Vol. 5 No. 2 (2026)

Pakistan Journal of Medical & Cardiological Review

https://pakjmcr.com/index.php/1/about

Online ISSN 3007-2387

Print ISSN 3007-2379

Machine Learning Based Early Prediction of Diabetes Mellitus Using Clinical Data: A Comparative Study

Halima Sadia
Faculty Of Medicine,Universite Laval, Email: halima-sadia.sadia.1@ulaval.ca

Dr. Shahzadi Saba
Family Medicine Specialist, NMC Health Care UAE, Oxford Medical Centre, Abudahbi Email: dr_saba75@yahoo.com

Tony T. Williams
MHA, Ed.S., PhD (H.C.), Department of Health and Human Services, Ashford University- UAGC Email: williamscaddola@gmail.com

Aishath Mala
Faculty of Health Sciences, Villa College Male Maldives, Email: aishath.mala@villacollege.edu.mv

Azzah Khadim Hussain
MPhil Pharmaceutics, University of Central Punjab Email: azzah.khadim@gmail.com

Naima Ibrahim Joo
Department of Computer science (Artificial Intelligence), MSCS, National University of Technology, Islamabad Email: naimaibrahimjoo@gmail.com

Author Details

Keywords: Diabetes Mellitus; Machine Learning; Early Prediction; Comparative Study; Gradient Boosting; Random Forest; Clinical Data; SMOTE; Feature Importance; AUC-ROC

Received on 15 May 2026
Accepted on 20 Jun 2026
Published on 30 Jun 2026

Corresponding E-mail & Author*:
Halima Sadia
Email: halima-sadia.sadia.1@ulaval.ca

Abstract

Background: Diabetes mellitus is one of the major public health problems of the twenty-first century, affecting more than 537 million adults worldwide, and projected to affect more than 780 million by 2045. Early and correct diagnosis of type 2 diabetes is essential for preventive intervention, patient stratification and optimization of healthcare resource. The challenges of creating predictive models from heterogeneous clinical data can be addressed by machine learning (ML).

Objective: The objective of this study is to systematically compare and evaluate 7 popular Machine Learning Classifiers namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Gradient Boosting (GB), and Artificial Neural Network (ANN) for early prediction of Diabetes Mellitus with Structured Clinical/Demographic Data.

Methods: Pima Indians Diabetes Dataset (PIDD) and secondary data from a hospital with a patient count of 1523 were used. Preprocessing techniques were applied to deal with missing values, outlier detection, feature scaling, and class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). The measures used to evaluate models included

accuracy, sensitivity, specificity, precision, F1 score, Matthew's correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC-ROC). Friedman and Wilcoxon signed-ranked tests were used to determine statistical significance.
Results: Gradient Boosting had the best classification performance in terms of accuracy (91.4%), AUC-ROC (0.958), F1 score (0.913) and MCC (0.819). Random Forest was the second best with an accuracy of 89.7% and AUC-ROC of 0.943. Logistic Regression was the most interpretable model and performed well with 79.6% accuracy. The plasma glucose concentration, body mass index (BMI), age and diabetes pedigree function were consistently identified as the most predictive variables by feature importance analysis.
Conclusions: Ensemble methods, especially Gradient Boosting and Random Forest algorithms, outperform the others in predictive performance for early diagnosis of diabetes and have a great clinical potential. The results highlight the significance of strict preprocessing and correction of class imbalance. Future studies should explore the possibility of federated learning approaches and real-time integration of clinical decision support.

Abbreviations:
ML, Machine Learning; DM, Diabetes Mellitus; RF, Random Forest; GB, Gradient Boosting; SVM, Support Vector Machine; k-NN, k-Nearest Neighbors; ANN, Artificial Neural Network; LR, Logistic Regression; DT, Decision Tree; AUC-ROC, Area Under the Receiver Operating Characteristic Curve; SMOTE, Synthetic Minority Over-sampling Technique; MCC, Matthew's Correlation Coefficient; PIDD, Pima Indians Diabetes Dataset; BMI, Body Mass Index

Introduction

Diabetes mellitus (DM) is a chronic metabolic disease in which the blood glucose is persistently elevated, due to defects in insulin secretion, insulin action or both. It is estimated that in 2025 there are some 537 million adults aged between 20 and 79 living with diabetes, which is 10.5% of the adult population worldwide [1]. Type 2 diabetes mellitus (T2DM) is the most prevalent form of diabetes, with about 90-95% of all cases diagnosed being of this type, and is known to be closely related to modifiable risk factors such as obesity, physical inactivity, unhealthy diet and genetic susceptibility [2].

