Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 July 2026
Revised 27 Aug 2026
Accepted 11 Sep 2026

Natural Resources for Human Health 2026, Vol.6(21s):1026–1050

KEYWORDS:
Diabetes Prediction, Hybrid Machine Learning, False Negative Analysis, Clinical Reliability, Ensemble Learning, Probability Calibration, Threshold Optimization, Medical Decision Support.

A Hybrid Machine Learning Framework for Diabetes Prediction with Enhanced False Negative Reduction and Clinical Reliability

Gurdeepika Kaur[1], Harsh Sadwarti[2]

[1]Department of CSA, Faculty of Engineering Technology and Computing , Desh Bhagat University, Mandi Gobindgarh, Punjab, India
[2]Department of Computer Science and Engineering, Faculty of Engineering Technology and Computing , Desh Bhagat University, Mandi Gobindgarh, Punjab, India
Email Id: gurdeepikakaur1804@gmail.com, vc@deshbhagatuniversity.in

ABSTRACT: Accurate and clinically reliable prediction of diabetes is essential for early diagnosis and prevention of severe complications. This paper proposes a hybrid machine learning framework for diabetes prediction with a specific focus on minimizing false negative errors, which are critical in clinical decision-making. The framework is evaluated on a large-scale dataset comprising 100,000 patient records with demographic, clinical, and biochemical features. A diverse set of models was employed, including Naive Bayes, Logistic Regression, K-Nearest Neighbors, Gradient Boosting, Support Vector Machine, Decision Tree, Random Forest, and advanced boosting algorithms such as LightGBM, XGBoost, and CatBoost, along with a Tabular Neural Network. Three hybrid models were developed to enhance predictive performance. Among them, the neural network and XGBoost hybrid achieved the best results with an accuracy 98.45%, PR-AUC of 0.883 and ROC-AUC 97.82%. Feature analysis identified HbA1c and blood glucose level as dominant predictors. Probability calibration and threshold tuning improved sensitivity to 90.29%, significantly reducing false negatives. Detailed error analysis revealed that misclassified cases exhibit borderline clinical values and overlapping feature distributions. The proposed framework demonstrates improved robustness and suitability for real-world clinical decision support systems.

1. INTRODUCTION

Diabetes is a chronic metabolic disorder affecting millions of people. It happens when the pancreas does not secrete adequate insulin, or the body is unable to respond adequately to the insulin produced leading to higher blood glucose levels [1]. The uncontrolled diabetes can lead to complications, some of which are nerve damage, heart disease, vision impairment, kidney failure, and amputation of limbs. Over the last few years, the number of diabetic patients has increased significantly, with millions of adults having the disease - and the number is projected to continue increasing over the decades ahead [2,3]. A large percentage of the population remains undiagnosed, which is a major problem because it implies that the disease will be silently progressing, and a threat of harmful complications will increase [4]. The typical tests in clinical practice to diagnose diabetes include fasting blood glucose, oral glucose tolerance tests, random glucose, and level of HbA1c. Even though these practices are undoubtedly popular, there are certain disadvantages. Some of these tests require certain testing like fasting that may not be performed accordingly. Others are either time-consuming or affected by the temporary health conditions [5,6]. Even HbA1c, as convenient as it is, might not give the correct results to all people because of biological variation or medical problems. Consequently, this can lead to the failure to identify some cases of diabetes, which will postpone treatment and deteriorate the health outcomes [7]. As technology is improving, machine learning and artificial intelligence are emerging as promising tools to enhance disease prediction. The methods can analyze data, such as clinical measurements, lifestyle, and demographic data, in large quantities, to determine complex patterns that could be missed in a traditional approach [8,9]. Some of the algorithms such as the support vector machine, logistic regression, random forests and boosting methods have shown promising results in improving prediction accuracy. However, this approach faces several challenges. The datasets that

are used in most of the research are small or limited datasets that might not be representative of the real world [10]. The diabetes data is often imbalanced with more non-diabetic subjects compared to the diabetic subjects. This implies that a given model may be very effective, but it may not detect numerous cases of diabetes. Specifically, the false negatives (misclassification of diabetic cases as non-diabetic) are of great concern [11]. These mistakes are especially dangerous because they may result in a late diagnosis and treatment, and complications. One of the trends in the existing literature is to pay more attention to accuracy, but not to such clinically relevant errors. The other issue is that models are typically evaluated using standard measures of performance and without considering clinical reliability [12]. Moreover, it is not sufficient to use one machine learning model to model diabetes as various models possess varying properties. To address these limitations, a combination of machine learning models to predict diabetes, with an emphasis on performance and reliability is proposed in the present study. It combines multiple models to take advantage of their respective strengths, uses resampling to handle class imbalance and threshold optimization to reduce false negatives. It also involves analysis of errors to understand why the model makes mistakes. This research aims at producing a a more effective and practical tool of early diagnosis of diabetes and assist in healthcare decision-making.

1.1 Contribution

This study presents a novel hybrid machine learning framework for diabetes prediction with a strong emphasis on clinical reliability and false negative reduction. The key contributions are summarized as follows:

- A clinically oriented prediction framework is proposed that prioritizes the reduction of false negative errors, addressing a critical limitation in existing diabetes prediction models where missed diagnoses are often overlooked.
- Development of three complementary hybrid architectures, integrating linear models, ensemble learning, boosting techniques, and deep learning, enabling effective modeling of complex and overlapping feature distributions.
- Incorporation of probability calibration techniques to align predicted probabilities with true clinical risk, enhancing the reliability of model outputs for decision-making.
- A comprehensive false negative analysis framework is proposed, distinguishing between borderline and high-confidence misclassifications, providing deeper insights into model behavior and diagnostic limitations.

2. RELATED WORK

Over the past few years, machine learning (ML) and artificial intelligence (AI) have experienced substantial growth in the field of diabetes prediction due to their ability to capture complex interrelationships between clinical and demographic variables. Various algorithms, datasets, and assessment plans have been investigated to enhance the predictive performance and aid in early diagnosis. Several studies have been conducted to compare traditional machine learning algorithms to diagnose diabetes. Often, a good example is that, as illustrated by Tasin et al. (2022) [13], the development of an automated prediction framework is carried out on the basis of multiple classifiers, i.e., decision trees, support vectors machine, and ensemble methods, proving the superiority of the ensemble models versus individual algorithms in relation to accuracy and robustness. In a similar fashion, Mao et al. (2022) [14] tested machine learning models using real-world clinical data and found that the methods like random forests and gradient boosting are more predictive than traditional statistical methods. Shin et al. (2022) [15] also highlighted the significance of large-scale electronic medical record datasets, demonstrating that large-scale datasets can be effectively trained to produce high predictive accuracy and enhanced generalization. In addition to conventional models, there is new literature that has investigated the effect of feature engineering and data preprocessing. Rahman et al. (2023) [16] emphasized the contribution of the exploratory data analysis and feature selection in enhancing prediction accuracy, especially when employing socio-demographic and clinical characteristics. Moreover, Wee et al. (2024) [17] made a comparison between machine learning and deep learning models and showed that deep learning models can capture more complex non-linear patterns, but need larger datasets and more computational resources. Later research has moved more to ensemble and hybrid methodologies in order to address the shortcomings of single-model systems. In Ahmed et al. (2024) [18], the authors tested several machine learning models and demonstrated that hybrid combinations of the classifiers increase the accuracy of diagnoses and decrease the number of misclassification errors. On the same note, hybrid frameworks that combine support vector machines, decision trees and random forests have been shown to have better performance through the advantageous features of various algorithms. Ayoade et al. (2025) [19] also established that hybrid and ensemble models offer a greater balance between accuracy and computational efficiency than do standalone models.

Simultaneously, a number of studies have explored issues related to real-world machine learning models deployment. Afolabi (2025) [20] pointed out the necessity to choose the right algorithm to capture non-linear relationships in clinical data, whereas Khokhar et al. (2025) [21] in their systematic review found the following issues, namely, class imbalance, limited data diversity, and lack of interpretability as key obstacles to clinical adoption. The relevance of machine learning models in forecasting the risk of diabetes with the help of real-world population data was proved by Lee et al. (2025) [22], which may find its application in preventive care. Also, El-Bashbishy et al. (2026) [23] suggested the use of a sophisticated prediction mechanism that includes feature-selection and imbalance-management methods, which supports the significance of managing data-related issues to ensure high performance. With these advancements, there are several very important limitations. Most of the current literature uses extensive benchmark data and this can be misleading of the real-world population diversity. Also, imbalance of classes in medical data frequently results in remove repetition towards the majority class and thus lack sensitivity to diabetic cases. Most significantly, most of the previous literature addresses general accuracy or AUC-related measures of evaluation, and there is little emphasis on minimization of false negatives. False negatives wherein diabetic patients are misdiagnosed as non-diabetic are especially perilous in the context of diabetes prediction because they postpone the diagnosis and treatment process. Consequently, more robust and clinically reliable strategies are required that can not only attain a high predictive accuracy but also explicitly consider such significant issues as class imbalance, model interpretability, and reduction of false negatives. The presence of such gaps is what drives the creation of hybrid machine learning frameworks, which combine various models and are focused on clinically meaningful results, which is the foundation of the proposed work.

Table 1 is a summarization of the existing research on machine learning in predicting diabetes. The table provides a description of the key characteristics of each of the studies, including the datasets, the methodologies used to perform the studies, the performance of the studies, and the limitations. This comparison demonstrates change in machine learning techniques in diabetes prediction and also demonstrates some problems in general (such as reliance on datasets, non-clinical validation, and emphasis on false negative reduction). The experiences based on this analysis allow finding the proposed work in the existing literature and justify the need to have a more solid and clinically valid framework.

