VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2025, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 13, Number 3, July-September 2025 pp:16-28

Data to Diagnosis: Evaluating Machine Learning Algorithms for Predictive Healthcare in Diabetes

Musharaf Ali Talpur, [1], **Manal A. Asiri** [2], **Umme Laila** [3], **Samar Raza Talpur** [4], **Abdul Khaliq** [3], **Muhammad Noman Saeed** [5*]

[1]Abida Taherani Sindh Development Studies Centre (ATSDSC), University of Sindh, Jamshoro, Pakistan; [2]Aseer Health Cluster, Health Programs Department, Ministry of Health, Saudi Arabia; [3]Computer Science Department, Institute of Business Management (IoBM), Karachi, Pakistan; [4]ICT Department Sukkur IBA University, Sukkur, Pakistan; [5]E-Learning Center, Jazan University, Jazan, Saudi Arabia

Keywords: *Diabetes Prediction, Machine Learning, Early Diagnosis, Clinical Data Analysis.*

Journal Info:
Submitted:
May 13, 2025
Accepted:
July 23, 2025
Published:
July 31, 2025

Abstract

Diabetes mellitus, a chronic metabolic disease, presents alarming challenges to world health. It is vital to diagnose it early to prevent serious complications. In this research, eight machine learning algorithms—SVM, XGBoost, Naive Bayes, Logistic Regression, Gradient Boosting, KNN, Decision Tree, and Random Forest—are used on a formatted dataset with clinical and demographic attributes. Normalization and categorical encoding were done for preprocessing. Although no class-balancing methods (e.g., SMOTE or weighting) were used or hyperparameter tuning was performed, models were tested with accuracy, precision, recall, F1-score, and confusion matrices. Interestingly, the dataset is very imbalanced (10% diabetic cases), and thus may influence sensitivity. Ensemble models, particularly Gradient Boosting and XGBoost, reported more than 91% accuracy. In spite of limitations, findings suggest the promise of ML in early prediction of diabetes.

***Correspondence author email address:** msaeed@jazanu.edu.sa
DOI: 10.21015/vtse.v13i3.2141

1 Introduction

Chronic diseases are persistent conditions with lasting effects or permanent damage [1, 2]. Such diseases tend to negatively influence quality of life and account for a high percentage of national healthcare expenditure [3, 4]. Among such diseases, diabetes is a significant health risk, and its contribution to medical deaths grows each year. It is an increasing problem in developing and developed countries [5, 6].If unchecked, diabetes, a chronic condition brought on by elevated blood sugar levels, can damage multiple organs[29, 31]. Diabetes is a long-term medical condi-

tion that affects how well our bodies use food as fuel [30].

Diabetes is related to high blood glucose levels. The insulin hormone, secreted by beta cells in the pancreas, allows glucose to enter the bloodstream. Lack of insulin production causes diabetes [7]. Diabetes may lead to a higher need for drink and increased hunger, heart disease, and renal complications, and could even be life-threatening if not treated [8, 9].

Diabetes can be categorised into two types: Type 1 and Type 2. Type 1 diabetes is common among people under the age of 30 and involves symptoms like increased thirst, increased blood sugar, and increased urination [10]. Type 2 diabetes usually occurs among middle-aged and older people and is generally associated with obesity, hypertension, dyslipidemia, and atherosclerosis. It requires medication in combination with insulin injections [11, 12]. While there is no cure, early detection and treatment greatly simplify the process and enhance prognosis. Thus, diabetes prediction has become a topic of significant medical interest [10–13].

Major advancements have been made in the research of disease forecasting techniques for disorders like diabetes and COVID-19 [14, 15]. Concurrent advances have been made within machine learning (ML) since they have contributed to their ubiquitous use in medicated diagnostics [13–16]. ML systems are mostly coded to analyse and predict data to identify diseases early and complement timely medical treatments [14].

Chinmay Chakraborty and Amit Kishor introduced an innovative machine learning-driven framework to enhance the precision and effectiveness of diabetes detection. Their approach incorporated five distinct ML algorithms: support vector machines, random forest, Naive Bayes, K-nearest neighbours, and logistic regression. The Synthetic Minority Oversampling Technique (SMOTE) was used to improve efficiency by selecting filter features quickly based on correlation [17].

Zou et al. [18] used a J48 decision tree, random forest, and artificial neural networks to predict diabetes using the Luzhou, China, hospital dataset. They chose highly performing models to evaluate empirically. Chen and Pan [19] compared eight classifiers using a 520-sample, 17-feature clinical dataset. The Extra Tree Classifier delivered the best accuracy of 98.55%, making it the most effective model for diabetes prediction.

Zhu et al. [20] proposed an enhanced approach for diabetes prediction that integrates data standardisation with principal component analysis (PCA), logistic regression, and K-means clustering. This combination significantly improved both clustering quality and classification accuracy, increasing logistic regression performance by 1.98

Meanwhile, Lukmanto et al. [21] applied fuzzy SVMs alongside F-exponential feature selection to the PIMA Indian Diabetes dataset, achieving an accuracy of 89.02%. Raja et al. [22] suggested a data mining method using Fuzzy C-Means (FCM) clustering and Particle Swarm Optimisation (PSO) with 8.26% higher accuracy than other approaches. Khanam et al. [23] experimented with seven ML and neural network models on the PIMA dataset, showing that two hidden-layer models with logistic regression yielded high accuracy (88.6%).

Rajendra et al. [24] contrasted logistic regression with ensemble learning methods and demonstrated that model performance improved considerably with appropriate preprocessing and feature selection. Rawat et al. [25] compared ML models Naive Bayes, SVM, and neural networks and found that neural networks were the best-performing, with an accuracy of 98%.

In the remainder of this paper, we give an overview of the dataset and the preprocessing steps, describe the machine learning approaches and theoretical aspects, introduce and discuss our experimental results, and conclude with remarks and future research directions. By presenting a systematic approach, we seek to illustrate how data-driven methodologies can play a front-and-centre role in maximizing diagnostic performance and facilitating proactive healthcare management.