The long-term complications of T2DM are micro and macrovascular, such as retinopathy, nephropathy, neuropathy, coronary artery disease, stroke and other complications, and are caused by the slow progression of T2DM which can be undiagnosed for years. These complications have significant morbidity and mortality risk and economic cost – the worldwide cost of diabetes care is estimated at USD 966 billion in 2025 [1]. Importantly, for those who are at high risk, early detection and lifestyle changes have been shown to significantly delay or prevent development of T2DM, and accurate early prediction is a public health priority [3].

Existing clinical screening methods include fasting plasma glucose testing and oral glucose tolerance testing, which are labor-intensive and require access to healthcare facilities, making them difficult to scale up at population level. While algorithmic risk scores, like the Finnish Diabetes Risk Score (FINDRISC), are simpler, they generally use a restricted number of variables and often do not have sufficient granularity in prediction for precision medicine [4]. The introduction of electronic health records (EHRs) and large-scale clinical databases has created enormous amounts of patient data, which can be used on a large scale to develop diabetes risk prediction using data-driven machine learning techniques [5].

The use of machine learning methods in clinical decision support is of great interest in research. Additionally, ML models are able to identify non-linear relationships, or

simply a more complex one, between the predictive features of the model and both can be applied to high-dimensional data and may incorporate different types of data (genomics, lifestyle, clinical biomarkers, etc.). Data mining has been employed in diabetes prediction by several ML techniques with varying success, such as decision trees, random forests, support vector machines and neural networks [6,7]. There are, however, variations with regard to the datasets used, the evaluation measures adopted, the pre-processing procedures and the experimental setup, which makes it difficult to draw any conclusions.

In this study, one of these gaps is filled and a methodologically homogeneous and rigorous comparative analysis of seven ML classifiers applied to the prediction of T2DM is carried out. We make four contributions: (1) a comprehensive pipeline for missing values, outlier detection, feature scaling, and class imbalance; (2) a systematic comparison of seven ML models, evaluated under the same conditions; (3) a thorough performance evaluation by means of a wide range of metrics, such as accuracy, AUC-ROC, F1-score, and MCC; and (4) feature importance analysis to pinpoint the most clinically relevant predictors. The rest of this paper will be organized as follows: Section 2 reviews the relevant prior work; Section 3 describes the datasets and methodology; Section 4 presents the results; Section 5 discusses the findings and their clinical implications; Section 6 outlines future directions and limitations; and Section 7 concludes the paper.

Literature Review

In the last twenty years, the field of diabetes prediction has received wide attention in the use of machine learning. In this section the important previous studies are reviewed and thematized according to the methodological approach.

Traditional Machine Learning Approaches

Kavakiotis et al. [8] performed a systematic review of studies for the application of ML and data mining techniques to diabetes research on 85 studies, which revealed that the classifiers used most were the support vector machine and neural network. The accuracy reported in different studies varied from 70% to 96%, with different factors reported in each study as the cause of the different accuracy. These factors included different characteristics of the datasets, different preprocessing and different evaluation protocols. The Pima Indians Diabetes Dataset turned out to be the most popular benchmark, used in more than 60% of the studies that were reviewed.

Sisodia and Sisodia [9] have tested 3 classifiers: Naive bayes, Decision Tree, Support Vector Machine on the PIDD. The highest accuracy was for Naive Bayes which scored 76.3%, followed by Decision Tree with 73.0% and SVM with 65.1%. As expected, the authors found BMI, glucose level and age to be the most predictive features. No imbalancing of the class was performed, though, which could introduce biases in favor of the larger class.

Zou et al. [10] carried out a comparison with seven ML algorithms for predicting T2DM in a population of 14,592 patients in China. The other algorithms had lower AUC values, with the highest being 0.814.Other algorithms performed worse than Random Forest with AUC value of 0.871. More importantly, the authors determined that the addition of lifestyle factors—including physical activity and nutrition to the clinical biomarkers significantly enhanced model accuracy, demonstrating the significance of incorporating multiple modalities.

Ensemble Methods and Boosting

In binary classification problems with medical data, ensemble methods have been shown to be effective and effective. Ensemble learning combines results of different base learners to minimize the variance, bias, or both [11]. Friedman's Gradient Boosting Machine [12] makes a series of approximations to residuals of previous

approximations, generating very accurate models. The diabetes prediction task was tackled by several groups and optimized implementations of gradient boosting, such as XGBoost, were used to achieve AUC values over 0.90 in several clinical datasets by Chen and Guestrin [13]..

Nnamoko and Korkontzelos [14] applied an ensemble of RF and AdaBoost to a dataset on an NHS primary care database with an 88.6% sensitivity and 85.3% specificity. A key finding of the study was that assortment diversity, in particular the utilization of algorithmically different base classifiers, was essential for performance improvements over individual classifiers. One limitation was the lack of external validation cohorts.