Table 1: Comparative analysis of recent studies on diabetes prediction using ML techniques

Authors	Dataset	Techniques	Reported Outcomes	Limitations
Tasin et al. (2022) [13]	PIMA + external dataset	RF, SVM, DT, XGBoost with SMOTE/ADASYN	Ensemble models outperformed single classifiers; RF and boosting methods showed best performance (Acc- 81%, AUC = 84%)	Relies on benchmark datasets; limited generalization
Mao et al. (2022) [14]	Real-world clinical cohort	RF, Gradient Boosting, Neural Networks	ML models achieved strong predictive performance (AUC-85.5%) and outperformed traditional statistical methods (84%)	Limited interpretability; no focus on clinical errors
Shin et al. (2022) [15]	Electronic Medical Records	Machine learning classifiers	High predictive capability using EMR data; demonstrated feasibility in clinical settings; Accuracy – 80.7%; AUC – 88%	Lack of external validation; dataset-specific bias
Rahman et al. (2023) [16]	Socio-demographic dataset	LR, SVM, Decision Tree	Feature engineering improved model performance; identified key risk factors; Accuracy – 99.36%	Absence of clinical biomarkers reduces reliability
Wee et al. (2024) [17]	Multiple datasets	Deep Learning + ML models	Deep learning captured complex patterns better than traditional ML	High computational cost; requires large datasets
Ahmed et al. (2024) [18]	Clinical dataset	Ensemble machine learning	Ensemble models improved accuracy and sensitivity compared	False negatives not explicitly minimized

			to single models	
Albadri et al. (2024) [19]	PIMA dataset	Hybrid (SVM + RF + DT)	Hybrid model outperformed individual classifiers; Accuracy – 80%; Precision – 82%; Error rate – 20%; Sensitivity – 88%	Small dataset; risk of overfitting
Afolabi et al. (2025) [20]	Public dataset	Logistic Regression, Random Forest, KNN	ML models effectively captured non-linear relationships KNN – Accuracy: 96.09%, Sensitivity – 98.54%, Specificity – 93.63%, Mean Test Error – 0.0391	Limited robustness and generalization analysis
Khokhar et al. (2025) [21]	Multiple studies (Review)	Survey of ML/DL methods	Identified ensemble learning and feature engineering as key trends	No experimental validation
Ayoade et al. (2025) [22]	Multiple datasets	ML, DL, Hybrid models	Hybrid models achieved better balance between accuracy and efficiency F1Score- AdaBoost (0.74), Random Forest (1.00), DNN (0.45)	Limited clinical validation
Lee et al. (2025) [23]	Real-world population data	Extreme Gradient Boosting	Effective in identifying high-risk individuals in clinical populations Accuracy – 84.88%, Precision – 77.92%, Recall -66.91%, F1 Score-72%, AUC – 79.57%	Limited interpretability
El-Bashbishy et al. (2026) [24]	Clinical dataset	Ensemble + feature selection	Improved robustness and predictive performance Accuracy – 98.87%, Precision – 98.61%, Recall -97.92%, F1 Score-98.51%, AUC – 99.99%	Class imbalance remains a challenge

3. MATERIALS AND METHODOLOGY

This section outlines the general approach that this research employs to predict diabetes, revealing the flow of the process, starting with the raw data and ending with significant clinical implications (Figure 1). The process begins with the diabetes dataset which is thoroughly prepared using preprocessing methods that include managing the class imbalance, exploratory data analysis, dimensionality reduction using PCA, and the most important clinical features. After preparing the data, it is employed in the training of various kinds of models such as baseline machine learning models and ensemble training models and the proposed hybrid models. Probability calibration and threshold optimization are then added to the predictions made by these models. The steps are useful in making the predictions more reliable, but more importantly, minimizing false negative cases, which is vital in making sure that diabetic patients are not missed in the diagnosis process.

Natural Resources for Human Health
eISSN: 2583-1194

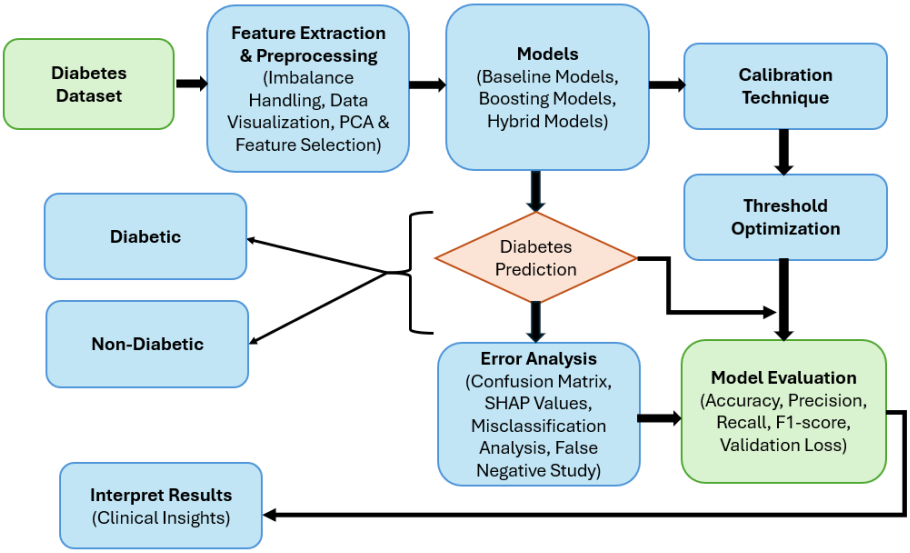


Figure 3.1: Proposed hybrid machine learning framework for diabetes prediction

Dataset Description: The data provided includes medical and demographic data of the patients and their diabetes as a binary outcome. It consists of several clinically significant characteristics as shown in Table 2. The records contain one patient each, and the target variable denotes whether the patient is diabetic or not. The dataset is designed for the prediction of diabetes through machine learning by integrating physiological measurements and lifestyle-related factors. Biochemical measures such as HbA1c and blood glucose level provide an indicator of blood glucose control. Lifestyle and other demographic factors - such as age, sex and smoking - also affect the risk of diabetes. Together, these factors make the data set ideal for developing models that incorporate both clinical state and lifestyle factors that affect diabetes. Consequently, it can help analyze risk more effectively of diabetes and create the models that can help healthcare professionals detect and intervene in a timely manner [25].

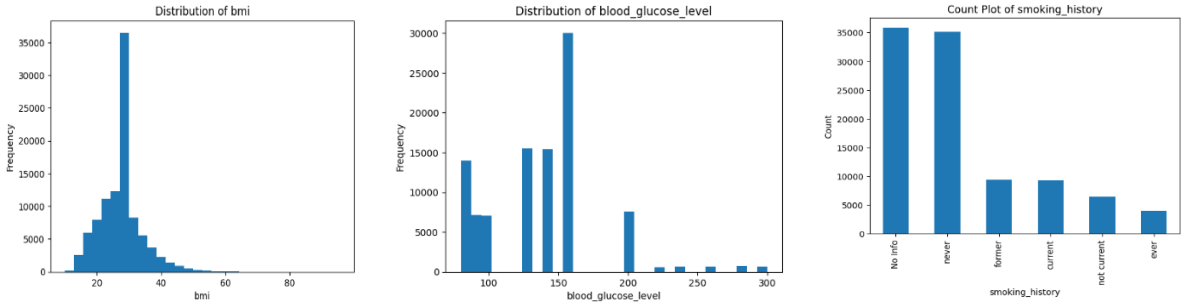
Table 2: Description of dataset features, including their types and clinical relevance in diabetes prediction

Feature	Type	Clinical Relevance
Age	Demographic	Increasing age is strongly associated with higher diabetes risk due to metabolic decline
Gender	Demographic	Helps analyze gender-based differences in diabetes prevalence
BMI	Clinical (Anthropometric)	High BMI is a major risk factor for insulin resistance and type 2 diabetes
Hypertension	Clinical (Comorbidity)	Strongly linked with metabolic syndrome and diabetes complications
Heart Disease	Clinical (Comorbidity)	Indicates cardiovascular risk often associated with diabetes
Smoking History	Behavioral	Smoking contributes to insulin resistance and increased diabetes risk
HbA1c Level	Biochemical	Key indicator of long-term blood glucose control (diagnostic marker)
Blood Glucose Level	Biochemical	Direct measure of current glucose level; critical for diagnosis
Diabetes (Target)	Outcome	Binary variable indicating presence or absence of diabetes

Data Preprocessing and Visualization: The dataset contains 100,000 records of patients with nine characteristics as demographic, clinical, and biochemical features. Before model development, a preprocessing steps were undertaken to guarantee the quality and consistency of data. The first step was to check the dataset for missing and non-congruent values.

The smoking history feature included the entries like "No Info", which were kept as a separate category to save the information that would otherwise be lost, as well as keep the possible patterns intact. One-hot encoding was used to encode categorical variables such as gender and smoking history to enable them to be used by machine learning algorithms. The standard scaling was applied to normalize numerical variables so that they make equal contributions when training the model. This was an essential step because the dataset has characteristics of different scales and units. There was also outlier analysis was performed on features such as BMI and glucose to determine the extreme values that may affect the performance of the model. Additionally, the issue of class imbalance between the diabetic and non-diabetic samples was resolved with the help of resampling methods, including SMOTE, which helped the model recognize the minority class cases. The dimensionality of the data was reduced using Principal Component Analysis (PCA) to capture the important information in the data and to avoid multicollinearity between features. Prior to these methods, Exploratory Data Analysis (EDA) was used to gain insights into the data, such as distribution of features, correlation between variables and problems with the data. EDA highlighted key characteristics such as skewed distributions, presence of outliers, and class imbalance. This guided the choice of suitable data preprocessing methods such as normalization, resampling, and feature transformation. The plots obtained from EDA also provides insights into the characteristics of the dataset and features that are important for predicting diabetes.

As shown in Figure 2(a), the distribution of the BMI values is skewed to the right, most of the people are concentrated between the range of about 20 to 35. The distribution exhibits a distinct peak around the normal to overweight range with slightly less samples in the high BMI range. It means that most of the people in the data are of normal or slightly overweight categories, and a smaller percentage are obese. The long tail to the right of the BMI values illustrates the occurrence of people with high body weights, which is clinically significant since obesity is a proven risk factor of type 2 diabetes. Moreover, the skewed distribution, may imply a lack of balance in the representation of features, thus, influencing model learning and necessitating suitable preprocessing methods like normalization or scaling to provide stable model behavior.



(a) (b) (c)

Figure 2: Exploratory data analysis of key features in the diabetes dataset

As illustrated in Figure 2(b), the distribution of blood glucose levels exhibits a multi-modal pattern which indicate a non-normal and partially discrete distribution. Most values are concentrated in the range of approximately 100 to 160, with additional peaks at specific intervals, suggesting clustering around common measurement points or clinically significant levels. There are also higher glucose levels that extend to 200 and above, which represent individuals with hyperglycemia, which is closely related to diabetes. These cases are extreme values, which are reflective of diabetic or high-risk cases and are very crucial in differentiating between classes. The distribution observed indicates the high predictive power of blood glucose level that complies with its high significance in the model and supports its clinical significance in the diagnosis of diabetes.

Figure 2(c) shows the count plot of the smoking history, which demonstrates the frequency of the categorical smoking behavior in the data set, with the most common categories being No Info and never, then there are the former, current and other categories. This means that a large percentage of the data either comprises of non-smokers or the smoking data is not captured. The large percentage of entries with No Info can also cause some uncertainty and possible bias since the lack of behavioral data can be a hindrance in the model to fully address the risk factors related to lifestyle. Although smoking related categories are not as predominant as other features, they are still of clinical importance as smoking is linked to insulin resistance and metabolic syndrome. The disproportion between the categories, however, indicates that this feature might not have a relatively strong predictive value in comparison to the most important biochemical variables including HbA1c and blood glucose levels.

Feature Importance: The analysis of the importance of features based on the Random Forest model shows that glycemic indicators are dominated in predicting diabetes (Figure 3). Among them, HbA1c is the most significant predictor with an importance score of 0.40 and then comes blood glucose level with an importance score of 0.33. This shows that long-term and short-term glucose values are the key variables affecting the model predictions and this is in line with known clinical diagnosis levels. Comparatively, the BMI (0.12) and age (0.10) have a moderate effect meaning that they are significant risk factors; however, it is not as powerful as direct biochemical indices. The effect of comorbid conditions, like hypertension and heart disease have minimal impact (around 0.02), whereas the effect of lifestyle and demographic factors, including gender and smoking history, is insignificant (around 0.01) to the overall prediction.

