonstrated excellent performance. The DT model exhibited perfect performance on the Mendeley dataset. These findings highlight the potential of ML in accurate diabetes detection, providing a valuable tool for early intervention and improved patient outcomes.

In our research, we included multiple datasets in our analysis to provide a comprehensive evaluation of the performance of ML models for diabetes detection. Each dataset represents a distinct population or data source, allowing us to assess the generalizability of the models across diverse scenarios. We evaluated the models on multiple datasets to gain insights into their strengths and limitations in different contexts. This approach helps experimenters understand the robustness of the models and identify potential challenges or biases that may arise when applying them to real-world scenarios. Additionally, it enables researchers to make informed decisions about which models are most suitable for specific datasets or patient populations. Regarding the conclusions, we acknowledge that the original discussion did not sufficiently elaborate on the insights gained from the extensive set of benchmarks. In light of the reviewer's comment, we will revise the conclusion section to provide a more comprehensive analysis of the results and their implications. We will discuss the key findings from each dataset, highlight the factors that contributed to successful performance, and address the challenges and considerations experimenters should be aware of when deploying these models in practice. We will also emphasize any new information or novel observations that emerged from our study. Although some previous research has examined ML models for diabetes detection, our study contributes by analyzing a diverse range of datasets and comparing the performance of multiple models. This allows us to provide a more comprehensive understanding of the strengths and weaknesses of different algorithms and their applicability in various scenarios.

Furthermore, a comparison analysis is illustrated in Table 8 and in Figure 14, where we can see that our proposed approach outperforms others, which proves the better prediction models. In Dataset 1, RF outperforms other ML models with an accuracy rate of 86%. Similarly, for Dataset 2, Dataset 3, and Dataset 4, RF, XGBoost, and DT outperform other ML models with an

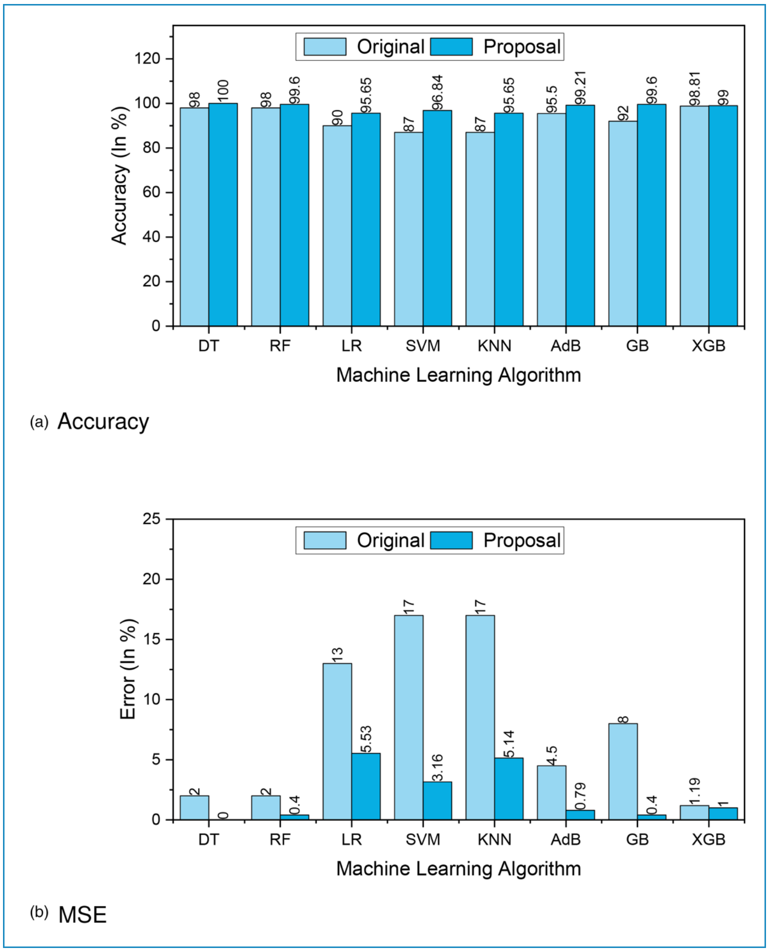


Figure 12. The performance results of Dataset 4: (a) accuracy and (b) mean-squared error (MSE).