Deep Learning Approaches

Longitudinal EHR data has been used for diabetes prediction using recurrent neural networks (RNNs) and long short term memory (LSTM) networks [15]. In settings with large amounts of data, Miotto et al. [16] have shown that the end-to-end representation learning methods based on deep learning models performed better in predicting risk than the models based on manually curated features. But deep learning models require more data and are not as interpretable as traditional ML which hinders their clinical translatability.

Feature Selection and Engineering

Feature selection is an essential and important pre-processing step that can impact the model's performance and interpretability. Several methods have been used to predict diabetes, including mutual information, recursive feature elimination (RFE) and LASSO regularization. Maniruzzaman et al. [17] showed that their feature subset of glucose, BMI, age, pedigree function, and blood pressure performed almost as well as the entire set of feature models, while being much more efficient to train. Feature selection follows principles of clinical parsimony, that is, those models that are accurate and actionable.

Handling Class Imbalance

In medical data, there is widespread imbalance in the number of positive (diabetic) patients compared to the number of negative (nondiabetic) patients, with only 30-40% of the population comprising the positive cases. Clinically unsafe, naive models trained on imbalanced data usually have high specificity and poor sensitivity, the latter is associated with higher costs of false negatives. The most popular re-balancing methods used in the ML health informatics literature are synthetic oversampling methods, such as SMOTE [18] which creates new instances of the minority class along interpolated feature vectors.

Materials and Methods

Datasets

Two datasets were used in the present study, for benchmarking and for generalizability evaluation.

Pima Indians Diabetes Dataset (PIDD)

The PIDD was curated in the beginning by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is publicly available in the UCI Machine Learning Repository [19]. It consists of 768 female Pima Indian patients, age 21 years or older. The clinical and demographic features are: number of pregnancies, plasma glucose concentration (2-hour OGTT), diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), BMI (kg/m2), diabetes pedigree function (hereditary risk score), and age (years). The dichotomous

dependent variable was the 5-year status of diabetes (268 cases (34.9%) and 500 cases (65.1%)).

Secondary Clinical Dataset

The secondary data was collected from a tertiary care hospital, in collaboration with the institutional review board (IRB-2024-ML-047), and included 1,523 patient records from 2019 to 2024. It includes 12 features: fasting blood glucose, HbA1c, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, BMI, waist circumference, physical activity index, and family history of diabetes (binary). The dependent variable is the diagnosis of T2DM (positive: 612 cases, 40.2%). Prior to analysis all patient data were de-identified in accordance with HIPAA and GDPR regulations.

Data Preprocessing

Both datasets were preprocessed using rigorous procedures that were consistently applied to both to guarantee data quality and model reliability.

Missing Value Imputation

Zero value in the glucose, blood pressure, skin thickness, insulin, and BMI columns were considered de facto missing values in the PIDD since they cannot be zero in living patients. There were 3.2% of the values of blood pressure, 4.6% of BMI, 29.7% of skin thickness, and 48.7% of insulin that were such values. Missing data was imputed with the K-Nearest Neighbors (KNN) imputation method, k=5 and stratified by outcome class to preserve conditional distributions (pooled) of each variable. Multivariate imputation by chained equations (MICE) was performed for the imputation of missing data in the secondary data set, which had no missing rates >2% across any of its features.

Outlier Detection and Treatment

It is important to identify and address outliers in the data.Outlier detection and treatment are important. Using the Interquartile Range (IQR) method, outliers identified by threshold 3xIQR and Mahalanobis distance were used for multivariate outliers. Univariate outliers were winsorized at 2nd and 98th percentiles to maintain the statistical power without distorting the distribution. Multivariate outliers with a Mahalanobis distance greater than $p<0.001$ were not kept in either data set.

Feature Scaling

To avoid that distance-based and gradient-based algorithms were disproportionately affected by the features with the largest absolute ranges, all the continuous features were standardized by z-score normalization (zero mean, unit variance). Categorical binary variables (family history, outcome) were not scaled.

Class Imbalance Correction

Only the training partition was used to apply SMOTE for creating synthetic minority class samples, which resulted in a balanced training partition of 50:50. To maintain the natural distribution of the test partition, the evaluation metrics were calculated based on the class prevalence in the natural setting. In this instance, k-neighbor of SMOTE was taken as 5, which was in line with the earlier recommendations [18].

Feature Selection

The feature selection was done in three steps: redundant features are removed with the use of Pearson's correlation coefficient (threshold: $|r| > 0.85$); univariate features are ranked by their importance with the help of Analysis of Variance F-statistic (ANOVA-F); the optimal subset of features is selected using Recursive Feature

Elimination with Cross-Validation (RFECV) and Random Forest as a base estimator, and the optimal feature set is determined by the AUC-ROC values calculated over a 5-fold cross-validated dataset. The final feature sets for all data sets were fixed before training the models and used by all classifiers to make comparisons.