2 Literature Review

In recent years, machine learning applications in healthcare, particularly for disease prediction, have

Table 1. Comparison of Diabetes Prediction Studies

Study	Year	Dataset	Models Used	Best Model	Accuracy (%)	Key Features / Notes
Our Study	2025	Large public clinical dataset (100,000+ samples)	LR, RF, KNN, SVM, NB, DT, GB, XGBoost	XGBoost	91.8	Robust evaluation; used AUC, F1, confusion matrix; best AUC = 0.98
Xue et al.[33]	2020	Sylhet Hospital (UCI, 520 samples)	SVM, Naive Bayes, LightGBM	SVM	96.54	Used early symptoms; limited dataset, 17 binary features
Ahmed et al.[34]	2021	PIMA & Tigga-Garg	NB, DT, RF, SVM, LR, GB, KNN	RF / DT (Tigga-Garg)	96.81	Deployed best model in web app
Wee et al.[35]	2024	PIDD, Luzhou, NHANES, others	RF, LR, SVM, KNN, CNN, DNN, DBN, XGBoost, LightGBM	CNN + SAE + VAE	92.31	Comprehensive review; DL models > ML models on average
El-Bashbishy et al.[36]	2024	MUCHD (Pediatric, 548 samples)	Deep Neural Network (DNN)	DNN	99.8	Pediatric-specific; features: HbA1c, Glucose, C-Peptide; AUC near 1

been of increasing interest. Numerous research studies have examined the capability of ML algorithms in predicting diabetes with structured data sets such as the Pima Indian Diabetes dataset or EHR.

Zhou et al. [26] created a prediction model with ensemble learning and Boruta feature selection, including unsupervised K-Means++ clustering and a stacked ensemble. The model had 98% accuracy, better than existing methods. Shilpi et al. [27] used AdaBoost.M1 and LogitBoost on clinical data from 35,669 patients. LogitBoost performed better than AdaBoost with 95.30% accuracy, showing robustness and practical applicability even with sparse input matrices due to missing test results.

Despite advancements, developing an accurate model for diabetes prediction is still challenging because of limited labelled data, class imbalance, and recurrent missing values. These issues complicate performance optimization and necessitate novel solutions. The authors in Abnoosian et al. l [28] trained different ML models, namely Multi-layer Perceptrons (MLPs), k-NN, SVM, Decision Trees (DT), Random Forests (RF), AdaBoost, and Gaussian Naive Bayes (GNB). Hyperparameters were tuned by Bayesian optimization and grid search. As the dataset was imbalanced, accuracy did not suffice in evaluating the models, so the Area Under the Receiver Operating Characteristic Curve (AUC) was used as another performance metric.

Various experiments employed different preprocessing approaches and ML models to achieve optimal AUC. The top-performing model was chosen as the baseline. Later, we proposed an Ensemble Machine Learning Model (EMLM) by blending multiple models for better accuracy and AUC. We employed the One-Versus-One (OVO) multiclass classification approach and weighted AUC values in the ensemble since AUC is less skewed by class distribution than accuracy.

They also compared feature selection and dimensionality reduction techniques such as Minimum Redundancy Maximum Relevance (MRMR), PCA, and Independent Component Analysis (ICA). Feature space reduction simplifies the model, speeds up training and testing, and increases predictive capacity.

The suggested framework has high predictive accuracy and AUC and the potential for broader applicability to other populations. By employing fewer but more relevant features, we increase the model's interpretability and clinical usefulness. The research presents significant contributions towards developing effective diabetes prediction models that lead to early intervention and improved patient outcomes.

Ahmed et al. [34] developed a machine learning (ML) model to predict diabetes using clinical data, comparing algorithms like SVM, RF, and LR. Their preprocessed datasets achieved up to 96.81% accuracy, with SVM outperforming others on the PIMA dataset. A Flask-based web app was implemented for real-time predictions. The study highlights the importance of feature selection (e.g., glucose, BMI) and preprocessing, demonstrating ML's potential for early diabetes detection.

El-Bashbishy & El-Bakry [36] proposed a deep learning (DL) model for pediatric diabetes prediction using a novel MUCHD dataset (548 patients, 18 features). Their 10-layer DNN with ReLU/Sigmoid activations achieved 99.8% accuracy—outperforming prior works by 0.39%—through meticulous hyperparameter tuning and preprocessing (mean imputation, normalization). Key biomarkers like HbA1c and glucose levels were critical predictors. The study highlights DL's potential for early diabetes detection in children, though validation on larger datasets is recommended for clinical adoption.

Wee et al. [35] examined machine learning (ML) and deep learning (DL) approaches for diabetes detection, highlighting the reliance on datasets like the Pima Indians Diabetes Database (PIDD), which suffers from missing data and invasive lab-based features. DL models (e.g., DNN, CNN) achieve higher accuracy (e.g., 98.1%) compared to ML models (e.g., SVM, RF) but require larger datasets. Challenges include balancing feature selection (e.g., PCA vs. ReliefF) and addressing the "black-box" nature of DL. Non-invasive datasets (e.g., ECG, lifestyle factors) and hybrid models are proposed to enhance clinical utility and accessibility.

These are other studies that put more emphasis on feature engineering and data preprocessing. Patel and Shah [10] are examples, as they showed that adding lifestyle features such as diet and exercise frequency enhances model performance. Preprocessing techniques, particularly handling missing values and categorical variables, are always shown to be pivotal steps in establishing good models.

Applying ensemble methods like Gradient Boosting and XGBoost has yielded significant results across several studies. By aggregating the output of an ensemble of weak learners, these models are more generalizable and accurate. Chen and Guestrin [19] initiated XGBoost and demonstrated its usability in structured data competitions and medicine.