Table 6. Performance result of Dataset 4 (Mendeley dataset).

	Precision		Recall		F1-score		MAE		RMSE	
ML	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal	Original	Proposal
DT	93.19	100	96.44	100	94.61	100	2	0	14.14	0
RF	92.94	99.63	96.83	99.55	94.45	99.58	2	0.4	14.14	6.29
LR	67.48	96.17	70.63	95.3	68.46	95.55	13	5.53	43.59	28.12
SVM	53.26	97.16	48.81	96.4	50.63	96.65	17	3.16	50	17.78
KNN	49.22	96.1	48.81	95.21	48.97	95.49	17	5.14	50	25.92
AdaBoost	79.44	99.26	98.27	99.1	85	99.17	4.5	0.79	21.21	8.89
GB	86.88	99.64	89.68	99.55	88.14	99.59	8	0.4	28.28	6.29
XGBoost	98.94	99.61	97.22	98.65	98.35	98.77	1.19	1	10.89	10

ML: machine learning; MAE: mean absolute error; RMSE: root mean-squared error; DT: decision tree; RF: random forest; LR: logistic regression; SVM: support vector machine; KNN: K-nearest neighbor; AdaBoost: adaptive boosting; GB: gradient boosting; XGBoost: extreme gradient boosting.

accuracy rate of 98.48%, 99.27%, and 100%, respectively. The higher accuracy rate of diabetes predictions proves the robustness of our proposed model.

The ML models RF, XGBoost, and DT exceed the performance indicators of other algorithms as well as research for Dataset 1, Dataset 2, Dataset 3, and Dataset 4. The proper efficient preprocessing upgrades the quality of data that helps to enhance the outcomes for various datasets. Yet, the RF, XGBoost, and DT algorithms had enhanced accuracy, and it is encouraged that they can be employed in the clinical categorization and prognosis of diabetes for greater performance.

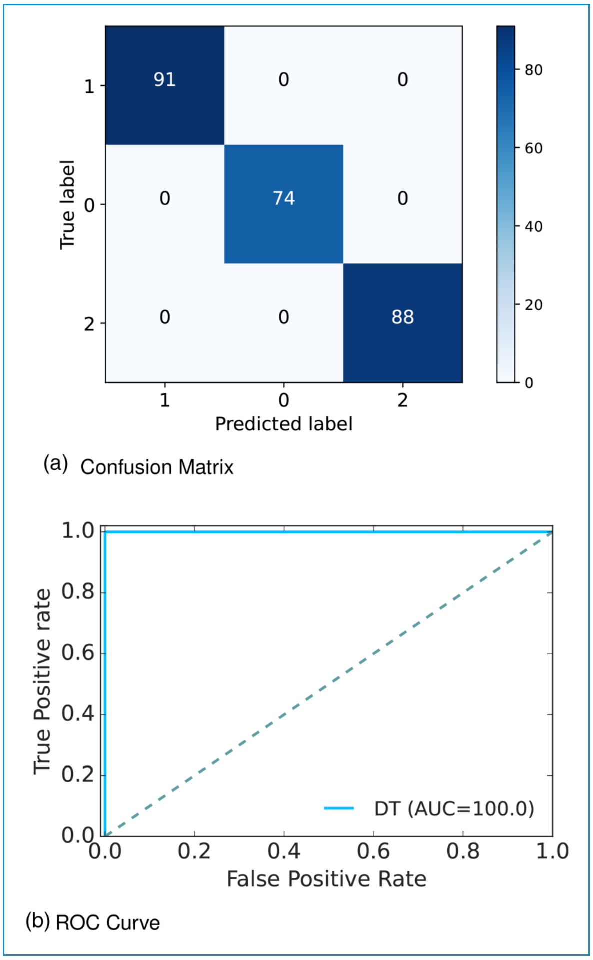


Figure 13. Confusion matrix and ROC curve for RF of Dataset 4: (a) confusion matrix and (b) ROC curve. ROC: receiver operative curve; RF: random forest.