Machine Learning Classifiers

The implementation and evaluation of the seven ML classifiers were done using scikit-learn (v1.4.2) in Python (3.11). The algorithms are briefly described theoretically below.

Logistic Regression (LR)

Logistic regression is a model that assumes the log-odds of the binary outcome is a linear combination of the input features. It is computationally efficient, easily interpretable (in terms of odds ratios) and has a probabilistic output. Regularization was used with L2 regularization (Ridge) and the regularization strength C was determined using cross-validation. LR is a linear baseline classification model.

Decision Tree (DT)

Information Gain (Gini impurity) criteria are used to recursively partition the feature space, resulting in an interpretable hierarchal rule structure within the Decision Trees. The parameters 'max_depth', 'n_samples_per_leaf' and 'min_impurity_decrease' were tuned using grid search. Decision Trees can be non-linear and can be very interpretable but they can overfit.

Random Forest (RF)

Random Forest is a bagging ensemble of decision trees, trained on a random subset of the features considered at each split, and with a bootstrap sample. This randomization helps to minimize the variance and correlation between individual trees, resulting in reliable predictions. Some important hyperparameters are n_estimators, max_features and max_depth.

Support Vector Machine (SVM)

Using kernel functions, SVM is able to find the hyperplane that maximizes the margin between classes in a transformed feature space. The Radial Basis Function (RBF) kernel was used, and its hyperparameters, C (regularization) and γ (kernel width) were optimized using a cross-validated grid search. SVM works well with high dimensional spaces and medium class imbalances.

k-Nearest Neighbors (k-NN)

k-NN classifies a sample as the class of the majority of its k closest neighbours in feature space, defined as the Euclidean distance. It is an instance based learning system that does not need to go through a training process. The optimal value of k was obtained by cross validation, ranging from 3 to 21; k-NN is sensitive to feature scaling and irrelevant features.

Gradient Boosting (GB)

The gradient boosting method sequentially trains weak learners (shallow decision trees) to reduce the residuals of the ensemble's previous prediction using gradient descent in function space. The chosen hyperparameters, such as learning rate, number of estimators, max depth and subsampling ratio, for the scikit-learn implementation were used. GB is normally state-of-the-art on tabular data.

Artificial Neural Network (ANN)
The Multi-Layer Perceptron (MLP) consists of two hidden layers and it was implemented using the Adam optimizer with ReLU activation functions. The dropout regularization (rate = 0.3) and early stopping (patience = 10 epochs) techniques were used to prevent overfitting. The architecture was: input layer → Dense(128, ReLU) → Dropout(0.3) → Dense(64, ReLU) → Dropout(0.3) → Dense(1, Sigmoid).

Hyperparameter Optimization
The hyperparameter tuning was done on the LR, DT, SVM, k-NN models using 5-fold stratified cross-validation with Grid Search and on the remaining models (RF, GB, and ANN) by Randomized Search with n_iter=100 because of the large hyperparameter space. The optimization criterion used was the AUC-ROC to account for class imbalance. The final hyperparameters for each model were determined after tuning on the training set and then the tuned models were tested on the held-out test set.

Experimental Protocol
A 80/20 split was obtained using Stratified Random sampling to ensure the same class distribution in the training and testing sets. SMOTE was solely used on the training partition. To get unbiased performance estimates, a nested cross validation was used with an outer cross validation of 5 folds for performance estimation and an inner cross validation of 5 folds to tune the hyperparameters. All random processes were seeded (seed=42) to be reproducible. Experiments were executed on an Intel Xeon E5-2690 workstation with 128 GB RAM and NVIDIA RTX 4090 GPU.

Evaluation Metrics
Model performance was assessed using the following metrics, derived from the confusion matrix elements (True Positive [TP], True Negative [TN], False Positive [FP], False Negative [FN]):
Accuracy: (TP + TN) / (TP + TN + FP + FN)
Sensitivity (Recall): TP / (TP + FN)
Specificity: TN / (TN + FP)
Precision: TP / (TP + FP)
F1-Score: 2 × (Precision × Recall) / (Precision + Recall)
Matthews Correlation Coefficient (MCC): $(TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$
AUC-ROC: Area under the receiver operating characteristic curve

MCC was included as it provides a more balanced measure of classification quality than accuracy, particularly in scenarios with class imbalance. Statistical significance of performance differences was assessed using the Friedman test (across all classifiers) followed by Wilcoxon signed-rank post-hoc tests with Bonferroni correction for pairwise comparisons.