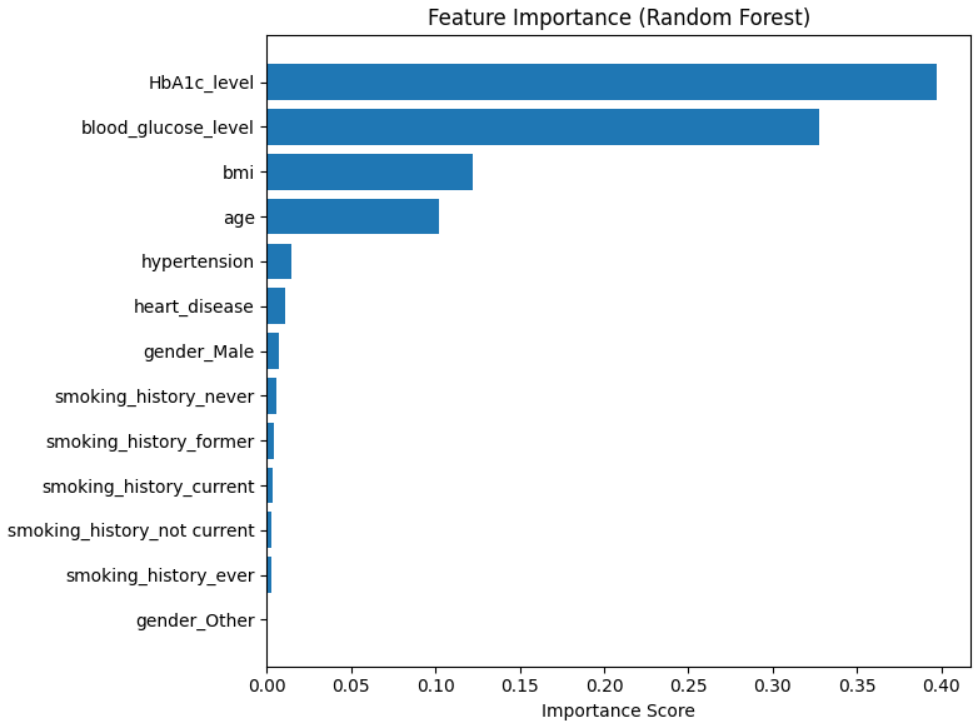


Figure 3: Feature importance analysis using the Random Forest model

This allocation of significance implies that the model is based on objective metabolic indicators as opposed to indirect or behavioural factors. In this case, it is especially critical due to the validation that the predictive framework is clinically based, and features that directly indicate glycemic control take precedence. Additionally, the high reliance of HbA1c and blood glucose is the reason why false negative cases are caused by borderline values of these features as seen in the error analysis. Therefore, feature importance not only determines important predictors but also gives a clue to the limitation of the model, when it is necessary to differentiate cases within the same range of biomarkers.

To gain a more insight into the structure of the underlying data, Principal Component Analysis (PCA) was used as a feature extraction method. The PCA feature loadings indicate the contribution of each variable to the various principal components and aid in revealing the structure of the dataset (Table 3). Age (0.56), BMI (0.43) and hypertension (0.39) are the key variables that impact the first principal component PC1, which implies that it reflects a general physiological and metabolic risk dimension. Conversely, HbA1c and blood glucose level have proportions more significant to PC3 (-0.36 and -0.37), indicating that glycemic variation is a distinct and significant dimension of the data. PC2 correlates mostly with smoking behavior, and a significant impact is made by smoking_history_never (0.77) and a moderate impact is made by smoking history current (−0.34) which denote lifestyle-related trends. Additional interactions between features are captured by other components (PC3 and PC5) which contain contributions by smoking categories and moderate effects of BMI and age. These loadings demonstrate that the data is spread among a variety of different dimensions that reflect physiological, biochemical, and behavioral characteristics, which proves that diabetes prediction is based on the effect of a combination of different variables, not one of them.

Table 3: PCA loadings showing feature contributions to the first five principal components

Feature	PC1	PC2	PC3	PC4	PC5
Age	0.56	0.08	0.12	0.06	0.22
BMI	0.43	0.15	−0.02	0.14	0.33
HbA1c_level	0.25	0.05	−0.36	0.13	−0.48
Blood_glucose_level	0.26	0.05	−0.37	0.13	−0.44
Hypertension	0.39	0.1	−0.04	−0.03	0.11
Heart_disease	0.34	−0.11	0.06	−0.15	−0.13
Smoking_history_current	0.02	−0.34	−0.67	−0.01	0.49
Smoking_history_never	−0.02	**0.77**	−0.04	−0.12	−0.04

The results of the Principal Component Analysis (PCA) explained variance shows that the variance in the dataset is shared between multiple components instead of being concentrated on one principal direction. Only 13.92% of the total variance is explained by the first principal component (PC1) and 10.38 and 8.49 by PC2 and PC3 respectively and the sum of the variance is about 49.40% after the first five components. This relatively slow accumulation of variance suggests that the dataset is not highly redundant or linearly dependent on features but is more of a multi-dimensional and complex structure. This finding is especially important in the framework of this study because it showcases the fact that prediction of diabetes cannot be well-modeled by a few linear combinations of features. The comparatively small percentage variance of the leading components suggests that the important information is distributed in several dimensions, encompassing biochemical indicators and demographic or clinical characteristics. Therefore, the use of dimensionality reduction alone to classify would lead to the loss of important information particularly to borderline or ambiguous ones. Moreover, the distributed variance structure also supports the necessity of sophisticated machine learning models that can reflect non-linear relationships and interactions among features. It also describes the apparent overlap of diabetic and non-diabetic samples because the distinction between classes is not restricted to a single feature space that dominates the data. In this work, the hybrid and ensemble methods, along with the threshold optimization techniques, can be justified by these findings to take advantage of the multi-dimensional nature of the data and better detect the difficult cases, specifically, false negatives.

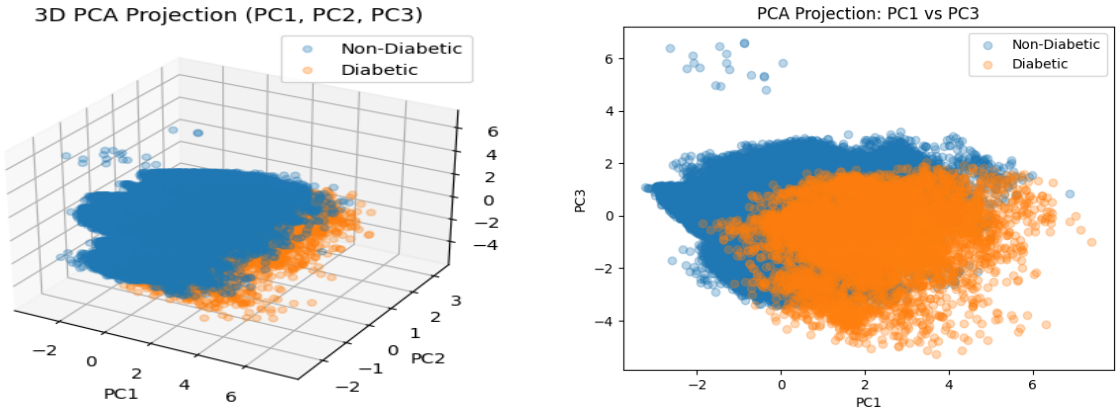


Figure 4: PCA-based visualization of feature space for diabetes classification

This is further observed by visualization of PCA projections, as shown in Figure 4. The 3D projection (PC1, PC2, PC3) and the 2D projection (PC1 vs PC3) shows some overlap of diabetic and non-diabetic samples, with some extent of separation along certain axes. Such an overlap implies that the feature space cannot be linearly separable, especially the borderline cases. In this research, this observation is very crucial, because it determines the reason behind some cases of diabetes being diagnosed as non-diabetic (false negatives). It also motivates the application of enhanced ensemble and hybrid models, and

threshold tuning to extract fine-tuning patterns which cannot be distinguished using simple linear transformations. Altogether, the process of feature extraction emphasizes the fact that although there are important predictors, the problem of diabetes prediction is high-dimensional and complicated, and the models should be able to operate in the context of overlapping feature distributions and non-linear dependencies.

Applied Classifiers: To capture the multi-dimensionality of diabetes prediction, a wide variety of classifiers was applied such as linear, distance-based, tree-based, ensemble, boosting, and hybrid classifiers (Figure 5). These models were selected based on the characteristics identified in the data that was found on the exploratory analysis, i.e., the existence of non-linear relationships, overlapping features distributions, and the imbalance of classes. More to the point, the analysis of these models was tuned to the clinical goal of reducing false negativity because false diagnoses are dangerous in the context of clinical practice.

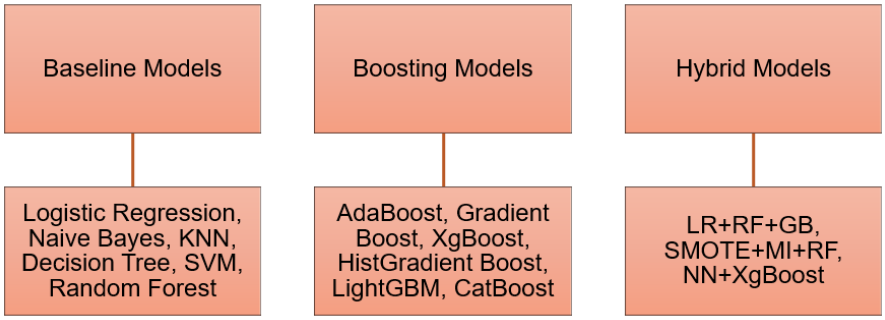


Figure 5: Categorization of machine learning models used in the study

Logistic Regression was used as a baseline probabilistic model to establish a linear decision boundary over the feature space [25]. The model estimates the probability of diabetes as a sigmoid transformation of a linear combination of features:

$$\hat{p}(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

While its simplicity enables interpretability, its linear nature limits its ability to capture complex interactions among features such as HbA1c, blood glucose, and BMI. However, in this work, Logistic Regression played a critical role beyond baseline comparison. Its probabilistic outputs were leveraged for threshold tuning, where the standard classification threshold of 0.5 was adjusted to a lower value ($\tau = 0.075$) to prioritize recall:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{p} \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