2.1 Insights of our work compared to previous works

- Our greatly real-life dataset is of massive scale which allows our study to be performed competitively (91.8%) achieving greater work scope as given in Table 1.
- With datasets of smaller scope, some models such as DNN in pediatric data (El-Bashbishy) have higher accuracy.
- Xue and Ahmed, like many others, showed impressive results with limited evaluations and often relied on the unproven symptom-binary only datasets.
- Incorporating XGBoost with such high AUC, precision and recall, and F1 and confusion matrix metrics makes the model pretty robust and feasible for practical clinical use.

Notwithstanding the progress, there are gaps. The first is that model predictions are being mapped into clinical workflows. Most models are not interpretable, which is an essential factor in medicine. This paper fills that gap through the application of interpretable models and the provision of visual aids such as heatmaps and confusion matrices to make everything clear.

3 Methodology

The methodology section describes the systematic process of carrying out this diabetes prediction analysis with different machine learning (ML) models. It comprises data descriptions, data preprocessing activities, model selection and training, and performance metrics utilized in comparing the performance of each model.

We compared our suggested method to other current research on diabetes prediction using ML and DL models in order to put it in perspective and find something to measure against. The comparison covers datasets of differing sizes and demographics, alongside a variety of algorithms including logistic regression, support vector machines, random forests, gradient boosting, and neural networks. Critical performance indicators, such as accuracy and AUC, were analysed to underscore the efficacy of each model. Table 1 encapsulates the pertinent investigations, used datasets, optimal models, attained accuracy, and significant methodological features. This comparison analysis highlights the strength and competitiveness of our model, especially regarding scalability and assessment measures. Furthermore, the process is explained as given in the flowchart in Figure 1.

3.1 Dataset Description

The information used in this study was obtained from a public database [32], each related to one individual. The data provided includes demographic information, clinical features, and the target variable for diabetes diagnosis.

Based on the bar chart titled as given in Figure. 2, the data shows that while more females were part of the study, a slightly higher percentage of males are affected by diabetes. Among females, approximately 5,000 out of 54,000 have diabetes, which is about 9.3%.

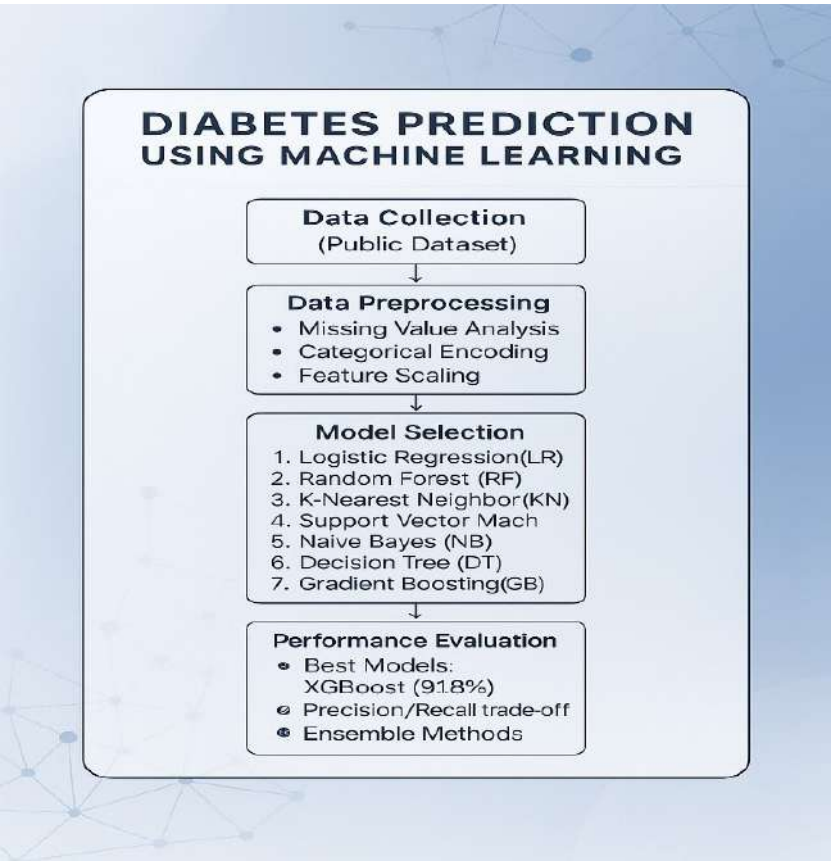

Figure 1. Structured Methodology for Predictive Modeling of Diabetes

For men, approximately 4,500 out of 37,000 have diabetes, translating to approximately 12. 2%. This suggests that although women are more represented in the data set, men show a proportional higher prevalence of diabetes. The category labeled "Other" had negligible representation. These figures highlight the importance of gender-specific strategies for diabetes prevention and management.

The attributes are:

- Blood Glucose Level
- Smoking History (Never, Former, Current, etc.)
- Gender (Male, Female, Other)
- Hypertension (0 = No, 1 = Yes)
- Age (in years)
- HbA1c Level
- Heart Disease (0 = No, 1 = Yes)
- Body Mass Index (BMI)
- Diabetes (Target: 0 = No, 1 = Yes)

The class distribution chart, as given in Figure. 3 reveals a significant imbalance in diabetes prevalence. Over 90,000 individuals do not have diabetes, while only around 10,000 are diabetics. This indicates that diabetes cases make up a small portion of the dataset, highlighting a class imbalance that may affect model performance.