Novelty and significance

This study acknowledges the well-established nature of optimizing preprocessing and oversampling techniques within the specific context of diabetes prediction using ML models. Our contribution lies in carefully implementing and evaluating these methods within the diabetes detection framework.

Context-specific application: Although oversampling techniques are not new, their effectiveness and impact on diabetes prediction may vary depending on the dataset and the specific problem being addressed. By applying these techniques in the context of diabetes detection, we aimed to explore their potential benefits and address the challenges associated with imbalanced class problems specific to this domain.

Comparative analysis: Our study not only employed oversampling techniques, but also performed a comparative analysis of different approaches for handling imbalanced class problems, such as undersampling and synthetic data generation. This analysis aimed to identify the most suitable technique for diabetes prediction and provide insights into the trade-offs and limitations of each approach.

Rigorous evaluation: We implemented a robust evaluation methodology by employing *k*-fold cross-validation. This approach ensures that our models are thoroughly tested on different subsets of the dataset, reducing the risk of overfitting and providing a more reliable assessment of their performance.

While our study introduces experimental evaluation techniques, we believe that the methodological and clinical insights gained from our rigorous analysis contribute to the field of diabetes prediction. By carefully implementing and evaluating oversampling techniques within the diabetes detection framework, we aim to provide practical guidance for researchers and practitioners working on similar problems.

Table 7. Best performance analysis for each dataset.

Dataset	Dataset name	ML model	Accuracy	Precision	Recall	F1-score	MAE	MSE	RMSE
1	Pima Indian	RF	85.53	86.96	81.29	83.06	14.47	14.47	38.04
2	Austin Public	RF	98.48	98.68	98.28	98.46	1.52	1.52	12.31
3	Tigga dataset	XGBoost	99.27	99.31	99.24	99.27	0.73	0.73	8.54
4	Mendeley	DT	100	100	100	100	0	0	0

ML: machine learning; MAE: mean absolute error; MSE: mean-squared error; RMSE: root MSE; RF: random forest; XGB: extreme gradient boosting; DT: decision tree.