This modification significantly increased the capability to diagnose diabetes cases, by directly addressing the issue of false negatives, at a cost of false positives. K-Nearest Neighbors (KNN) was used to explore distance-based learning, and instances are classified using local neighborhood structures in the feature space. The similarity-based model implicitly refers to the similarity-based model which labels the classes and, therefore, is sensitive to the differences in the scales and density of the features [26]. KNN analysis in the present research demonstrated that observed local overlap of the diabetic and non-diabetic sample was observed especially in the areas where the HbA1c and glucose levels were felt within the borderline ranges. Its inability to generalize effectively in high-dimensional and imbalanced settings however hampered its performance. Tree-based models, like Decision Trees and Random Forest [27,28], played a role in non-linear relationship models and feature interaction. Decision trees sequentially divide the feature space into hierarchical decision rules using the best splitting criteria. But because of this overfitting, a more potent one, Random Forest, was used. Random Forest takes a combination of the decisions of multiple decision trees, which are built based on bootstrap sampling and randomness of features:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

where $h_t(x)$represents the prediction of the t-th tree. Random Forest in this work not only demonstrated high predictive power but also the ability to give interpretable scores of features, which showed the significant role of HbA1c and blood

glucose levels. This observation was instrumental in learning about model behavior and informing further investigation of false negative cases. The Support Vector Machine (SVM) plus radial basis function (RBF) kernel was employed to approximate the complicated non-linear decision boundaries by projecting the input features in a higher dimensional space [29]. This enabled the model to be able to separate overlapping classes. Yet, because of the class imbalance and the existence of borderline cases, SVM was very sensitive but with a relatively poor precision which implies a trade-off between the ability to detect diabetic cases and false alarms. Ensemble boosting techniques were added to further enhance the performance, such as Gradient Boosting [30], AdaBoost [31], HistGradientBoosting [32], XGBoost [33], LightGBM [34], and CatBoost [35]. These models are sequentially learned with a learning paradigm where the successive models aim at correcting the errors of earlier ones. The overall boosting formulation can be formulated as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $h_m(x)$represents the weak learner at iteration m, and $\gamma_m$is the learning rate. In this study, boosting algorithms demonstrated superior performance in terms of ROC-AUC and PR-AUC, as they effectively concentrated on difficult samples, including those near the decision boundary. Advanced implementations such as XGBoost, LightGBM, and CatBoost further enhanced efficiency and accuracy through optimized gradient-based learning and handling of categorical features. To leverage the complementary strengths of different models, hybrid architectures were proposed. One such approach involved combining predictions from Logistic Regression, Random Forest, and Gradient Boosting using a weighted ensemble:

$$\hat{p}_{hybrid} = \alpha p_{LR} + \beta p_{RF} + \gamma p_{GB}, \text{where } \alpha + \beta + \gamma = 1$$

This approach improved performance that comprised linear interpretability, ensemble stability and refinement by boosting. In addition, a combination of Tabular Neural Networks and XGBoost was created based on the neural network that would guarantee the learning deep feature representations and organization of gradient boosting patterns. To have more reliable predictions, predicted probabilities were calibrated with probability calibration techniques:

$$\hat{p}_{calibrated} = f(\hat{p}_{raw})$$

This was done to make sure that the probabilities that were predicted were more consistent with the actual risk levels, which is essential in threshold-based decision-making. A combination of calibration and threshold tuning allowed the model to obtain a more favorable compromise between the sensitivity and precision, which in the end minimized the number of false negatives, at the cost of a reasonable overall performance. In general, the architecture used in this paper represents a layer-based strategy with the initial linear models, then the non-linear and ensemble strategies, and finally some hybrid frameworks with the calibration and threshold optimization. This design is specifically adapted to clinical prediction tasks, and the aim is both high accuracy and consistent detection of high-risk patients and reduction of critical diagnostic mistakes. Three hybrid modeling approaches were created to overcome the shortcomings of single classifiers and provide increased clinical reliability in diabetes prediction. The methods are aimed at combining complementary learning methods, such as linear modeling, ensemble learning, boosting, feature selection, and deep learning, and clearly addressing class-imbalance, feature-redundancy, and false-negative errors.

Algorithm 1 shows the first hybrid model which is a weighted ensemble of Logistic Regression, Random Forest, and Gradient Boosting. It is a combination of linear probability estimation, non-linear ensemble learning and boosting-based error correction. This architecture takes advantage of the interpretability of the Logistic Regression, as well as the strength of the Random Forest, and the ability of the Gradient Boosting to refine predictions in sequence to bring about a balanced and dependable prediction.

Algorithm 1: Hybrid Blend (Logistic Regression + Random Forest + Gradient Boosting)

Input: Preprocessed dataset $D = \{X, y\}$

Output: Final prediction $\hat{y}$

Run data preparation module:

1. Transform categorical variables and normalize numerical features.

2. Split dataset into training and testing sets.

Run parallel model learning module:

3. Train Logistic Regression to learn global linear decision boundary and produce probability output p_{LR}.
4. Train Random Forest to capture non-linear feature interactions using multiple decision trees and produce p_{RF}.
5. Train Gradient Boosting sequentially, where each learner corrects previous residual errors, producing p_{GB}.

Run fusion module:

6. Aggregate model outputs using weighted blending:

$$p_{hybrid} = \alpha p_{LR} + \beta p_{RF} + \gamma p_{GB}$$

7. Adjust decision threshold $\tau$to prioritize recall for diabetic cases.

Run decision module:

8. Classify input sample based on aggregated probability and optimized threshold.

Return: Final prediction results combining linear, bagging, and boosting behaviors

The combination of SMOTE-resampling, Mutual Information feature selection, and Random Forest classification is used in algorithm 2. In this approach, the goal is to boost minority class representation by eliminating irrelevant or redundant characteristics before training the model. This model achieves data-level balancing combined with feature-level optimization, and ensemble learning to improve sensitivity to diabetic cases and generalize.

Algorithm 2: Hybrid (SMOTE + Mutual Information + Random Forest)

Input: Original dataset $D = \{X, y\}$

Output: Prediction $\hat{y}$

Run imbalance handling module:

1. Apply SMOTE to generate synthetic minority (diabetic) samples.

2. Form balanced dataset $D' = \{X', y'\}$.

Run feature extraction module:

3. Compute Mutual Information between each feature and target variable.
4. Rank features based on dependency strength.
5. Select top informative features to form reduced feature space X^*.

Run ensemble learning module:

6. Train Random Forest on X^*using bootstrapped sampling.
7. Each decision tree independently learns feature splits based on selected features.

Run aggregation module:

8. Combine predictions from all trees using probability averaging.

Run decision module:

9. Assign final class label based on aggregated probability.

Return: Prediction results with improved minority class detection and reduced feature redundancy

In Algorithm 3, a Tabular Neural Network is combined with XGBoost to obtain a combination of deep feature representations and structured decision boundaries. The neural network element is used to learn the complex non-linear correlations between the features, and XGBoost is used to optimize predictions with gradient boosting. The combination of these two enables the model to enjoy the advantages of both representation learning and error correction.

Algorithm 3: Hybrid (Neural Network + XGBoost) for Diabetes Prediction

Input: Preprocessed dataset $D = \{X, y\}$

Output: Final prediction $\hat{y}$

Run representation learning module:

1. Feed input features into a Tabular Neural Network.

2. Pass data through multiple dense layers with non-linear activations to learn feature representations.

3. Generate probability output p_{NN}capturing complex feature interactions.

Run structured learning module:

4. Train XGBoost model using gradient boosting over decision trees.
5. Sequentially minimize residual errors to produce probability output p_{XGB}.

Run hybrid fusion module:

6. Combine outputs from both models:

$$p_{hybrid} = \lambda p_{NN} + (1 - \lambda) p_{XGB}$$

Run calibration and decision module:

7. Calibrate probability scores to align with true risk distribution.

8. Apply optimized threshold $\tau$for final classification.

Return: Final prediction results leveraging deep feature learning and boosting-based refinement

4. RESULTS AND DISCUSSION

Different models are used to evaluate the diabetes prediction system which include baseline models, advanced boosting models and hybrid models. The evaluation cannot be simply measured as the cost of not detecting a true case of diabetes (false negative) is high in healthcare. Instead, it goes to such metrics as class-wise performance, confusion matrix, probability calibration and analysis of the types of errors made by the models. The results show the models do not work well in the presence of common problems in the data, including unbalanced classes, overlap of features, and non-linear relationships between clinical variables. In this section, we will attempt to identify the most plausible approaches to the credible and clinically safe detection of diabetes through a comparative analysis of behavior of the various models.

4.1 Baseline Models

In this section, the performance of the traditional machine learning models is assessed to form a baseline benchmark of diabetes prediction. They are linear, probabilistic, distance based and tree-based models which offer another perspective of the way the dataset can be modeled. The purpose of this analysis is to identify how far individual algorithms can reflect the underlying clinical patterns, particularly when there is a class imbalance, and overlapping feature spaces. Their capability to identify cases of diabetes is also considered because baseline models tend to be poor at balancing their precision and recall on the minority group. The lessons gained in this section can be used as a benchmark in evaluating the gains made by more developed and hybrid modeling strategies. The findings in Table 4 indicate apparent variations in the way the different baseline models are doing in this clinically structured and imbalanced diabetes data.

Table 4: Overall Performance of Baseline Models for Diabetes Prediction

Model	Accuracy	Log Loss	ROC–AUC	PR–AUC
Logistic Regression	0.8892	0.2449	0.9630	0.8212
Naive Bayes	0.4682	2.4513	0.9128	0.5088
KNN	0.8925	0.4994	0.9475	0.7435
Decision Tree	0.9466	1.9238	0.8567	0.5194
Random Forest	0.9568	0.1223	0.9672	0.8581
SVM (RBF)	0.8943	0.2227	0.9692	0.8220

Ensemble-based algorithms, especially Gradient Boosting and Random Forest, perform the best overall, with a high accuracy (>95%), a high ROC-AUC (>0.96), and a high PR-AUC value. This means that they could successfully identify non-linear associations and complicated interactions among important biomedical attributes including HbA1c and blood glucose concentrations and have a reasonable precision/recall balance. By contrast, the linear and margin-based models (Logistic Regression and SVM) have high ROC–AUC scores, indicating that they have good separability between classes, but poorer accuracy and PR–AUC values indicate that they are poor at addressing class imbalance and tuning the models to favor the minority (diabetic) class. The performance of KNN is moderate, since it depends on the local neighborhood structures, and thus it is delicate to local variations in feature scaling and density of data. Decision Tree is highly accurate with a relatively low ROC–AUC and PR–AUC which means it is overfitting, with limited generalization. Naive Bayes has the worst performance in all measures because its model of feature independence is not true in the case of clinical variables that are correlated. The findings indicate that the models that can represent more complex, non-linear patterns and interactions between features are more appropriate in this dataset, especially when there is a class imbalance. In Figure 6, Decision Tree and Random Forest confusion matrices give more insight into the behavior of these models beyond aggregate measures, especially in addressing class imbalance and clinical critical errors.