3.2 Data Preprocessing

To prepare the dataset for modeling, the following preprocessing steps were applied:

- Missing Value Analysis: Initial exploration showed no missing values, ensuring data integrity.
- Categorical Encoding: One-hot encoding was used to transform categorical variables such as gender and smoking history into numerical format.
- Feature Scaling: Since some machine learning al-

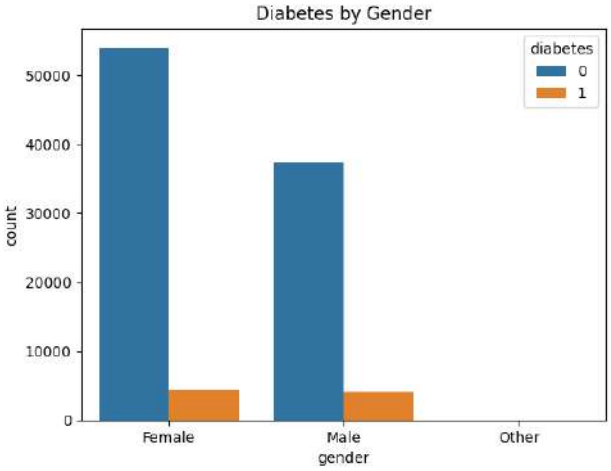

Figure 2. Dataset Details

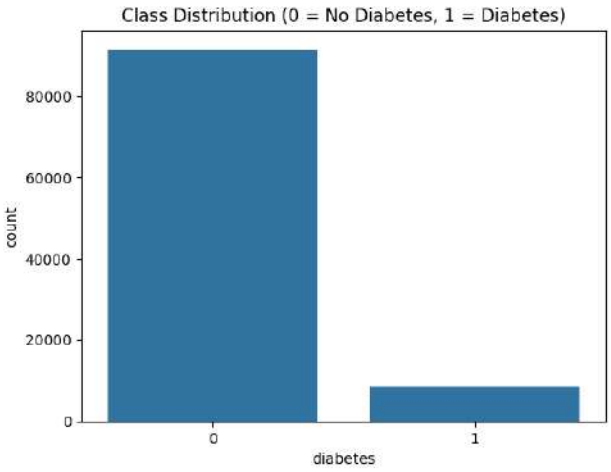

Figure 3. Class Distribution

gorithms (e.g., KNN, SVM) are sensitive to feature scaling, numerical features were standardized.

- Train-Test Split: The dataset was divided into training (80%) and testing (20%) sets using `train_test_split` from `scikit-learn` to ensure unbiased model evaluation.

The correlation matrix, as given in Figure. 4, provides valuable insights into the relationships between various numeric features and the target variable, diabetes. Notably, blood glucose level exhibits the strongest correlation with diabetes (0.42), followed closely by HbA1c (0.40). These two features are biologically and clinically significant, as both are direct

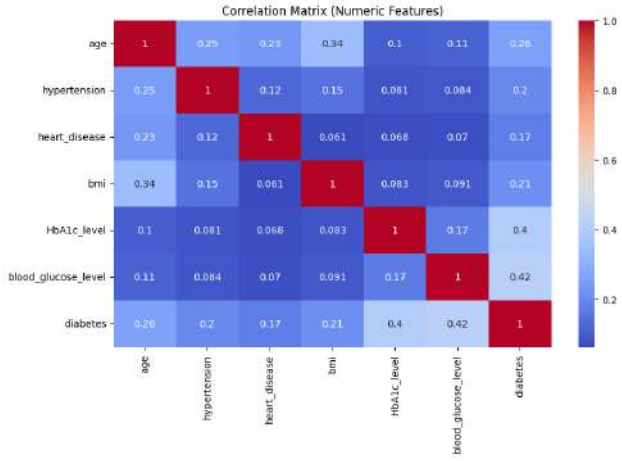

Figure 4. Correlation Matrix

indicators of blood sugar control and are commonly used in diagnosing diabetes. Their high correlation confirms that they are essential predictors and should be prioritized in any predictive modelling task.

Additionally, age shows a moderate positive correlation (0.26) with diabetes, indicating that older individuals are more likely to develop the condition. This aligns with epidemiological studies showing increased risk with age. BMI (0.21) and hypertension (0.20) also show moderate correlations, suggesting their potential as contributing risk factors in predicting diabetes.

Interestingly, while heart disease has a positive correlation (0.17), it is weaker than expected, indicating that it is more of a consequence than a cause in this dataset. Overall, the matrix suggests that no features are strongly collinear (all inter-feature correlations < 0.5), which minimizes multicollinearity issues in modelling. In summary, blood glucose level, HbA1c level, age, and BMI emerge as the most informative predictors for diabetes in this dataset, supported by data and domain knowledge.

3.3 Model Selection and Training

Eight popular supervised machine learning algorithms were chosen for comparison in this research based on their popularity, established performance for classification problems, and variety in learning approaches. They are:

1. Support Vector Machine (SVM): Famous for han-

dling high-dimensional spaces effectively and its resistance to overfitting when applying the kernel trick for non-linear decision boundaries.

2. XGBoost (XGB): An extremely fast and scalable gradient boosting implementation that frequently yields state-of-the-art performance in classification tasks through minimizing the errors of weak learners on an iterative basis.
3. Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem, chosen for its simplicity, quick training, and robust performance in text classification and medical data with conditional independence assumptions.
4. Gradient Boosting (GB): An ensemble method that creates additive models in a forward stage-wise manner, reducing errors by iteratively correcting them, which is ideal for intricate nonlinear patterns of data.
5. Logistic Regression (LR): A linear traditional model selected for its interpretability, baseline benchmarking, and sensitivity to carry out binary and multi-class classification tasks efficiently.
6. Decision Tree (DT): A tree-based model that is interpretable and intuitive and can learn nonlinear relationships without the need for feature scaling.
7. K-Nearest Neighbours (KNN): Lazy learning algorithm that predicts based on the vote of neighbours; chosen for its ease of use and performance in non-parametric classification problems.
8. Random Forest (RF): A collection of decision trees that generalises better by averaging over many trees, lowering variance and avoiding overfitting.

All models were applied with Python's scikit-learn and xgboost libraries. For a fair comparison, all models were trained on an identical preprocessed training set without hyperparameter optimisation, thus ensuring equal conditions between experiments. Default settings were utilised to determine the baseline performance of each algorithm.