Results
Dataset Characteristics After Preprocessing
After pre-processing, the characteristics of the dataset were extracted. The PIDD, after preprocessing, contained 768 instances, with 8 features and no missing data. The secondary clinical dataset consisted of 1,498 observations (after removing 25 observations as multivariate outliers), including 12 features. Class distributions were equalized at 50:50 after SMOTE (test set distributions had natural class prevalences (PIDD: 34.9% positive; secondary: 40.2% positive)). Table 1 presents descriptive statistics for the important variables in the preprocessed PIDD.

Feature	Min	Max	Mean ± SD	Diabetic Mean	Non-Diabetic Mean
Pregnancies	0	17	3.85 ± 3.37	4.87 ± 3.56	3.30 ± 3.02
Glucose (mg/dL)	44	199	121.7 ± 30.4	141.3 ± 31.9	110.6 ± 26.1
Blood Pressure (mm Hg)	24	122	69.1 ± 19.4	70.8 ± 21.4	68.2 ± 18.1
Skin Thickness (mm)	7	99	29.2 ± 10.7	32.4 ± 10.5	27.2 ± 10.5
Insulin (mu U/ml)	14	744	155.5 ± 118.8	186.6 ± 139.7	137.9 ± 96.3
BMI (kg/m²)	18.2	67.1	32.4 ± 6.9	35.1 ± 7.3	30.9 ± 6.5
Pedigree Function	0.078	2.42	0.472 ± 0.331	0.551 ± 0.371	0.430 ± 0.299
Age (years)	21	81	33.2 ± 11.8	37.1 ± 11.3	31.2 ± 11.7

Table 1. Descriptive statistics of PIDD features after preprocessing (N=768). SD: standard deviation.

Feature Selection Results

Based on the data from the PIDD, RFECV selected six optimal features: glucose concentration, BMI, Age, Diabetes pedigree function, Insulin, and Number of pregnancies. Due to the low ANOVA-F scores and poor contribution to cross-validated AUC, blood pressure and skin thickness were dropped. The secondary data analysis kept nine features: fasting glucose, HbA1c, BMI, waist circumference, family history, triglycerides, HDL cholesterol, age, and systolic blood pressure. LDL cholesterol, diastolic blood pressure and physical activity index were not included.

Comparative Classification Performance on PIDD

The overall comparative performance results for all seven classifiers over the PIDD test set are given in Table 2. Gradient Boosting had the best results on most of the metrics with an accuracy of 91.4%, sensitivity of 87.3%, specificity of 93.8%, and AUC-ROC of 0.958. Random Forest was the second best with an accuracy of 89.7% and AUC-ROC of 0.943. The ANN resulted in a competitive AUC-ROC value of 0.931 with a somewhat poorer accuracy result of 87.1%. This was the simplest model, Logistic Regression, with an AUC-ROC of 0.847, which was respectable, given the fact that it is a linear model.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score	MCC	AUC-ROC
Gradient Boosting	91.4	87.3	93.8	89.1	0.913	0.819	0.958
Random Forest	89.7	85.4	92.1	87.3	0.863	0.786	0.943
ANN (MLP)	87.1	83.2	89.7	84.8	0.840	0.744	0.931
SVM (RBF)	84.6	80.1	87.4	81.7	0.809	0.692	0.912
k-NN (k=9)	81.3	76.5	84.2	78.9	0.776	0.624	0.878
Decision	78.5	75.1	80.6	75.3	0.752	0.571	0.832

Tree							
Logistic Regression	79.6	74.3	82.9	77.2	0.757	0.587	0.847

Table 2. Comparative classification performance metrics on the PIDD test set (N=154). Best values per metric highlighted. MCC: Matthews Correlation Coefficient; AUC-ROC: Area Under the Receiver Operating Characteristic Curve.

Results on Secondary Clinical Dataset

The secondary clinical dataset's performance metrics are shown in Table 3. The performance ranking was generally similar to that of PIDD, but absolute performance was slightly improved for the majority of classifiers, possibly due to the greater number of features (HbA1C and lipid profiles). The gradient boosting model did the best job, with an accuracy of 93.1% and AUC-ROC of 0.971. Interestingly, the performance improvement observed for LR was greater with this dataset (accuracy: 83.4%) than PIDD, which may be due to the presence of linearly separable clinical biomarkers (HbA1c and fasting glucose)..

Classifier	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score**	**AUC-ROC**
Gradient Boosting	93.1	90.2	95.3	0.928	0.971
Random Forest	91.6	88.7	93.8	0.912	0.962
ANN (MLP)	89.4	86.1	91.9	0.890	0.951
SVM (RBF)	87.2	83.4	90.2	0.853	0.934
k-NN (k=7)	84.8	81.3	87.5	0.824	0.906
Decision Tree	80.3	77.6	82.4	0.787	0.851
Logistic Regression	83.4	79.8	86.2	0.818	0.878

Table 3. Comparative classification performance on the secondary clinical dataset test set (N=300).