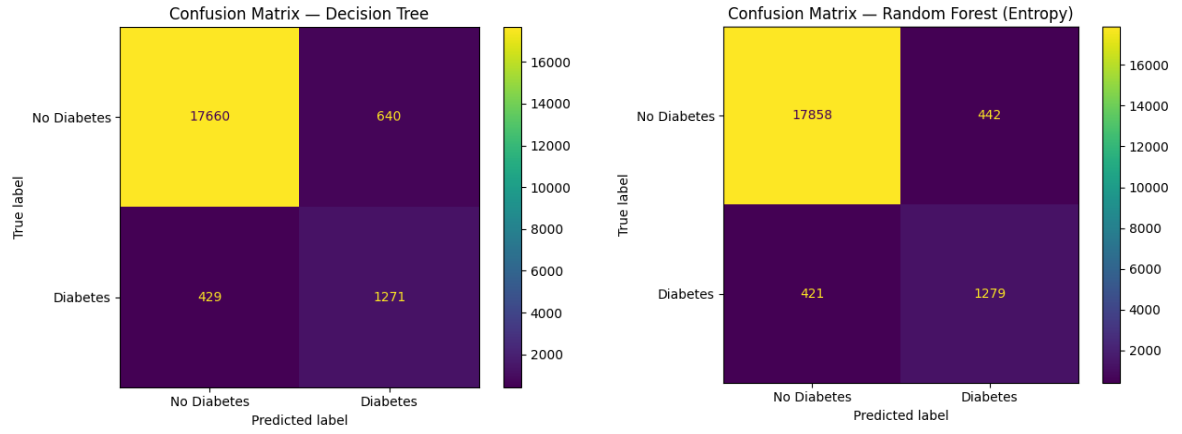


Figure 6: Confusion matrix comparison of Decision Tree and Random Forest models

In the case of the Decision Tree, a significant proportion of non-diabetic cases are correctly identified (true negatives = 17,660) but the proportion of false positives is also relatively high (640), meaning that the model is more likely to mislabel as diabetic a significant number of non-diabetic cases. In the case of the diabetic group, the model detects 1,271 true positives and fails to detect 429 cases (false negatives), which in clinical practice is a factor in clinical settings where undiagnosed diabetes can have severe outcomes. This implies that although the Decision Tree can generalize on the overall patterns, it does not cope well with accurate delimiting boundaries in the areas where the data overlap. The Random Forest model, on the contrary, demonstrates better results in both classes. It boosts the number of properly identified non-diabetic cases (true negatives = 17,858) and greatly decreases false positives to 442. When it comes to diabetics, it is true positive (1,279) and

reduces the false negatives to 421. This improvement is reflective of the ensemble nature of Random Forest wherein a set of decision trees is used to reduce overfitting and enhance generalization. Overall, the Random Forest has shown to have more balanced and reliable classification in comparison to the Decision Tree, particularly false positives and false negatives. This makes it more relevant in diabetes prediction where minimization of misclassification, especially false negatives may be of utmost importance to the early diagnosis and clinical intervention.

Table 5 reflects the internal decision-making of each model and its decision-making process, along with the way the decisions are translated into precision, recall, and F1-score of both classes. These values do not come out of thin air; they are a result of the way each algorithm models the data distribution, treats class imbalance and separates overlapping feature spaces.

Table 5: Class-wise Performance of Baseline Models

Model	Class	Precision	Recall	F1-score
Logistic Regression	No Diabetes	0.9888	0.8890	0.9362
	Diabetes	0.4272	0.8912	0.5776
Naive Bayes	No Diabetes	0.9978	0.4197	0.5909
	Diabetes	0.1368	0.9900	0.2404
KNN	No Diabetes	0.9863	0.8950	0.9384
	Diabetes	0.4339	0.8665	0.5782
Decision Tree	No Diabetes	0.9763	0.9650	0.9706
	Diabetes	0.6651	0.7476	0.7040
Random Forest	No Diabetes	0.9770	0.9758	0.9764
	Diabetes	0.7432	0.7524	0.7477
SVM (RBF)	No Diabetes	0.9917	0.8920	0.9392
	Diabetes	0.4415	0.9194	0.5966

In case of Logistic Regression and SVM, the models are trained on a global decision boundary over the feature space. The fact that they recall the diabetes class better (0.89 and 0.91) means that the learned boundary is moved towards the positive (diabetic) cases. But the change causes numerous non-diabetic samples to be mistakenly considered diabetic hence poor precision (approximately 0.4244). This is common when the linear or margin-based models are used to attempt to correct overlapping distributions in imbalanced data. Naive Bayes assumes that there is independence between features and that class probabilities are computed under this assumption. This assumption is invalid in this dataset in which correlations exist between such features as HbA1c and blood glucose. Consequently, the model has a large probability of the diabetes class, with very high recall (0.99) and very low precision (0.13), which represents high numbers of false positives. KNN is based on local structure of neighborhoods. Its average accuracy and recall of the diabetes category (around 0.43 and 0.87) imply that the samples of diabetes are usually enclosed in mixed neighborhoods, which results in inconsistent classification. This indicates the existence of overlapping regions of features in the dataset. Decision Tree, Random Forest on the other hand are hierarchical decision rules learned using feature splits. Such models can more easily isolate areas of non-diabetic and diabetic sample differences. Random Forest specifically, combine several decision paths, decreasing variance and enhancing generalization. This results in a more balanced precision-recall trade-off of the diabetes class (precision 0.74, recall 0.75), and higher F1-scores. These values indicate that model that can capture non-linear relationships and feature interactions result in more balanced and reliable performance in classes, whereas more simple models or assumption-based models are biased to either high recall or high precision and not both.

4.2 Advanced Boosting Models

In this section, advanced boosting algorithms are evaluated, which are aimed to improve the prediction based on the focus on misclassified and difficult samples. In contrast to baseline models, boosting methods use sequential learning to improve the decision boundaries and include more complex non-linear interactions among features like HbA1c and blood glucose levels. In this section, the assessment of these models does not only focus on their improvement of discrimination but also on probability estimation, especially in the presence of skewed clinical data. Moreover, their sensitivity and specificity are critically evaluated because the improvement of the models is usually enhanced, and their overall performance does not decrease markedly. This discussion gives an understanding of how to increase the effectiveness of boosting strategies as an intermediary between the traditional models and more advanced hybrid ones. Table 6 illustrates the behavior of advanced boosting models which always use more information to classify samples compared to baseline methods because of their sequential learning structure as well as their capacity to learn hard-to-classify samples. In contrast to traditional models, boosting algorithms can be used to repeatedly refine predictions, minimizing the residual errors to enable them to more closely represent the complex, non-linear relationships found in clinical data.

Table 6: Overall Performance of Boosting Models

Model	**Accuracy (%)**	**Log Loss**	**ROC–AUC (%)**	**PR–AUC (%)**
AdaBoost	88.83	0.3680	97.13	83.76
Gradient Boosting	96.60	0.1002	97.81	88.03
HistGradientBoosting	96.89	0.0863	97.79	87.98
XGBoost	97.04	0.0848	97.78	88.09
LightGBM	97.02	0.0834	97.69	87.59
CatBoost	97.18	0.0815	97.84	88.19

CatBoost has the best overall performance, highest accuracy (97.18%), least log loss (0.0815), and highest ROC -AUC (97.84%) and PR -AUC (88.19%). What this means is good classification power as well as well-calibrated and highly certain probability estimates. It is efficient in distinguishing between categorical features, which explains its effectiveness, and decreasing prediction bias. The performance of XGBoost and LightGBM is also rather competitive with the accuracy of approximately 97%, PR-AUC values near 0.88. Their optimization by gradient and efficient tree construction allows them to capture the interactions of features of interest, especially between important variables, including HbA1c and blood glucose levels. The fact that the log loss values are low also shows that these models are reliable in the production of probability. Gradient Boosting and HistGradientBoosting have a similar performance with a little lower ROC-AUC and PR-AUC scores. The performance of their implementation can indicate good error correction between iterations, albeit at a possible lower level of optimization than more sophisticated implementations. AdaBoost, in contrast, has a lower accuracy and a higher log loss, although it still has a high ROC–AUC. This is an indication that although it can separate classes reasonably, it does not make stable and well-calibrated predictions in this dataset. In sum, the findings validate that boosting models are effective to this problem due to their progressive specialization on hard examples and their superior discrimination and probability estimation, which suits well imbalanced, and clinically complex data. CatBoost and XGBoost confusion matrices in Figure 7 offer a more in-depth perspective of how enhanced boosting algorithms tackle the classification, especially the balancing of false positives and false negatives in a clinical-sensitive scenario.

Natural Resources for Human Health
eISSN: 2583-1194

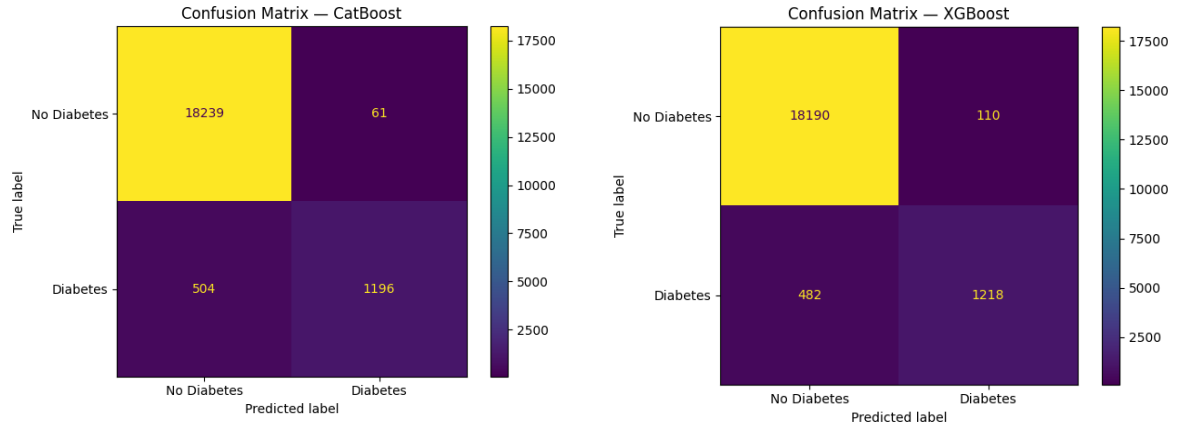


Figure 7: Confusion matrix comparison of CatBoost and XgBoost

In the case of CatBoost, the model has a high proportion of correctly identified non-diabetic cases (TN = 18, 239) with a low number of false positives (61), which implies that this algorithm has high specificity and is capable of preventing unnecessary false alarms. Nevertheless, it generates 504 false negatives, that is, there are still quite several instances of diabetic cases that it failed to detect, although it identified 1,196 diabetic cases correctly. This hints at the fact that although CatBoost is very conservative and accurate, it slightly underestimates the diabetic category. Comparatively, the XGBoost behaves slightly differently. It properly clusters 18,190 non-diabetic cases with 110 false positives, which means minor decrease in specificity as compared to CatBoost. Nevertheless, it enhances the ability to detect cases of diabetes by cutting false negatives to 482 and boosting true positives to 1,218. This implies improved sensitivity as well as specificity. On the whole, the two models perform well, with the only difference being that the trade-off CatBoost focuses more on precision and minimization of false positives, and XGBoost is more sensitive by minimizing false negatives. Clinically, XGBoost can be preferable, because it is common that minimizing false alarms is not as important as minimizing missed diabetic cases. Table 7 shows the distribution of the predictive performance of the advanced boosting models between the two classes, which is clearly more balanced and clinically meaningful than the baseline models.