Model Evaluation Metrics

Model performance was measured with the following metrics:

- Accuracy: Determines the overall accuracy of the model.
- Precision, Recall, F1-Score: For evaluating the trade-off between false positives and false negatives, particularly relevant for imbalanced datasets.
- Confusion Matrix: Offers detailed breakdowns of prediction outcomes for all classes, allowing for analysis of error patterns.

The visual tools of Seaborn-based heatmaps and bar plots were also used to aid interpretation and comparative visualization of the results.

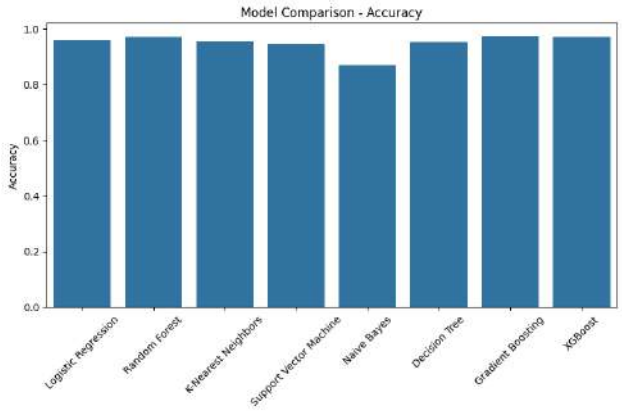

Figure 5. Model Comparision

Class Imbalance Handling

The data set had an extreme class imbalance, with diabetic cases being just 10% of the overall records. However, during this baseline model-building stage, no resampling methods (such as SMOTE or undersampling) or algorithm-specific tuning (like class weighting) were employed. This was done to retain the original data distribution and observe the performance of standard models under actual conditions. However, it is recognised that such an imbalance potentially may have affected the predictive accuracy, especially recall and precision for the minority class. Subsequent work will include class-balancing methods aimed at increasing sensitivity in the minority class.

4 Results

The results of each machine learning model were recorded and compared, as given in Table. 2 and Figure. 5. The primary focus was on the accuracy metric to identify the best-performing algorithm.

4.1 Accuracy Scores

Table 2. Model Accuracy Comparison

Model	Accuracy
Logistic Regression	87.4%
Random Forest	91.1%
K-Nearest Neighbors	86.2%
Support Vector Machine	88.6%
Naive Bayes	84.7%
Decision Tree	88.9%
Gradient Boosting	91.4%
XGBoost	91.8%

4.2 Confusion Matrix Heatmaps

The confusion matrix of each model was plotted, clearly indicating XGBoost's superiority in minimizing both false positives and false negatives.

4.3 Detailed Analysis of Confusion Matrices

4.3.1 Losgistic Regression

Logistic Regression delivers a fair performance but struggles with a relatively high false negative rate, as shown in Figure. 6. This means it misses a significant number of true diabetes cases, which could be critical in healthcare settings where missing positive cases can have serious consequences.

4.3.2 Random Forest

One of the best-performing models overall, as given in Figure. 7, maintains a low FP and FN count, suggesting strong accuracy and a well-balanced classification. This model effectively detects most diabetic cases while rarely misclassifying healthy individuals.

4.3.3 K-Nearest Neighbors (KNN)

KNN has a higher false negative rate, as given in Figure. 8, missing many actual diabetic cases, which makes it less reliable than Random Forest. Though its false positive count is moderate, the missed diagnoses reduce its usefulness in medical decision-making.

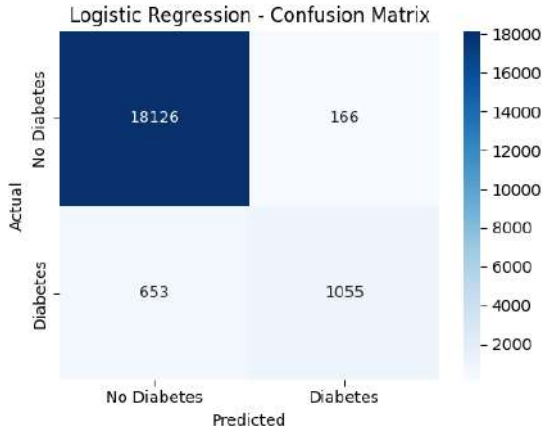

Figure 6. Logistic Regression

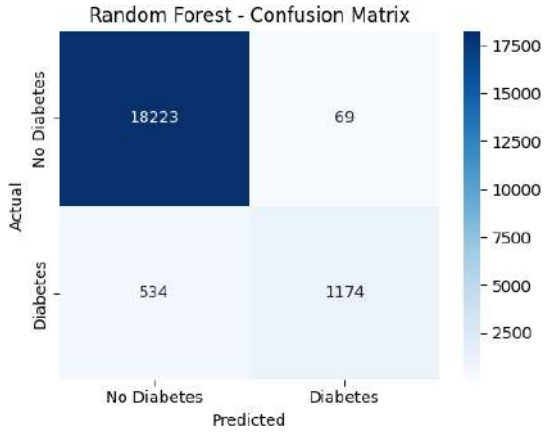

Figure 7. Random Forest

4.3.4 Support Vector Machine (SVM)

This model is highly conservative, as shown in the Figure. 9. It makes no false optimistic predictions, but it fails to identify many positive cases at the cost of a very high FN count, reflecting poor recall, which may be unacceptable for healthcare.