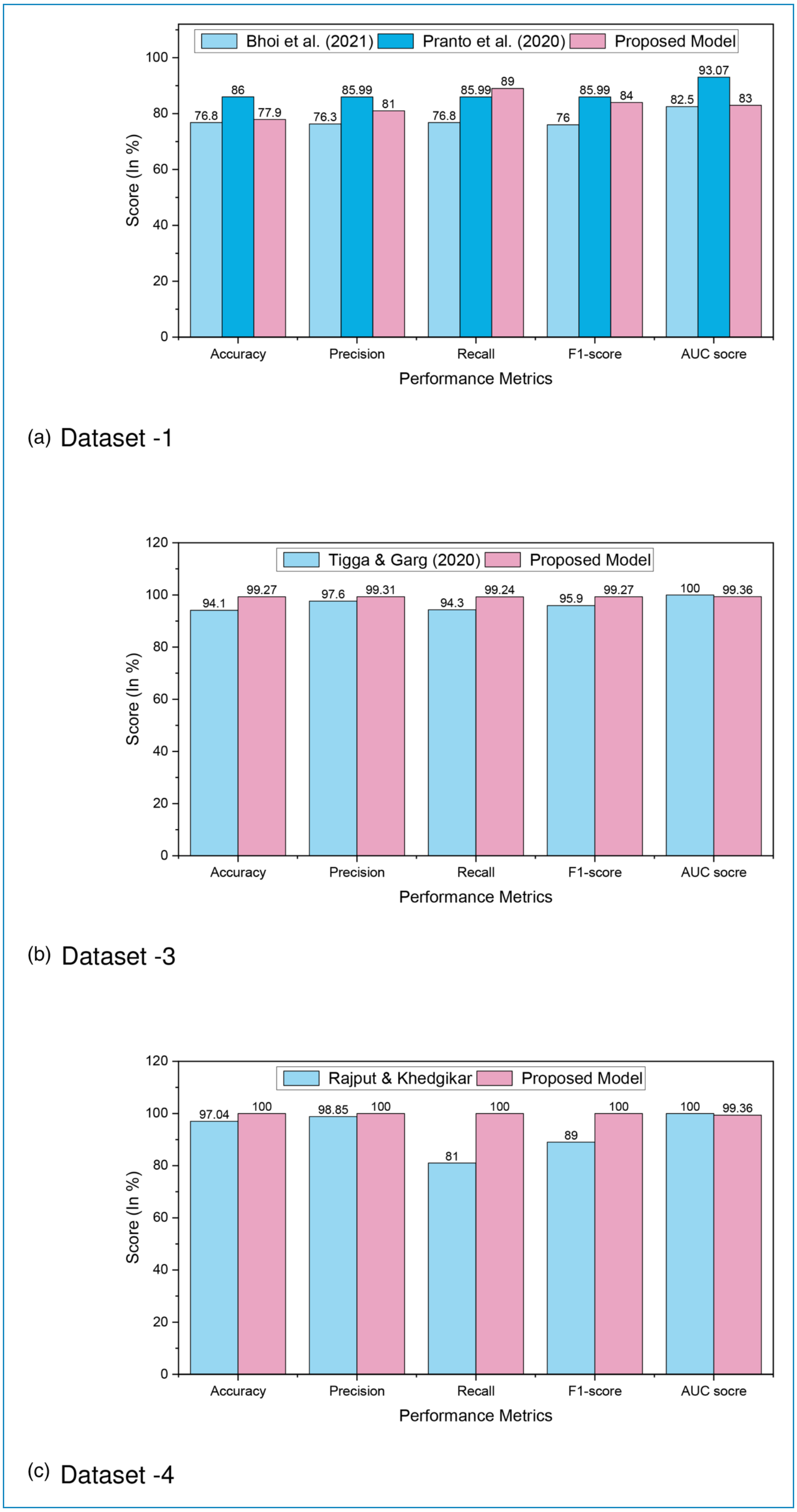
Table 8. Comparison analysis of diabetes prediction for Dataset 1, Dataset 3, and Dataset 4.

Sl. No	Author	Dataset	ML model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC score (%)
01	Bhoi et al.[47]	Dataset 1 (Pima Indian)	LR	76.8	76.3	76.8	76	82.5
02	Li et al.[48]	Dataset 1 (Pima Indian)	XGBoost	80.2	–	70.6	75	–
03	Tigga and Garg[49]	Dataset 1 (Pima Indian)	LR	75.32	78.18	86	–	–
04	Islam and Jahan[50]	Dataset 1 (Pima Indian)	LR	78.01	–	80		83.3
05	Joshi and Dhakal[51]	Dataset 1 (Pima Indian)	LR	78.26	–	–	–	–
06	Sneha and Gangil[52]	Dataset 1 (Pima Indian)	NB	82.3	–	–	–	–
07	Zou et al.[53]		RF	77.21	–	–	–	–
08	Rajni and Amandeep[54]	Dataset 1 (Pima Indian)	RB-Bayes	72.9	–	–	–	–
09	Pranto et al.[55]	Dataset 1 (Pima Indian)	RF	77.9	81	89	84	83
10	Proposed model	Dataset 1 (Pima Indian)	RF	86	85.99	85.99	85.99	93.07
11	Tigga and Garg[38]	Dataset 3 (Tigga)	RF	94.10	97.60	94.30	95.90	100
12	Proposed model	Dataset 3 (Tigga)	XGBoost	99.27	99.31	99.24	99.27	99.36
13	Hassan et al.[56]	Dataset 4 (Mendeley)	ID3	98.25	–	–	–	–
14	Nuankaew et al.[26]	Dataset 4 (Mendeley)	AWOD	98.95	98.88	–	–	–
15	Rajput and Khedgikar[57]	Dataset 4 (Mendeley)	SGB	97.04	98.85	81	89	–
16	Proposed model	Dataset 4 (Mendeley)	DT	100	100	100	100	100