Table 7: Class-wise Performance of Boosting Models

Model	Class	Precision	Recall	F1-score
AdaBoost	No Diabetes	0.9900	0.8869	0.9356
	Diabetes	0.4261	0.9035	0.5791
Gradient Boosting	No Diabetes	0.9758	0.9874	0.9816
	Diabetes	0.8442	0.7365	0.7867
HistGradientBoosting	No Diabetes	0.9743	0.9922	0.9832
	Diabetes	0.8958	0.7182	0.7973
XGBoost	No Diabetes	0.9742	0.9940	0.9840
	Diabetes	0.9172	0.7165	0.8045
LightGBM	No Diabetes	0.9728	0.9953	0.9839
	Diabetes	0.9326	0.7000	0.7997
CatBoost	No Diabetes	0.9731	0.9967	0.9847
	Diabetes	0.9515	0.7035	0.8089

In the case of the No Diabetes class, all boosting models have a high level of precision and recall (usually greater than 0.97), which means that such models are very high in detecting non-diabetic people. This is mostly because the ensemble boosting mechanism enhances the decision boundaries by refining them through the iterative process and lowering misclassification of the majority class. In the case of Diabetes class, the behavior is varied with respect to models but it is much more accurate than the base models. The high precision of advanced models, like XGBoost, LightGBM, and CatBoost (more than 0.91 and up to 0.95), is that, once it predicts diabetes, the model itself is confident and, therefore, generates fewer false positives. This however, is at the expense of reduced recall (approximately 0.70-0.72) meaning that not all diabetic cases are detected. This indicates a narrower decision boundary that is more conservative and values correctness over coverage. Conversely, AdaBoost has a different trend, showing very high recall (0.9035) but low precision (0.4261) which means that it is overly aggressive in classifying the diabetic as a diabetic and thus giving many false positives. HistGradientBoosting and Gradient Boosting offer a more moderate trade-off, with moderate precision and recall. All in all, these findings indicate that highly evolved boosting models transition to recall-light behavior to precision-based predictions, making them more reliable but with further potential to enhance their ability to detect all diabetic cases. This trade-off is a direct motivation to hybrid models and threshold optimization to increase sensitivity without losing precision.

4.3 Proposed Hybrid Models

In this section, the performance of the proposed hybrid models is provided which combine various learning paradigms to address the shortcomings which have been witnessed in the baseline and boosting methods. The models will be combined to achieve improved predictive power and improve the detection of diabetic cases incorporating complementary capabilities; linear interpretability, ensemble robustness, deep feature representation, and boosting-based refinement. The evaluation process is not only based on the overall performance measures but also balance, minimizing the false negative, and reliability of the estimates of the probability by classes. In addition, the hybrid architectures are explored in terms of their ability to handle high-dimensional and complex clinical data and overlapping distributions of features. This section will subsequently demonstrate why hybridization will lead to more clinically effective and strong diabetes prediction systems.

As shown in Table 8, the Hybrid (LR + RF + GB) model has a good overall performance, with high ROC-AUC (0.9751) and PR-AUC (0.8737). This suggests that a combination of linear, bagging and boosting techniques enables the model to explain both general trends, and complicated non-linear interactions. The log loss is also relatively low, indicating that the model is well-calibrated to make probability estimates.

Table 8: Performance of Hybrid Models for Diabetic Detection

Model	Accuracy	Log Loss	ROC–AUC	PR–AUC
Hybrid (LR + RF + GB)	95.46 %	0.1250	0.9751	0.8737
Hybrid (SMOTE + MI + RF)	95.73%	0.1178	0.9676	0.8580
Hybrid (NN + XGBoost)	98.45%	0.1091	0.9782	0.8830

The Hybrid (SMOTE + MI + RF) model is aimed at the balancing of data and optimization of features. The fact that it can better address the issue of class imbalance (95.73%) and has lower log loss (0.1178) is indicative of the advantages of dealing with class imbalance as well as eliminating less informative features. Nevertheless, its slightly worse ROC-AUC and PR-AUC than other hybrids imply that although feature selection is more efficient, it might restrict the ability of the model to reveal all the multifaceted patterns. The Hybrid (NN + XGBoost) model has the best overall performance, the highest accuracy (98.45%), ROC -AUC (0.9782), and PR -AUC (0.8830). This means that a combination of deep feature representation of neural networks and the structured learning of XGBoost allows to model more complicated relationships and work with borderline cases more effectively. The enhanced PR-AUC is significant especially because it indicates improved performance in the minority (diabetic) group. The findings overall indicate that hybridization is an effective tool in harnessing the strengths of the complementary models and is associated with better discrimination, probability estimation, and the capacity to deal with skewed clinical information. Table 9 will give a closer look at the balance of the proposed hybrid models between the two classes, which exhibit evident improvements compared to each of the baseline and boosting models, especially in the minority (diabetes) class.

Table 9: Class-wise Performance of Hybrid Models

Model	Class	Precision	Recall	F1-score
Hybrid (LR + RF + GB)	No Diabetes	0.9802	0.9701	0.9751
	Diabetes	0.7099	0.7888	0.7473
Hybrid (SMOTE + MI + RF)	No Diabetes	0.9770	0.9763	0.9766
	Diabetes	0.7466	0.7524	0.7495
Hybrid (NN + XGBoost)	No Diabetes	0.9787	0.9826	0.9806
	Diabetes	0.8039	0.7694	0.7863

In the case of the No Diabetes class, the precision and recall of all of the hybrid models (c. 0.97 -0.98) are continuously high, meaning that the combination of various strategies in learning does not lead to a negative impact on the capability of the model to recognize non-diabetic individuals. This stability indicates the strength of hybrid architectures in maintaining majority class performance with the enhancement of the minority class detection. In the case of Diabetes class, appreciable improvement is noted over previous models. The Hybrid (LR + RF + GB) model provides a reasonable compromise between recall (0.7888) and precision (0.7099), i.e. it can identify larger share of diabetic cases at the expense of false positives. The Hybrid (SMOTE + MI + RF) model even enhances this balance, with almost identical precision (0.7466) and recall (0.7524), and illustrates the power of balancing data and feature selection to minimize both false positives and false negatives. The Hybrid (NN + XGBoost) model has the best overall performance in the diabetes classification, as well as the highest precision (0.8039) and a good recall (0.7694), which leads to the highest F1-score (0.7863). This implies that the model is not just precise in detecting the cases of diabetes, but predictable and is advantageous in the combination of deep feature learning and boosting-based refinement. The findings indicate that hybrid models can greatly enhance the trade-off between the precision and the recall of the diabetes class and overcome the drawbacks identified in the baseline and boosting models. Such a balanced performance is especially crucial in clinical environments, where both proper detection and reliability are crucial. The Hybrid (Neural Network + XGBoost) model confusion matrix presents a clear insight into the balanced and clinically viable performance of this model (Figure 9).

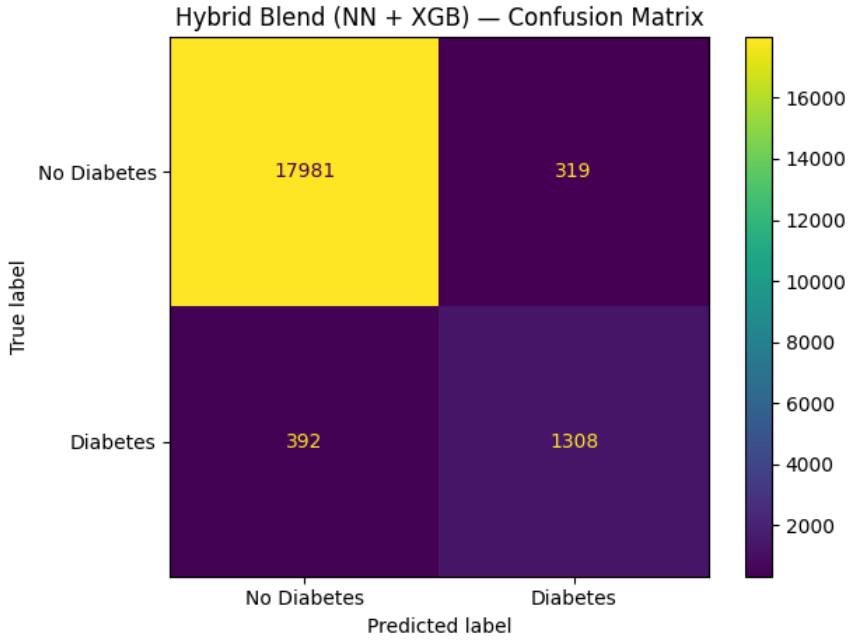


Figure 9: Confusion matrix of Hybrid Model (Neural Network + XgBoost)

The model has a significantly large number of non-diabetic cases (TN = 17,981) classified correctly with a small number of false positives (FP = 319), which is indicative of very high specificity and allows to control the unwanted false alarms. More importantly, the model manages to detect a large proportion of cases of diabetics (TP = 1,308), and minimize false negatives to 392. This is a very important reduction in false negatives, as it shows that this is a much smaller number of diabetic patients being missed than with baseline and individual boosting models. The fact that the hybrid architecture is better able to detect a higher percentage of true diabetic cases is indicative of the favorable properties of the hybrid architecture in capturing complex and overlapping feature patterns. The trade-off between false negativity and false positivity points to the effectiveness of the combination of deep learning and boosting methods. The neural network part acquires the sophisticated feature representations and XGBoost narrows down the decision boundaries by means of error correction. Consequently, the model can establish a more balanced trade-off of sensitivity to specificity. Overall, this confusion matrix supports the idea that the hybrid model does not only improve overall predictive performance but also clinical reliability since it will minimize critical misclassification errors, especially false negatives.

Probability calibration is employed to bring the predicted probabilities of a model into consistency with the true probability of outcomes, so that the confidence scores generated by the model are meaningful and interpretable. This is especially crucial in clinical usage where decisions are typically made based on risk thresholds instead of raw class names [36].

Table 10 illustrates that calibration enhances the accuracy of predicted probabilities to detect diabetes in the provided dataset, which consists of both clinical and biochemical variables.

Table 10: Calibration Performance (Before vs After Calibration)

Model	Setting	Accuracy	Log Loss	Brier Score	ROC–AUC	PR–AUC
Tabular NN	Uncalibrated	0.9366	0.1521	0.0455	0.9720	0.8518
Tabular NN	Calibrated	0.9644	0.1014	0.0283	0.9720	0.8518
XGBoost	Uncalibrated	0.9697	0.0867	0.0248	0.9780	0.8820
XGBoost	Calibrated	0.9706	0.0921	0.0239	0.9780	0.8820
Hybrid (NN + XGB)	Uncalibrated	0.9667	0.1133	0.0312	0.9774	0.8826
Hybrid (NN + XGB)	Calibrated	0.9708	0.0876	0.0236	0.9774	0.8826

Because this dataset has overlapping distributions of features, particularly between borderline diabetic and non-diabetic people, uncalibrated models can make overconfident or underconfident predictions. The largest change can be seen in Tabular Neural Network, where the log loss decreased by 0.1521 to 0.1014 and Brier score decreased by 0.0455 to 0.0283, which means that predicted probabilities are more consistent with actual probability of diabetes. This implies that patients who have higher predicted risk scores based on HbA1c and blood glucose levels can better be classified as diabetic. In the case of XGBoost, as well as Hybrid model which involves Neural Network with XGBoost, calibration also narrows down the estimates of probabilities, especially on patients with marginal clinical values. Such cases are prevalent in this dataset, and the HbA1c and blood glucose levels are close to diagnostic limits, which is necessary to estimate the probability accurately. Calibration, in general, improves the predictive interpretation and reliability of predictions by making the predicted risk scores more reflective of actual diabetic conditions. This not only favors better threshold-based decision making, but also leads to better identification of high-risk patients, and less uncertainty in clinically ambiguous cases.