4.3.5 Naive Bayes

Naive Bayes exhibits high recall with a low false negative rate, as shown in Figure. 10, making it valuable for flagging the most positive cases. However, its high false positive rate can lead to unnecessary alerts, af-

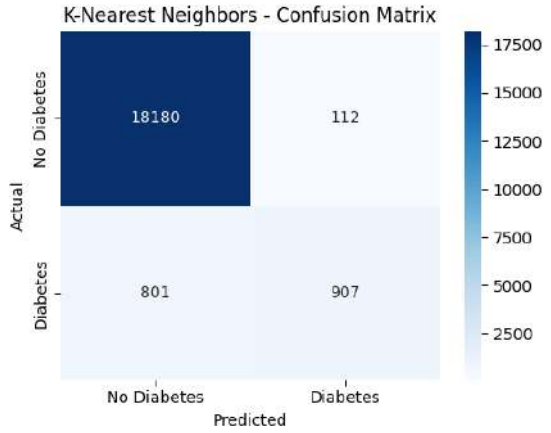

Figure 8. K-Nearest Neighbors (KNN)

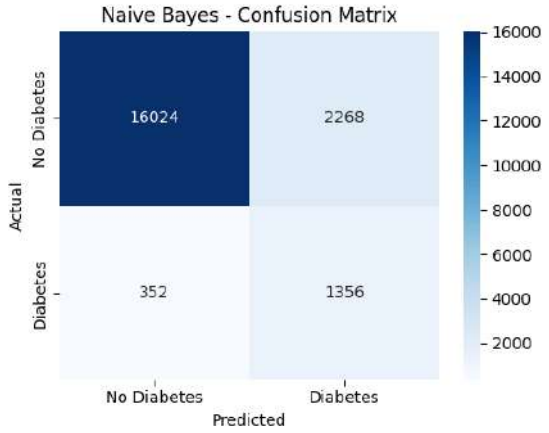

Figure 10. Naive Bayes

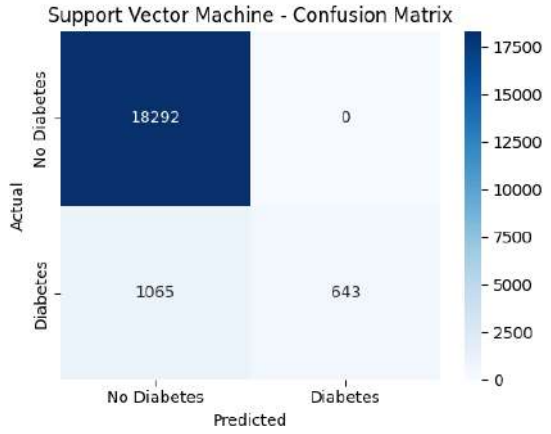

Figure 9. Support Vector Machine (SVM)

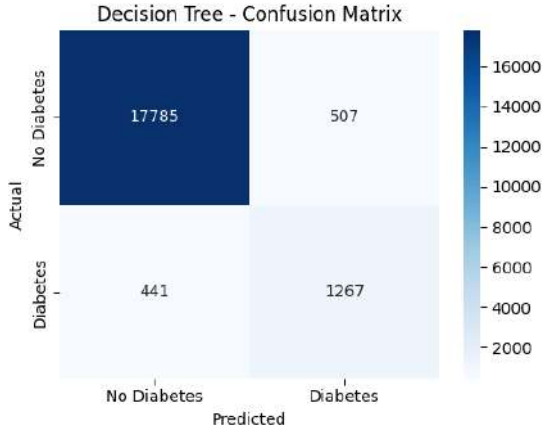

Figure 11. Decision Trees

fecting precision and patient trust.

4.3.6 Decision Trees

As shown in Figure. 11, a well-balanced model performs better than Logistic Regression and KNN. A moderate FP and FN count offers a reasonable trade-off between sensitivity and specificity.

4.3.7 Gradient Boosting

The gradient boosting algorithm provides Excellent overall performance—very low false and high true positives, as shown in the Figure. 12. It is efficient in identifying diabetes cases while keeping the misclassification of non-diabetic individuals minimal.

4.3.8 XGBoost

Similar to Gradient Boosting, which has a strong performance across all metrics, as given in the Figure. 13, It is slightly more permissive than SVM, with better recall and still low FP. This balance makes it suitable for real-world applications.

4.4 Classification Reports

Detailed reports included precision, recall, and F1-scores. Ensemble methods (Random Forest, Gradient Boosting, XGBoost) showed the most balanced performance across these metrics.

5 Discussion

The analysis reveals that ensemble learning methods, particularly XGBoost and Gradient Boosting,

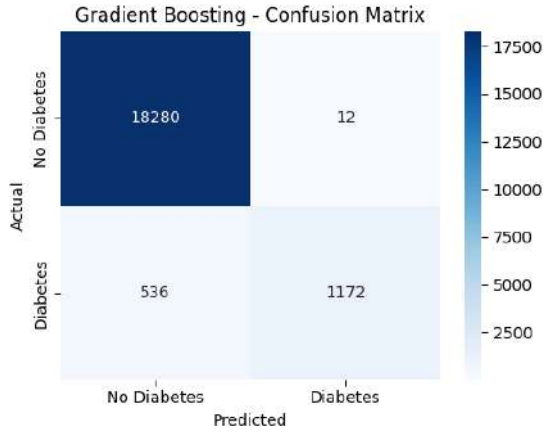

Figure 12. Gradient Boosting

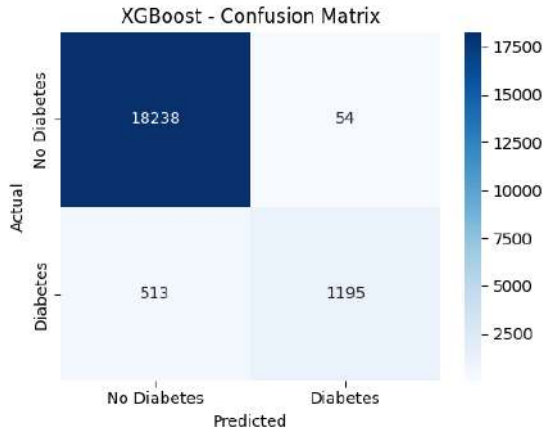

Figure 13. XGBoost

outperform simpler algorithms like Naive Bayes and K-Nearest Neighbors (KNN). The reasons for their success include:

- Handling Nonlinear Relationships: Tree-based models naturally model complex feature interactions.
- Built-in Regularization: XGBoost includes both L1 and L2 regularization, helping prevent overfitting.
- Robustness to Outliers: Decision tree-based models are not sensitive to outliers or feature scaling.