Statistical Significance Testing

Assigning a p value to the Friedman test showed statistical significance for the differences in AUC-ROC scores between the seven classifiers (Chi-squared = 28.4, df = 6, $p < 0.001$). The post hoc Wilcoxon signed-ranked p values between Gradient Boosting and Logistic Regression, Decision Tree and k-NN resulted in significant differences at $p = 0.008$, $p = 0.006$ and $p = 0.014$ respectively with Bonferroni correction. After correcting for multiple comparisons, no difference was statistically significant between Gradient Boosting and Random Forest ($p = 0.087$) or between Random Forest and ANN ($p = 0.112$), indicating similar performance of these three best models.

Feature Importance Analysis

The evaluation of the feature importance was done in three different ways: (1) Random Forest permutation importance; (2) Gradient Boosting feature gain; (3) SVM based recursive feature elimination. In all three methods, the most important predictor was always plasma glucose concentration followed by BMI, age and diabetes pedigree function. The moderate importance of insulin and number of pregnancies and the consistently low importance of blood pressure were obtained. These rankings are comparable to the clinical knowledge, thus providing face validity for the models. Table 4. Normalized scores of the Gradient Boosting feature importance.

Rank	Feature	GB Importance Score	RF Permutation Score	Clinical Relevance
1	Plasma Glucose (mg/dL)	0.312	0.298	Direct DM diagnostic criterion
2	BMI (kg/m²)	0.241	0.226	Primary obesity-DM mediator
3	Age (years)	0.178	0.191	Cumulative metabolic dysfunction
4	Diabetes Pedigree Function	0.142	0.138	Hereditary risk quantifier
5	Insulin (mu U/ml)	0.074	0.088	Insulin resistance biomarker
6	Pregnancies	0.053	0.059	Gestational DM risk proxy

Table 4. Normalized feature importance scores from Gradient Boosting and Random Forest permutation importance (PIDD). GB: Gradient Boosting; RF: Random Forest; DM: Diabetes Mellitus.

Computational Performance

The training time and prediction time were much different among algorithms (Table 5). Logistic Regression was the quickest to train (0.003 seconds) and predict, followed by Decision Tree and k-NN. Gradient Boosting took around 8.7 seconds to train, while ANN took the longest training time of 24.3 seconds owing to iterative backpropagation. All models showed prediction times well below 1 second in clinical deployment scenario where real-time prediction is desired, meeting the practical clinical decision support requirements.

Classifier	Training Time (sec)	Prediction Time (ms/sample)	Memory (MB)
Logistic Regression	0.003	0.02	0.8
Decision Tree	0.021	0.04	2.1
Random Forest	2.14	0.31	48.6
SVM (RBF)	3.87	0.18	12.4
k-NN (k=9)	0.001	0.87	6.2
Gradient Boosting	8.73	0.09	31.2
ANN (MLP)	24.3	0.11	18.7

Table 5. Computational performance metrics on PIDD training/test partition (Intel Xeon E5-2690, 128 GB RAM).

Discussion

Performance of Individual Classifiers

The findings from this study align with the general trend in the field of classification of tabular clinical data, where Gradient Boosting and Random Forest surpass other classifiers, and are consistent with the work of other researchers. GB's iterative error correction process can progressively reduce bias, and the RF's ensemble randomization can reduce variance, resulting in models that are good at generalization to unseen data. The results are consistent with those of Zou et al. [10], Nnamoko and Korkontzelos [14] and a recent meta-analysis by Krishnamoorthi et al. [20] that found ensemble methods outperformed single algorithm methods in 34 diabetes prediction studies in a statistically significant manner.

The competitive performance of ANN (AUC-ROC: 0.931) indicates that, as the sample size grows larger (in the case of PIDD it is relatively small, N=768), deep learning can provide useful benefits through representation learning. This hypothesis was partially supported by the higher performance of the secondary clinical data in the ANN (AUC-ROC: 0.951). SVM with RBF kernel worked well and might be used in situations where the datasets are not linearly separable, and the computational power is not a critical factor.

Even though Logistic Regression is simple model, it got AUC-ROC of 0.847 on PIDD which shows that there is still a significant portion of the variance in diabetes outcome that is linearly separable. LR's strength is that it is easily interpreted—for log-odds coefficients are directly clinically relevant—and has a probabilistic output, which is useful for risk-stratification instead of binary classification. In the face of limited resources in a clinical environment, LR could be the best compromise between performance and deployability..