The reliability diagrams in Figure 10 explain the extent to which the predicted probabilities are related to the actual likelihood of diabetes in each model, which gives a better insight into the calibration with reference to this dataset.

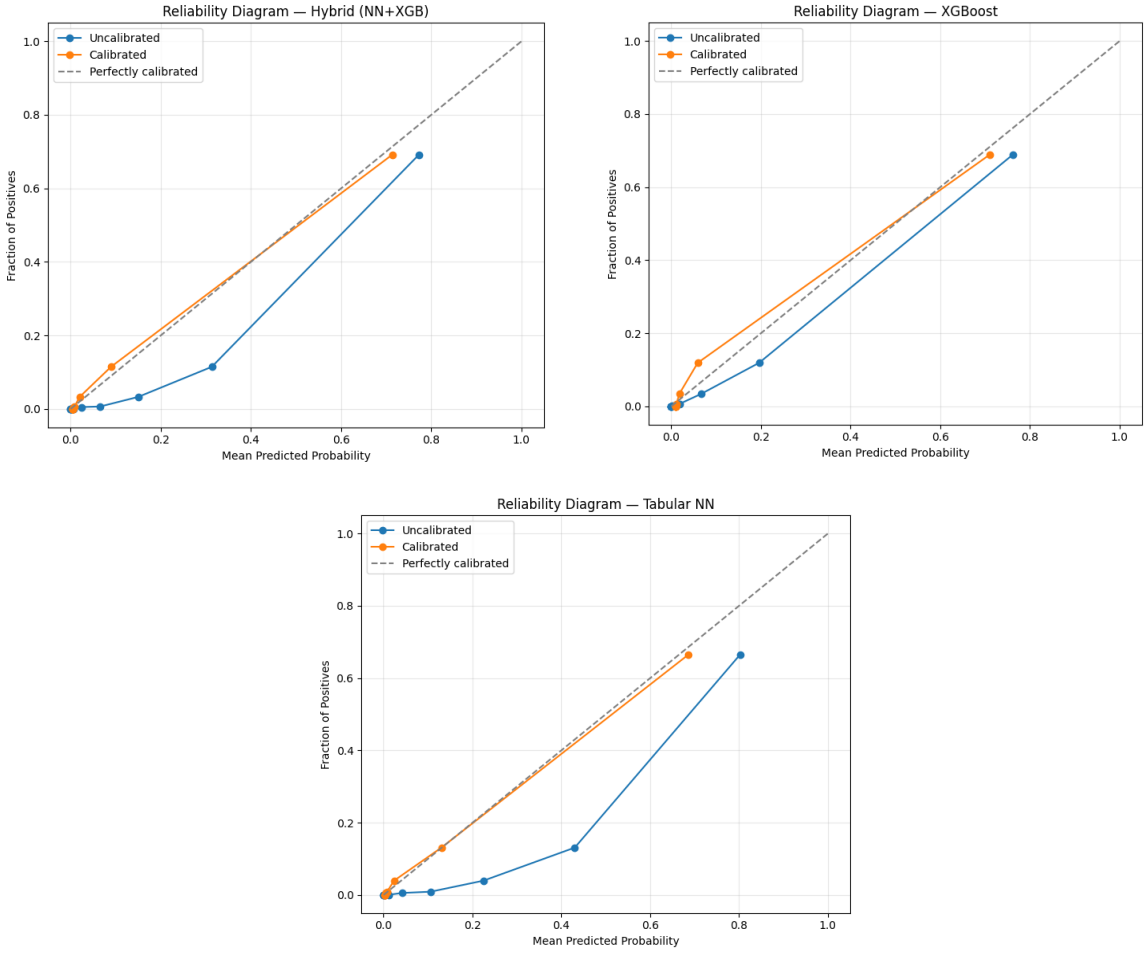


Figure 10: Reliability diagrams illustrating probability calibration for different models

The uncalibrated curve in the case of the Hybrid (NN + XGBoost) model does not follow the diagonal reference line accurately and especially in the middle range of probabilities, this means that the model will not be able to accurately predict the actual risk of diabetes. Upon calibration, the curve fits a lot nearer to the diagonal, which indicates that the predicted probabilities are now more accurately representative of the real percentage of diabetic cases. This is particularly noteworthy within patients who score moderately on risk where precise estimation of probability is paramount to clinical decision making. The same trend is followed with XGBoost, where the uncalibrated model predicts less risk in lower probability and a little more in higher ranges. Calibration helps correct this deviation which makes the closer alignment to the ideal diagonal line. It means that it is more reliable in giving risk scores to patients, especially those who are close to diagnostic thresholds. The Tabular Neural Network has the greatest improvement as the uncalibrated predictions are much far off the ideal line indicating weak probability estimation. The curve is nearly parallel to the diagonal after calibration, meaning that there is a significant change in risk predicted. This is especially true in this dataset where feature distributions may overlap, causing uncertainty in prediction. Overall, these diagrams verify that calibration increases reliability of probability estimates in all models. This guarantees that the risk scores predicted are more reliable in the context of diabetes detection, particularly in borderline cases, and facilitates a more reliable threshold-based decision as well as uncertainty in the identification of high-risk patients. Table 11 measures the effects of calibration on the performance of the model, showing that it improves the probability reliability and its influence on the detection of diabetes. The Tabular Neural Network has the largest improvement in accuracy (2.78% better) and log loss and Brier score are significantly reduced. This suggests that calibration is a significant boost to the capacity of the model to give meaningful probability scores, especially when it comes to falsely identifying diabetic patients whose clinical appearance can be on the borderline of diagnostic thresholds.

Table 11: Improvement After Calibration

Model	Δ Accuracy	Δ Log Loss	Δ Brier Score
Tabular NN	+0.0278	−0.0507	−0.0172
XGBoost	+0.0010	+0.0053	−0.0009
Hybrid (NN + XGB)	+0.0041	−0.0257	−0.0076

In the case of XGBoost, the modifications are not significant, and there is an insignificant improvement in the accuracy and a minor decrease in the Brier score, although log loss increases slightly. This implies that XGBoost was already well calibrated, and hence, calibration benefits are minimal. Nevertheless, the minimal increase in the Brier score indicates even more conformity between the forecasted probabilities and the real diabetic results. The Hybrid (NN + XGBoost) model is steadily improving, and its accuracy is higher, and both log loss and Brier score are visibly decreasing. This means that calibration further enhances the already robust hybrid predictions, rendering them more accurate in estimating diabetes risk.

The SHAP summary plot in Figure 11 gives a detailed understanding of the impact of various features on the way the model predicts diabetes and their significance as well as direction of influence.

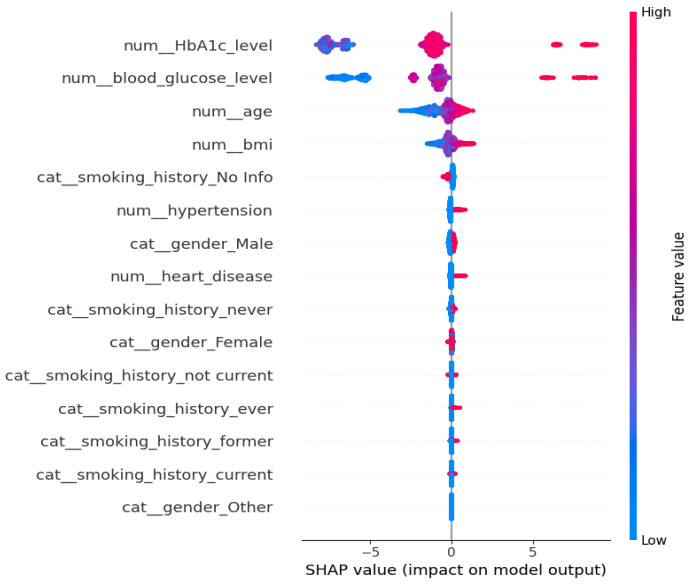


Figure 11: SHAP summary plot illustrating the impact of features on diabetes prediction

The strongest of them are HbA1c level and the level of blood glucose, which exhibit the biggest dispersion of SHAP values. These features have higher values (red points) that strongly push the prediction towards the diabetic class whereas lower values (blue points) push it to the non-diabetic one. This is more than coincidental with regards to clinical knowledge because high levels of HbA1c and glucose are among the top diagnostic signs of diabetes. The contribution of age and BMI is not as strong, but it is still significant. The BMI and advanced age values are likely to be associated with increased chances of having diabetes, which are positive SHAP values, meaning that they are not diagnostic values but rather a significant risk factor. The positive contributions of features like hypertension, heart disease are smaller yet still significant, indicating that comorbid conditions have a slight yet significant positive impact on the risk of diabetes. Conversely, categorical variables, such as gender and smoking history, have little effect and the SHAP values are close to zero, which means that they have a weak effect on the model predictions in this dataset. One notable finding is that SHAP values of key features are widely disseminated, indicating patient variability. Cases whose features values are similar do not contribute to predictions with the same degree meaning there is an overlapping feature distribution, particularly in borderline cases. In general, the plot substantiates that the model is mainly based on clinically relevant biochemical features, but the secondary factors include

demographic and lifestyle factors. This not only increases the predictive performance, but the model is also more interpretable and hence suitable in clinical decision support.

Table 12 shows the best false negatives that are nearest to the decision threshold, and this provides valuable information on the borderline misclassifications when predicting diabetes. The cases in the table are all true diabetic cases which were falsely classified as non-diabetic and the predicted probabilities were all very near to the classification cutoff of 0.5. The probabilities between 0.47 and 0.50 show that the model was unsure and these cases are close to the boundary of decision.

Table 12: Top 20 False Negative Cases Near Decision Threshold

Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c	Glucose	True	Pred	P(Diabetes)
Female	76	1	0	Never	32.57	6.2	160	1	0	0.4990
	42	0	0	Never	56.36	6.1	126	1	0	0.4989
Male	53	1	0	No Info	49.22	6.1	160	1	0	0.4898
Female	72	1	0	Former	31.15	6.1	130	1	0	0.4895
Male	70	1	0	Former	29.29	6.5	130	1	0	0.4857
Female	68	0	0	Former	46.13	5.7	155	1	0	0.4834
	77	0	0	Former	41.88	6.0	130	1	0	0.4800
Male	79	0	0	Former	30.67	6.6	130	1	0	0.4756
	60	1	0	Never	26.92	6.6	130	1	0	0.4741
Female	77	1	0	Never	30.12	5.7	130	1	0	0.4740

The notable finding is that most of the individuals in this table have moderately high HbA1c levels (between 5.7 and 6.6) and blood glucose levels ranging between 126 and 160 that are close to clinical diagnostic levels of diabetes. It can be implied that these cases are borderline physiological conditions, in which the difference between diabetic as well as non-diabetic is less obvious. Moreover, not only do most of them have high BMI scores and are older, but these are also known risk factors, and the model does not have enough certainty to categorize them as diabetic. These are also clinically relevant high-risk patients as the presence of comorbidities including hypertension in some of the cases also shows. Nevertheless, the model generates probabilities almost at the threshold, because of overlapping feature distributions in the dataset, and misclassification. In general, this table indicates that false negatives in this population are not caused by extreme errors but by decision boundary sensitivity causing slight changes in probability or threshold choice to appropriately label such patients. This supports the significance of threshold optimization and probability calibration in enhancing the detection of such borderline diabetic cases. Table 13 shows the most pressing false negative instances in which the model makes a very high confidence mistake of classifying diabetic people as such with very high confidence as shown by extremely low predicted probabilities (around 0.005-0.006). These cases are well below the decision level unlike borderline cases which indicates that the model is very confident that the patients are non-diabetic, though they are diabetic.