ML: machine learning; AUC: area under curve; LR: logistic regression; XGBoost: extreme gradient boosting; NB: Naive Bayes; RF: random forest; AWOD: average weighted objective distance; SGB: stochastic gradient boosting; DT: decision tree.

The significance of this study lies in its contributions to the field of diabetes research and its potential impact on clinical practice.

1. **Improved diagnostic accuracy:** By developing an optimized data preprocessing pipeline and implementing advanced ML techniques, our study enhances the accuracy of diabetes prediction. This can assist healthcare professionals in making more precise and timely diagnoses, leading to better patient outcomes.
2. **Personalized diabetes management:** The robust models developed in this study have the potential to



(a) Dataset -1

(b) Dataset -3

(c) Dataset -4

Figure 14. Comparison analysis of diabetes prediction for (a) Dataset 1; (b) Dataset 3; and (c) Dataset 4.

enable personalized diabetes management. By accurately predicting diabetes prognosis, healthcare providers can tailor treatment plans and interventions to individual patients, optimizing their care and reducing the risk of complications.

3. **Early intervention and prevention:** Early detection of diabetes is crucial for effective intervention and prevention. Our study contributes to the development of predictive models that can identify individuals at high risk of developing diabetes. This enables early intervention strategies, such as lifestyle modifications and targeted preventive measures, to be implemented, reducing the burden of the disease.
4. **Advancement in ML techniques:** Through the exploration and evaluation of various ML algorithms and preprocessing techniques, our study contributes to the advancement of the field. The insights gained from this research can inform the development of more robust and interpretable ML models for diabetes prediction and prognosis.

Overall, this study holds significant implications for clinical practice, offering improved diagnostic accuracy, personalized management strategies, early intervention opportunities, and advancements in ML techniques. The findings have the potential to enhance diabetes care, contribute to preventive healthcare, and ultimately improve patient outcomes.

Conclusion

In this research, we have conducted a comprehensive study on diabetes detection using ML techniques, aiming to underscore both the scientific value added by our work and the applicability of our findings in clinical practice. Our contributions encompass the outcome of an optimized preprocessing pipeline, addressing dataset imbalance, preventing overfitting, and demonstrating superior performance through extensive experimentation. We have rigorously evaluated various ML models on four different datasets: Pima Indian, Austin Public, Tigga, and Mendeley. Our results showcase notable improvements in accuracy, precision, recall, and F1-score metrics compared to existing methods.

Specifically, the RF model achieved an accuracy of 85.53% on the Pima Indian dataset, with balanced precision and recall values. On the Austin Public dataset, the RF model excelled with an exceptional accuracy of 98.48%, along with high precision, recall, and F1-score values. The XGBoost model demonstrated outstanding performance on the Tigga dataset, achieving an accuracy of 99.27% with minimal errors in predictions. Notably, the DT model achieved perfect accuracy and precision on the Mendeley dataset, indicating flawless classification of diabetes instances. Our study reveals significant improvements over existing methods, with accuracy rates ranging from 86% to 100% across different datasets. Specifically, our suggested method outperforms previous works by 4.95% to 12.15% for Dataset 1, 0% to 8.74% for Dataset 2, 2.25% to 10.75% for Dataset 3, and 0.19% to 9.84% for Dataset 4.

However, it is essential to acknowledge the limitations of our study. Further investigation is needed to assess the generalizability of our approach to diverse datasets, feature selection, ensemble models and DL techniques. Additionally, the lack of interpretability in ML models poses a challenge in understanding the underlying factors driving predictions.