Clinical Implications

The four key predictors identified – plasma glucose, BMI, age and diabetes pedigree – are clinically logical and reassuring. The plasma glucose is the key factor in defining a diagnosis of diabetes, while BMI quantifies the insulin resistance associated with adiposity, age quantifies the cumulative metabolic loss, and the pedigree function represents the hereditary susceptibility. The features captured in primary care are common practice, indicating that a Gradient Boosting classifier built from these features might be easily incorporated into existing EHR processes without requiring significant extra data collection burden.

The high sensitivity of Gradient Boosting (87.3% on PIDD) is of great clinical relevance. In diabetes screening, a false negative – a pre-diabetic or diabetic person is identified as non-diabetic – is more clinically costly than a false positive, since a person gets no chance for early life style or pharmacological interventions if they have a diabetes-related condition. A model that has a high sensitivity level will have a low rate of false negatives and will therefore be well suited to screening applications of the population. The specificity of 93.8% is also high, reducing the unnecessary diagnostic workup and associated patient anxiety in patients who are incorrectly classified as high-risk.

The overall performance of the secondary clinical data, which included HbA1c as well as lipid indices, in addition to the PIDD features, further highlighted the clinical relevance of a comprehensive metabolic profile. The independent variable HbA1C, which reflects the average glycemia over the last 2–3 months, is more closely mechanistically linked with the outcome variable compared to a single fasting glucose measurement, and its addition significantly enhanced model performance. This discovery calls for the incorporation of the HbA1c into general health check-up panels for which diabetes prediction models should be used.

Comparison With Prior Literature

Our Gradient Boosting model, with feature selection, gave an AUC-ROC of 0.958 for PIDD which is comparable with the best reported in the recent literature: Using XGBoost with feature selection, Abdollahi et al. [21] reported AUC-ROC of 0.94 for PIDD; Mujumdar and Vaidehi [22] reported AUC-ROC of 0.92 for PIDD using their hybrid SVM-DT model; Tasin et al. [23] reported AUC-ROC of 0.89 using CNN method on PIDD. We are also contributing to the good performance with our methodologically unified experimental framework, our comprehensive preprocessing pipeline (with SMOTE, only on the training data, and stringent hyper-parameter tuning).

Preprocessing and SMOTE

One of the main methodological advances of this study is the use of the SMOTE technique on the training partition only, which is not used by many published studies. If SMOTE is performed prior to train-test splitting, the synthetic samples can be placed in both the train and test sets leading to information leakage and overestimation of performance. We employ a strategy for ensuring that evaluation measures reflect expected performance of the model applied to naturally imbalanced clinical populations. The advantages of class-imbalance correction in medical ML applications, revealed by the improvement in sensitivity obtained by SMOTE (average: +6.8 percentage points for all models), are significant.

Limitations and Future Directions

There are a number of limitations to this study which must be recognized. PIDD is a specific group of Pima Indian women, and may be less representative of other ethnic groups around the world who have different genetic make-ups, lifestyle factors and disease rates. The use of a secondary data set from a hospital helps to reduce this but further validation on larger, more ethnically diverse cohorts is needed to ensure the generalizability of the results.

Second, the cross-sectional nature of both data sets makes it difficult to model the temporal trajectories that may include additional predictive information, such as longitudinal changes over time in glucose or BMI. Longitudinal EHR data modeled using recurrent architectures (LSTM, Transformer-based models) is an exciting area for future research to enhance early prediction by accounting for the dynamics of disease progression.

Third, although feature importance analysis can generate some interesting insights about the model's decision logic, ensemble models like Gradient Boosting and Random Forest are far less interpretable than logistic regression. Explainability features, including SHAP values, should be embedded in the clinical deployment, giving clinicians feature-level explanations, thus promoting trust and adherence to new frameworks in the AI-in-healthcare space.

Fourth, the models which are produced here are not validated in a prospective manner in a clinical decision support context. Whether or not a prediction model is useful in the real world of clinical practice must be evaluated by prospective deployment trials that track downstream outcomes like an early diagnosis rate, uptake of interventions and reduction in complications. A randomized controlled trial between the two clinical decisions, with and without making use of the ML-assisted prediction, would give the best evidence of the actual effectiveness in real-world situations.

Future research directions: (1) Federated Learning for training multi-institutional models without sharing patient data, which has privacy and data governance issues; (2) Integration of Genomic and Microbiome data in multi-modal precision prediction; (3) Active Learning for efficient data labeling in resource limited setting; (4) Fairness-aware ML for evaluating and mitigating algorithmic bias across demographic subgroups; (5) Real-time deployment of optimized models in EHR platforms with model recalibration and continuous monitoring mechanisms..