Table 13: Top 20 False Negative Cases with High Confidence Misclassification

Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c	Glucose	True	Pred	P(Diabetes)
Male	4	0	0	No Info	17.23	5.8	200	1	0	0.0053
Female	13	0	0	No Info	17.34	6.2	140	1	0	0.0057
	6	0	0	No Info	27.32	6.6	130	1	0	0.0058

	32	0	0	Current	13.73	6.0	160	1	0	0.0059
Male	26	0	0	No Info	27.32	5.7	130	1	0	0.0060
	16	0	0	Never	23.30	5.7	160	1	0	0.0061
Female	15	0	0	No Info	19.88	6.5	145	1	0	0.0064
	27	0	0	No Info	27.32	5.7	145	1	0	0.0064
	17	0	0	Ever	22.90	6.1	130	1	0	0.0065
	23	0	0	Never	22.14	6.2	130	1	0	0.0066

One of the interesting trends that can be observed in this table is that most of them are young (as young as 4, 6 and 13 years old) and demonstrate low or normal BMI levels, which greatly biases the model towards non-diabetic outcomes. This would indicate that the model has acquired a bias that a younger age and lower BMI are associated with lower risks of diabetes, despite the biochemical indicators indicating otherwise. Notably, a number of cases continue to have high HbA1c levels (as high as 6.6) and high blood glucose levels (as high as 200) which are medically important signs of diabetes. These cues however are overwhelmed with demographic attributes resulting to confident misclassification. The other observation is that in most cases, there are no comorbidities like hypertension and heart disease, which adds to the confidence of the model that it will predict a non-diabetic outcome. Also, the high cases of No Info in smoking history suggest the absence or less informative behavioural information, which can decrease the sensitivity of the model to risk patterns. Overall, this table underscores a more severe shortcoming of the model where the feature dominance and acquired biases result in systematic errors, especially to less typical patients, including young people with abnormal glucose levels. Such false negatives with high confidence are clinically dangerous, because they are not likely to be amenable to the effects of threshold adjustment. This highlights the necessity of enhanced feature weighting, enhanced representation learning, and integration of domain knowledge to make sure that essential biochemical indicators are not underemphasized in such situations.

5. CONCLUSION

This paper introduced a hybrid machine learning model for predicting diabetes with a heavy bias towards false negative reduction and clinical reliability. Several baseline, boosting, and hybrid models were tested using a structured dataset with key clinical features. The research shows that the boosting models are more effective than the traditional strategies. The hybrid models and especially the Neural Network-XGBoost model have presented the optimal balance between predictive performance and the capability to predict diabetes correctly. In addition, the probability calibration and threshold tuning methods have increased the quality of the predictions and have enabled the detection of the high at-risk patients of diabetes. Further, realistic diagnostic challenges of missed cases and a distribution overlap of features was also shown in false negative analysis at the case-by-case level. The performance is good, but the study has limitations. Although the data set that is utilized in this research is gigantic, the data is merely in tabular form, which does not capture the change over time or the medical history of specific patients. Other characteristics like lifestyle information and genetic predisposition are also not recorded. Furthermore, because a lot of features are shared by both cases of diabetes and non-diabetes, there will always be cases which the model may fail to predict, irrespective of the model. In future studies, longitudinal data can be used, and multiple streams of data can be incorporated, including electronic health records and sensor-based data, to enhance the predictive capabilities. More sophisticated deep learning and attention-based models can also be considered to learn more complex feature relationship. Moreover, it can be supplemented with medical knowledge and explainable AI techniques to enhance the level of trust and acceptance of the models in practice. The methodology can also be extended to real-time clinical decision-making models, and it can be tested on other population-based datasets to make sure that it can be generalised in clinical settings.

DECLARATIONS AND STATEMENT

Funding: Not applicable

Competing Interests: The authors confirm that no conflicts of interest exist.

Ethics Approval: Not Applicable

Availability of data and material: The dataset used in this study is publicly available and can be accessed at the following link: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Consent to Participate: Not Applicable

Consent to publish: Not Applicable

REFERENCES

[1] Antar, S. A., Ashour, N. A., Sharaky, M., Khattab, M., Ashour, N. A., Zaid, R. T., ... & Al-Karmalawy, A. A. (2023). Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments. *Biomedicine & Pharmacotherapy*, *168*, 115734.

[2] Hippisley-Cox, J., & Coupland, C. (2016). Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care. *bmj*, *352*.

[3] Tomic, D., Shaw, J. E., & Magliano, D. J. (2022). The burden and risks of emerging complications of diabetes mellitus. *Nature Reviews Endocrinology*, *18*(9), 525-539.

[4] Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., ... & Robinson, K. A. (2023). Diagnostic errors in the emergency department: a systematic review.

[5] Lennartsson, A. K., Jonsdottir, I. H., Jansson, P. A., & Dahlman, A. S. (2026). Insulin sensitivity in individuals with burnout is associated with physical activity level–a study using oral glucose tolerance test. *Frontiers in Psychiatry*, *17*, 1664680.

[6] Swaby, R., Scudder, C., Randell, T., Marcovecchio, M. L., Gillespie, K., Liu, Y. F., ... & Besser, R. E. (2026). A study to determine a capillary alternative to the gold standard oral glucose tolerance test-Protocol. *Wellcome Open Research*, *9*, 601.

[7] Mukherjee, S., Ray, S. K., Jadhav, A. A., & Wakode, S. L. (2024). Multi-level analysis of HbA1c in diagnosis and prognosis of diabetic patients. *Current Diabetes Reviews*, *20*(7), 85-92.

[8] Zhou, J., Park, S., Dong, S., Tang, X., & Wei, X. (2025). Artificial intelligence-driven transformative applications in disease diagnosis technology. *Medical Review*, *5*(5), 353-377.

[9] Khalifa, M., & Albadawy, M. (2024). Artificial intelligence for clinical prediction: exploring key domains and essential functions. *Computer Methods and Programs in Biomedicine Update*, *5*, 100148.

[10] Asif, R., Upadhyay, D., Zaman, M., & Sampalli, S. (2025). Enhancing diabetes risk prediction: A comparative evaluation of bagging, boosting, and ensemble classifiers with SMOTE oversampling. *Informatics in Medicine Unlocked*, *57*, 101661.

[11] Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, *2*(2), 22.

[12] Health–Europe, T. L. R. (2023). Misdiagnosis of type 1 and type 2 diabetes in adults. *The Lancet Regional Health-Europe*, *29*, 100661.

[13] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), 1-10.

[14] Mao, Y., Zhu, Z., Pan, S., Lin, W., Liang, J., Huang, H., ... & Chen, G. (2023). Value of machine learning algorithms for predicting diabetes risk: a subset analysis from a real-world retrospective cohort study. *Journal of Diabetes Investigation*, *14*(2), 309-320.

[15] Shin, J., Kim, J., Lee, C., Yoon, J. Y., Kim, S., Song, S., & Kim, H. S. (2022). Development of various diabetes prediction models using machine learning techniques. *Diabetes & Metabolism Journal*, *46*(4), 650-657.

[16] Rahman, M. A., Abdulrazak, L. F., Ali, M. M., Mahmud, I., Ahmed, K., & Bui, F. M. (2023). Machine learning-based approach for predicting diabetes employing socio-demographic characteristics. *Algorithms*, *16*(11), 503.

[17] Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, *83*(8), 24153-24185.

[18] Ahmed, A., Khan, J., Arsalan, M., Ahmed, K., Shahat, A. A., Alhalmi, A., & Naaz, S. (2024, December). Machine learning algorithm-based prediction of diabetes among female population using pima dataset. In *Healthcare* (Vol. 13, No. 1, p. 37). MDPI.

[19] Albadri, R. F., Awad, S. M., Hameed, A. S., Mandeel, T. H., & Jabbar, R. A. (2024). A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm. *Mathematical Modelling of Engineering Problems*, *11*(8).

[20] Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, *2*(1), 9-16.

[21] Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. *Artificial intelligence in medicine*, *164*, 103132.

[22] Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine learning and deep learning approaches for predicting diabetes progression: A comparative analysis. *Electronics*, *14*(13), 2583.

[23] Lee, H., Park, M. B., & Won, Y. J. (2025). AI machine learning–based diabetes prediction in older adults in South korea: cross-sectional analysis. *JMIR Formative Research*, *9*(1), e57874.

[24] El-Sayyid El-Bashbishy, A., & El-Bakry, H. M. (2026). Pediatric diabetes prediction using machine learning. *Scientific Reports*, *16*(1), 1979.

[25] Mustafa, M. (2023). *Diabetes prediction dataset*. Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[26] Abdalrada, A. S., Neamah, A. F., & Murad, H. (2024). Predicting diabetes disease occurrence using logistic regression: An early detection approach. *Iraqi Journal For Computer Science and Mathematics*, *5*(1), 6.

[27] Permana, B. A. C., Ahmad, R., Bahtiar, H., Sudianto, A., & Gunawan, I. (2021, April). Classification of diabetes disease using decision tree algorithm (C4. 5). In *Journal of Physics: Conference Series* (Vol. 1869, No. 1, p. 012082). IOP Publishing.

[28] Noviyanti, C. N., & Alamsyah, A. (2024). Early detection of diabetes using random forest algorithm. *Journal of Information System Exploration and Research*, *2*(1).

[29] Harini, P., Priyanka, A. L., & Narayana, N. L. (2025, December). Hybrid SVM–Random Forest Framework for Early Detection and Risk Prediction of Diabetes Mellitus. In *2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* (pp. 1889-1894). IEEE.

[30] Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, *11*(9), 1714.

[31] Jasim, A. A., Hazim, L. R., Mohammedqasim, H., Mohammedqasem, R. A., Ata, O., & Salman, O. H. (2024). e-Diagnostic system for diabetes disease prediction on an IoMT environment-based hyper AdaBoost machine learning model. *The Journal of Supercomputing*, *80*(11), 15664-15689.

[32] Alqahtani, S. A. M. (2025). Comparative machine learning models for hypertension prediction in a cohort of patients with diabetes using routine clinical variables. *Archives of Endocrinology and Metabolism*, *69*(6), e250168.

[33] Prabha, A., Yadav, J., Rani, A., & Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Computers in Biology and Medicine*, *136*, 104664.

[34] Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, *11*(9), 1714.

[35] Kumar, P. S., Kumari, A., Mohapatra, S., Naik, B., Nayak, J., & Mishra, M. (2021, January). CatBoost ensemble approach for diabetes risk prediction at early stages. In *2021 1st Odisha international conference on electrical power engineering, communication and computing technology (ODICON)* (pp. 1-6). IEEE.

[36] Verma, P., Elango, S., & Singh, K. (2025). Diabetic retinopathy detection using a well-calibrated uncertainty aware convolutional neural network. *Multimedia Tools and Applications*, *84*(19), 20779-20801.