In conclusion, our study highlights the need for further research to address limitations and enhance the reliability and applicability of the proposed approach for diabetes detection using ML. Moving forward, potential avenues for future research include:

1. Feature selection: Exploring advanced feature selection techniques to improve the efficiency and accuracy of diabetes detection models.
2. Ensemble models: Investigating the integration of ensemble learning techniques to combine multiple models for enhanced predictive performance.
3. DL algorithms: Exploring the application of DL algorithms, such as convolutional neural networks and recurrent neural networks, to improve the prediction accuracy of diabetes detection models.

By pursuing these future directions, we aim to advance the field of diabetes detection using ML and contribute to improved healthcare outcomes.

Acknowledgement: The authors would like to extend their sincere appreciation to the Researchers Supporting Project Number (RSP2024R301), King Saud University, Riyadh, Saudi Arabia.

Availability of data and materials: The UCI ML Diabetes Dataset 1 is available on https://archive.ics.uci.edu/ml/datasets/diabetes; The Austin Public Health Diabetes Dataset 2 is available on https://data.austintexas.gov/Health-and-Community-Services/Austin-Public-Health-Diabetes-Self-Management-Educ/48iy-4sbg; The T2D Dataset 3 is available on https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019; The Mendeley Diabetes Dataset 4 is available on https://data.mendeley.com/datasets/wj9rwkp9c2/1.

Consent to participate: Not applicable.

Consent to Publish: Not applicable.

Contributorship: Md. Alamin Talukder contributed to Conceptualization, data curation, methodology, software, resource, visualization, formal analysis, and writing—original draft, review, and editing; Md. Manowarul Islam, Md Ashraf

Uddin, Arnisha Akther, and Mohammad Ali Moni contributed to methodology, visualization, validation, investigation, and writing—review and editing; Majdi Khalid and Mohsin Kazi contributed to supervision, formal analysis, visualization, investigation, and writing–review and editing.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics approval: Not applicable.

Funding: The authors received no financial support for the research, authorship and/or publication of this article.

Guarantor: Md. Alamin Talukder.

ORCID iDs: Md. Alamin Talukder https://orcid.org/0000-0002-3192-1000
Md Ashraf Uddin https://orcid.org/0000-0002-4316-4975