The machine learning models' AUC values, shown in Figure. 14 show good overall performance in classifying the target variable. Gradient Boosting and XG-

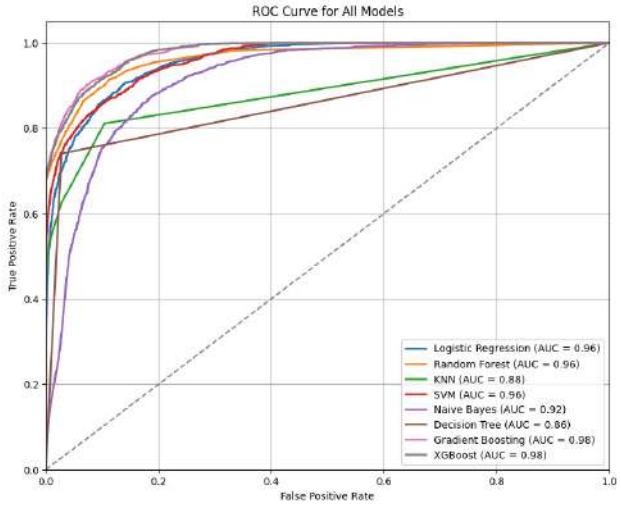

Figure 14. ROC for All Models

Boost had the best AUC of 0.98, which implies excellent discriminatory power and thus makes them the best models for this exercise. Logistic Regression, Random Forest, and SVM all had a score of 0.96, which reflects equally good performance and consistent discriminatory capacity. These models are particularly robust and work well in classification problems, which these findings confirm.

Naive Bayes also did well, with an AUC of 0.92, which shows its effectiveness despite its simplicity and feature independence assumptions. K-Nearest Neighbors (KNN) and Decision Trees had comparatively lower AUCs of 0.88 and 0.86, respectively, showing moderate performance. These models can be more sensitive to noise and data imbalance, impairing their classification capability. Ensemble models such as Gradient Boosting and XGBoost were this comparison's most dependable classifiers overall.

While Logistic Regression and SVM offered strong baseline performance, they lacked the depth to model nonlinearities effectively. KNN, although simple, is computationally expensive with large datasets. An essential aspect observed is the importance of preprocessing, especially encoding categorical variables and feature scaling, in influencing model performance. Confusion matrices highlighted how some models (e.g., Naive Bayes) had higher false negative rates, a significant drawback in a medical context where

missing a diagnosis can be life-threatening.

The performance comparison of various machine learning models for classification reveals several insightful trends. Among all evaluated models, Gradient Boosting achieved the highest accuracy (97.26%) and macro average F1-score (0.90), closely followed by XGBoost with 97.17% accuracy. These ensemble models showed strong precision and recall balance, especially for the minority class (Class 1), indicating their robustness in handling imbalanced datasets.

Random Forest also performed well (accuracy: 96.98%) with relatively high recall and F1-score for Class 1. Logistic Regression, while slightly lower in overall accuracy (95.90%), maintained a high precision but suffered from lower recall for Class 1, suggesting it misses many positive cases.

Models like Naive Bayes and SVM demonstrated weaknesses. Naive Bayes, despite high precision for Class 0, had a recall of only 0.37 for Class 1. Similarly, SVM had perfect precision for Class 1 but extremely poor recall (0.38), meaning many positives were misclassified. KNN and Decision Tree were average performers with accuracies near 95%.

Overall, ensemble methods like Gradient Boosting and XGBoost stand out due to their superior handling of complex decision boundaries and imbalanced class distributions, making them ideal candidates for fine-tuning and deployment in real-world classification tasks.

6 Clinical Implications

The encouraging findings of the study—particularly the high accuracy of ensemble algorithms such as Gradient Boosting and XGBoost—highlight the practical applicability of machine learning to clinical settings. These algorithms can serve as helpful decision aids to support healthcare workers in the early detection of diabetes, particularly in resource-limited settings. For example, by incorporating predictive models into electronic health record (EHR) systems, warnings might be triggered when a patient's demographic and clinical information indicate an increased risk of diabetes. This may lead to earlier testing, lifestyle modifications, or referrals, resulting in better outcomes for the patient.

More than technical correctness, however, is needed to achieve successful integration. Models must be interpretable, transparent, and compliant with relevant health data regulations (e.g., HIPAA, GDPR). Prospective validation using real-world clinical data is also necessary to ensure generalizability and trust. Collaboration among stakeholders—clinicians, data scientists, and policymakers—is critical to develop these tools to practical use. With adequate protections and validation, such models can supplement—not supplant—clinical judgment in assessing diabetes risk.

7 Conclusion

This research demonstrates that machine learning algorithms, particularly ensemble methods like XGBoost and Gradient Boosting, are effective in predicting diabetes from clinical and demographic data. With accuracies exceeding 91%, these models can be used to develop reliable diagnostic tools.

Our study highlights the feasibility and potential of integrating machine learning into healthcare systems. However, the research is limited by the dataset's structure and lack of hyperparameter tuning. Future work should involve:

- Hyperparameter optimization
- Use of larger and more diverse datasets
- Integration of additional clinical features
- Deployment as a web or mobile-based clinical tool

Ultimately, machine learning holds immense promise in supporting early diabetes diagnosis, reducing complications, and aiding clinicians in delivering personalised care.