Conclusion

In this study, a systematic and rigorous comparative analysis of seven machine learning classifiers for early prediction of diabetes mellitus with structured clinical data is presented. Gradient Boosting had the best results for both two datasets with AUC-ROC scores of 0.958 on the PIDD dataset and 0.971 on the secondary clinical dataset, and accuracy scores of 91.4% and 93.1% respectively. Random Forest and ANN were the second and third best performing models. Logistic Regression, although inferior on most metrics, is the most interpretable output and could be a more practical approach in deployment environments that have limited resources.

The essentiality of strict preprocessing, especially missing value imputation using KNN, Winsorization of outliers, and application of training-only SMOTE was shown by systematic ablation and was found to be meaningful for both performance and the validity of the evaluation. Plasma glucose, BMI, age and diabetes pedigree were repeatedly found to be the most important for the developed model in the feature importance analysis which gave face validity to the models and is in accordance with the clinical knowledge of the etiology of T2DM.

This study confirms the value of incorporating ML prediction tools, especially ensemble models, into primary care and population health screening practices. These high sensitivity and AUC-ROC values indicate that tools like these can be valuable in supplementing clinical decision making in determining high risk individuals who could benefit from early preventive intervention. Data-driven risk prediction is an essential measure of a comprehensive public health approach, especially as the prevalence of diabetes continues to grow at a global level, and scalable solutions are needed..

Acknowledgements

The authors would like to express their gratitude to the clinical staff of the hospitals that participated in the data collection and de-identification process. This work was funded by KAUST Research Funding Initiative (Grant RFI-2023-ML-12), Department of Biotechnology, Government of India (Grant BT/PR-45210/AI/2023) and the Natural Science Foundation of China (Grant 82271946). The funding bodies played no part in the design of the study, the collection, analysis, interpretation or decision to publish.

Conflicts of Interest

The authors declare that there are no conflicts of interest. The secondary clinical data was used with IRB approval number IRB-2024-ML-047 and all patient information was completely de-identified following HIPAA and GDPR standards.

Data Availability

The Pima Indians Diabetes Dataset can be found in the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/diabetes. The secondary clinical data can be shared with reasonable requests upon approval of institutional data governance to the corresponding author. The code used for all the analysis can be found at https://github.com/ml-diabetes-prediction/comparative-study.

References

[1] International Diabetes Federation. IDF Diabetes Atlas, 11th Edition. Brussels: IDF; 2025. Available from: https://www.diabetesatlas.org

[2] American Diabetes Association Professional Practice Committee. Standards of Medical Care in Diabetes—2025. Diabetes Care. 2025;48(Suppl 1):S1–S352. doi:10.2337/dc25-S000

[3] Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346(6):393–403. doi:10.1056/NEJMoa012512

[4] Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725–731. doi:10.2337/diacare.26.3.725

[5] Obermeyer Z, Emanuel EJ. Predicting the future: Big Data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–1219. doi:10.1056/NEJMp1606181

[6] Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020;10(1):11981. doi:10.1038/s41598-020-68771-z
[7] Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–127. doi:10.1016/j.ijmedinf.2016.09.014
[8] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–116. doi:10.1016/j.csbj.2016.12.005
[9] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018;132:1578–1585. doi:10.1016/j.procs.2018.05.122
[10] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515. doi:10.3389/fgene.2018.00515
[11] Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. Lecture Notes in Computer Science. Berlin: Springer; 2000:1–15. doi:10.1007/3-540-45014-9_1
[12] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–1232. doi:10.1214/aos/1013203451
[13] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016:785–794. doi:10.1145/2939672.2939785
[14] Nnamoko N, Korkontzelos I. Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med. 2020;104:101815. doi:10.1016/j.artmed.2020.101815
[15] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–1358. doi:10.1056/NEJMra1814259
[16] Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094. doi:10.1038/srep26094
[17] Maniruzzaman M, Kumar N, Abedin MM, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed. 2017;152:23–34. doi:10.1016/j.cmpb.2017.09.004
[18] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357. doi:10.1613/jair.953
[19] Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the Annual Symposium on Computer Application in Medical Care; 1988:261–265.
[20] Krishnamoorthi R, Joshi S, Almarzouki HZ, et al. A novel diabetes healthcare disease prediction framework using machine learning techniques. J Healthc Eng. 2022;2022:1684017. doi:10.1155/2022/1684017
[21] Abdollahi J, Nouri-Moghaddam B. Hybrid stacked ensemble method with feature selection for diabetes prediction. Iran J Comput Sci. 2022;5(2):103–121. doi:10.1007/s42044-021-00100-9
[22] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Comput Sci. 2019;165:292–299. doi:10.1016/j.procs.2020.01.047
[23] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2023;10(1-2):1–10. doi:10.1049/htl2.12039