References

1. Al-Goblan AS, Al-Alfi MA and Khan MZ. Mechanism linking diabetes mellitus and obesity. *Diabet Metab Syndr Obesity: Targets Ther* 2014; 7: 587–591.
2. Scheen A. Pathophysiology of type 2 diabetes. *Acta Clin Belg* 2003; 58: 335–341.
3. Association AD et al. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2006; 29: S43.
4. Mathieu C and Badenhoop K. Vitamin D and type 1 diabetes mellitus: State of the art. *Trends Endocrinol Metab* 2005; 16: 261–266.
5. Ershow AG. Environmental influences on development of type 2 diabetes and obesity: Challenges in personalizing prevention and management. *J Diabetes Sci Technol* 2009; 3: 727–734.
6. Røder ME, Porte D Jr, Schwartz RS, et al. Disproportionately elevated proinsulin levels reflect the degree of impaired B cell secretory capacity in patients with noninsulin-dependent diabetes mellitus. *J Clin Endocrinol Metab* 1998; 83: 604–608.
7. Han E, Cho NH, Kim MK et al. Diabetes fact sheets in korea, 2018: An appraisal of current status. *Diabet Metab J* 2019; 43: 487–494.
8. Parim B, Sathibabu Uddandrao V and Saravanan G. Diabetic cardiomyopathy: Molecular mechanisms, detrimental effects of conventional treatment, and beneficial effects of natural therapy. *Heart Fail Rev* 2019; 24: 279–299.
9. Kim TM, Kim H, Jeong YJ et al. The differences in the incidence of diabetes mellitus and prediabetes according to the type of HMG-CoA reductase inhibitors prescribed in Korean patients. *Pharmacoepidemiol Drug Saf* 2017; 26: 1156–1163.
10. Fagherazzi G and Ravaud P. Digital diabetes: Perspectives for diabetes prevention, management and research. *Diabet Metab* 2019; 45: 322–329.
11. Alberti KGMM, Zimmet P and Shaw J. International diabetes federation: A consensus on type 2 diabetes prevention. *Diabet Med* 2007; 24: 451–463.
12. Talukder MA, Islam MM, Uddin MA, et al. An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. *Expert Syst Appl* 2023; 230: 120534.
13. Talukder MA, Layek MA, Kazi M, et al. Empowering COVID-19 detection: Optimizing performance through fine-tuned efficientNet deep learning architecture. *Comput Biol Med* 2023; 168: 107789.
14. Bhattacharjya A, Islam MM, Uddin MA et al. Exploring gene regulatory interaction networks and predicting therapeutic molecules for hypopharyngeal cancer and EGFR-mutated lung adenocarcinoma. *FEBS Open Bio* 2024; 14: 1166–1191.
15. Talukder MA, Islam MM, Uddin MA, et al. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert Syst Appl* 2022; 205: 117695.
16. Talukder MA, Islam MM, Uddin MA et al. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *J Big Data* 2024; 11: 1–44.
17. Hasan MK, Alam MA, Das D, et al. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access* 2020; 8: 76516–76531.
18. Maniruzzaman M, Rahman MJ, Ahammed B, et al. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inform Sci Syst* 2020; 8: 1–14.
19. Kopitar L, Kocbek P, Cilar L, et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* 2020; 10: 1–12.
20. Saeedi P, Petersohn I, Salpea P et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabet Res Clin Pract* 2019; 157: 107843.
21. Maniruzzaman M, Kumar N, Abedin MM et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Comput Meth Prog Biomed* 2017; 152: 23–34.
22. Ahmed N, Ahammed R, Islam MM et al. Machine learning based diabetes prediction and development of smart web application. *Int J Cogn Comput Eng* 2021; 2: 229–241.
23. Howlader KC, Satu M, Awal M et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. *Health Inform Sci Syst* 2022; 10: 1–13.
24. Deepajothi S, Juliana R, Aruna S, et al. Hereditary factor-based multi-featured algorithm for early diabetesDetection using machine learning. *Artif Intell Tech Wirel Commun Netw* 2022; 15: 235–253.
25. Rajagopal A, Jha S, Alagarsamy R, et al. A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures. *Math Comput Simul* 2022.
26. Nuankaew P, Chaising S and Temdee P. Average weighted objective distance-based method for type 2 diabetes prediction. *IEEE Access* 2021; 9: 137015–137028.
27. Wei H, Sun J, Shan W et al. Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus. *Sci Total Environ* 2022; 806: 150674.
28. Sivaranjani S, Ananya S, Aravinth J, et al. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: *2021 7th international conference on advanced computing and communication systems*

(ICACCS), vol. 1, Coimbatore, India, 19–20 March 2021, pp.141–146. IEEE.