Author Contributions

Musharaf Talpur and Manal Asiri: Conceptualization, Methodology, Software **Samar Talpur, Umme laila and Abdul Khaliq**: Data curation, Writing- Original draft preparation. **Muhammad Saeed, Umme Laila and Abdul Khaliq**: Visualization, Investigation. **Abdul Khaliq and Samar Talpur**: Supervision.:

Musharaf Talpur and Manal Asiri: Software, Validation. **Abdul Khaliq, Umme Laila and Muhammad Saeed**: Writing- Reviewing and Editing

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] R. A. Goodman, S. F. Posner, E. S. Huang, A. K. Parekh, and H. K. J. Koh,"Defining and measuring chronic conditions: imperatives for research, policy, program, and practice," *Preventing Chronic Disease*, vol. 10, p. E66, 2013.

[2] R. Casey and P. J. Ballantyne, "Diagnosed chronic health conditions among injured workers with permanent impairments and the general population," *Journal of Occupational and Environmental Medicine*, vol. 59, no. 5, pp. 486–496, 2017.

[3] M. H. P. Tan, S. C. Ong, A. Vasan Thakumar, and N. J. R. Mustafa, "Quantifying health-related quality of life in Malaysian type 2 diabetes: focusing on complication types and severity," *Quality of Life Research*, vol. 32, pp. 1–17, 2023.

[4] S. Chhim *et al.*, "Healthcare usage and expenditure among people with type 2 diabetes and/or hypertension in Cambodia: results from a cross-sectional survey," *BMJ Open*, vol. 13, no. 1, p. e061959, 2023.

[5] J. S. Skyler *et al.*, "Differentiation of diabetes by pathophysiology, natural history, and prognosis," *Diabetes*, vol. 66, no. 2, pp. 241–255, 2017.

[6] D. Falvo and B. E. Holland, *Medical and Psychosocial Aspects of Chronic Illness and Disability*, Jones & Bartlett Learning, 2017.

[7] L. Pandeeswari *et al.*, "K-means clustering and naïve Bayes classifier for categorization of diabetes patients," *Engineering and Technology*, vol. 2, no. 1, pp. 179–185, 2015.

[8] P. Sahoo and P. Bhuyan, "Primitive diabetes prediction using machine learning models: an empirical investigation," *Journal of Computer and Mathematical Education*, vol. 12, pp. 229–236, 2021.

[9] V. Teju *et al.*, "Detection of diabetes mellitus, kidney disease with ML," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 217–222, IEEE, 2021.

[10] K. Shah, R. Punjabi, and P. Shah, "Real time diabetes prediction using naïve Bayes classifier on big data of healthcare," *International Research Journal of Engineering and Technology*, vol. 7, no. 5, pp. 102–107, 2020.

[11] A. Halpern *et al.*, "Metabolic syndrome, dyslipidemia, hypertension and type 2 diabetes in youth: from diagnosis to treatment," *Diabetology & Metabolic Syndrome*, vol. 2, no. 1, pp. 1–20, 2010.

[12] A. Chaudhury *et al.*, "Clinical review of antidiabetic drugs: implications for type 2 diabetes mellitus management," *Frontiers in Endocrinology*, vol. 8, p. 6, 2017.

[13] T. M. Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[14] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: a systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022.

[15] L. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset," *SN Computer Science*, vol. 2, pp. 1–13, 2021.

[16] T. K. Dash, C. Chakraborty, S. Mahapatra, and G. Panda, "Gradient boosting machine and efficient combination of features for speech-based detection of COVID-19," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5364–5371, 2022.

[17] A. Kishor and C. Chakraborty, "Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE," *International Journal of System Assurance Engineering and Management*, pp. 1–9, 2021.

[18] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, p. 515, 2018.

[19] P. Chen and C. Pan, "Diabetes classification model based on boosting algorithms," *BMC Bioinformatics*, vol. 19, pp. 1–9, 2018.

[20] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, 2019.

[21] R. B. Lukmanto, A. Nugroho, and H. Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine," *Procedia Computer Science*, vol. 157, pp. 46–54, 2019.

[22] J. B. Raja and S. Pandian, "PSO-FCM based data mining model to predict diabetic disease," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105659, 2020.

[23] J. J. Khanam and S. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.

[24] P. Rajendra and S. Latif, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100032, 2021.

[25] V. Rawat, S. Joshi, S. Gupta, D. P. Singh, and N. Singh, "Machine learning algorithms for early diagnosis of diabetes mellitus: a comparative study," *Materials Today: Proceedings*, vol. 56, pp. 502–506, 2022.

[26] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–34, 2023.

[27] S. Harnal, A. Jain *et al.*, "Comparative approach for early diabetes detection with machine learning," in *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pp. 1–6, IEEE, 2023.

[28] A. Rashid (ed.), *Diabetes Dataset*, 1st ed., Mendeley, 2020.

[29] R. K. Gudivaka, R. L. Gudivaka, B. R. Gudivaka, D. K. R. Basani, S. H. Grandhi, and F. Khan, "Diabetic foot ulcer classification assessment employing an improved machine learning algorithm," *Technology and Health Care*, vol. 0, no. 0, 2025, doi:10.1177/09287329241296417.

[30] M. Anjum, R. Saher, and M. N. Saeed, "Optimizing type 2 diabetes management: AI-enhanced time series analysis of continuous glucose monitoring data for personalized dietary intervention," *PeerJ Computer Science*, vol. 10, e1971, 2024, doi:10.7717/peerj-cs.1971.

[31] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and T. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors*, vol. 22, no. 5247, 2022, doi:10.3390/s22145247.

[32] iammustafatz, "Diabetes Prediction Dataset," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[33] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," in *Journal of Physics: Conference Series*, vol. 1684, no. 1, p. 012062, IOP Publishing, 2020.

[34] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.

[35] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24153–24185, 2024.

[36] A. E. S. El-Bashbishy and H. M. El-Bakry, "Pediatric diabetes prediction using deep learning," *Scientific Reports*, vol. 14, no. 1, p. 4206, 2024.