29. Ramesh J, Aburukba R and Sagahyroon A. A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthc Technol Lett* 2021; 8: 45.
30. Ravaut M, Sadeghi H, Leung KK et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *npj Digit Med* 2021; 4: 1–12.
31. Naz H and Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J Diabet Metab Dis* 2020; 19: 391–403.
32. Hassan MM, Billah MAM, Rahman MM, et al. Early predictive analytics in healthcare for diabetes prediction using machine learning approach. In: *2021 12th international conference on computing communication and networking technologies (ICCCNT)*, Kharagpur, India, 6–8 July 2021, pp.1–5. IEEE.
33. Gupta H, Varshney H, Sharma TK, et al. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex Intell Syst* 2021; 8: 1–15.
34. Gupta D, Choudhury A, Gupta U, et al. Computational approach to clinical diagnosis of diabetes disease: A comparative study. *Multimed Tools Appl* 2021; 80: 1–26.
35. Majhi SK. How effective is the Moth-Flame optimization in diabetes data classification. In: Kalita J, Balas, V, Borah, S (eds) *Recent developments in machine learning and data analytics: IC3 2018*, vol. 740. Singapore: Springer, 2019, pp.79–87.
36. Repository UML. Diabetes data set, https://archive.ics.uci.edu/ml/datasets/diabetes (1998, accessed 20 March 2021).
37. Riquelme L. Austin public health diabetes self-management education participant demographics 2015–2017, https://data.austintexas.gov/Health-and-Community-Services/Austin-Public-Health-Diabetes-Self-Management-Educ/48iy-4sbg (2018, accessed 20 March 2021).
38. Tigga NP and Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Comput Sci* 2020; 167: 706–716.
39. Rashid A. Diabetes Dataset, https://data.mendeley.com/datasets/wj9rwkp9c2/1 (2020, accessed 20 March 2021).
40. Susilawati DS and Riana D. Optimization the naive Bayes classifier method to diagnose diabetes mellitus. *IAIC Trans Sustain Digit Innov (ITSDI)* 2019; 1: 78–86.
41. Sharmin S, Ahammad T, Talukder MA, et al. A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. *IEEE Access* 2023.
42. Talukder MA, Hasan KF, Islam MM et al. A dependable hybrid machine learning model for network intrusion detection. *J Inf Sec Appl* 2023; 72: 103405.
43. Talukder MA, Sharmin S, Uddin MA, et al. MLSTL-WSN: Machine learning-based intrusion detection using SMOTETomek in WSNs. *Int J Inf Sec* 2024; 23: 2139–2158.
44. Akhter A, Acharjee UK, Talukder MA, et al. A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Nat Lang Process J* 2023; 4: 100027.
45. Hang OY, Virgiyanti W and Rosaida R. Diabetes prediction using machine learning ensemble model. *J Adv Res Appl Sci Eng Technol* 2024; 37: 82–98.
46. Reza MS, Hafsha U, Amin R, et al. Improving SVM performance for type II diabetes prediction with an improved nonlinear kernel: Insights from the Pima dataset. *Comput Method Progr Biomed Update* 2023; 4: 100118.
47. Bhoi SK and et al. Prediction of diabetes in females of Pima Indian heritage: A complete supervised learning approach. *Turk J Comput Math Educ (TURCOMAT)* 2021; 12: 3074–3084.
48. Li M, Fu X and Li D. Diabetes prediction based on XGBoost algorithm. *IOP Conf Ser: Mater Sci Eng* 2020; 768: 072093.
49. Tigga NP and Garg S. Predicting type 2 diabetes using logistic regression. In: Nath V and Mandal JK (eds) *Proceedings of the fourth international conference on microelectronics, computing and communication systems*, vol. 673. Singapore: Springer, 2021, pp.491–500.
50. Islam MA and Jahan N. Prediction of onset diabetes using machine learning techniques. *Int J Comput Appl* 2017; 180: 7–11.
51. Joshi RD and Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int J Environ Res Public Health* 2021; 18: 7346.
52. Sneha N and Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* 2019; 6: 1–19.
53. Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Front Genet* 2018; 9: 515.
54. Rajni R and Amandeep A. RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset. *Int J Electr Comput Eng* 2019; 9: 4866.
55. Pranto B, Mehnaz S, Mahid EB et al. Evaluating machine learning methods for predicting diabetes among female patients in bangladesh. *Information* 2020; 11: 374.
56. Hassan S, Karbat AR and Towfik ZS. Propose hybrid KNN-ID3 for diabetes diagnosis system. *Int J Sci Eng Res* 2014; 5: 1087–1104.
57. Rajput MR and Khedgikar SS. Diabetes prediction and analysis using medical attributes: A machine learning approach. *J Xi'an Univ Archit Technol* 2022; 14: 98–103.