Journal of Computer Science

Research Article

Improving Diabetes Risk Prediction Using Ensemble Boosting and SMOTE-Based Class Balancing

Kittipol Wisaeng[1], Pankom Sriboonlue[1] and Benchalak Muangmeesri[2]

[1]Department of Technology and Business Information System Unit, MSU Research Laboratory of Blockchain and Artificial Intelligence for Interdisciplinary Innovation, Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand
[2]Department of Engineering Management, Suan Sunandha Rajabhat University, 1 U-Thong nok Road, Dusit, Bangkok 10300, Thailand

Article history
Received: 17-06-2025
Revised: 31-07-2025
Accepted: 04-08-2025

Corresponding Author:
Kittipol Wisaeng
Department of Technology and Business Information System Unit, MSU Research Laboratory of Blockchain and Artificial Intelligence for Interdisciplinary Innovation, Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand
Email: kittipol.w@acc.msu.ac.th

Abstract: Accurate diabetes prediction is vital for early intervention, optimized resource allocation, and minimizing long-term complications. This study presents a comparative evaluation of traditional and advanced machine learning models for diabetes classification using a structured clinical dataset. Seven baseline algorithms were assessed against five advanced ensemble methods: CatBoost, LightGBM, XGBoost, Voting Ensemble, and Stacking Ensemble. To improve algorithm learning, the Synthetic Minority Over-sampling Technique (SMOTE) and feature normalization were employed. The algorithm's effectiveness was carefully evaluated using accuracy, precision, recall, and the F1 score. Results show that advanced models substantially outperformed traditional ones, with CatBoost achieving the highest F1 score of 0.7625. Feature importance analysis identified glucose, BMI, and age as the most influential indicators, consistent with clinical evidence. These findings demonstrate the potential of ensemble learning and boosting strategies for building interpretable, scalable, and effective diagnostic support tools in healthcare settings.

Keywords: Diabetes Prediction, Ensemble Learning, Voting Classifier, SMOTE

Introduction

Diabetes mellitus is a persistent metabolic disorder characterized by elevated blood glucose levels, known as hyperglycemia. This occurs due to a complete lack of insulin (absolute insulin deficiency) or a reduced response to insulin (insulin resistance), both of which disrupt glucose regulation and lead to various health complications (Roglic, 2016). This prolonged elevation in blood glucose levels poses serious health risks, as it contributes to a host of debilitating complications, including diabetic retinopathy. These complications significantly impair organ function, diminish quality of life, and elevate both morbidity and mortality rates (DeFronzo *et al*., 2015). Early identification and prediction of diabetes and its associated complications are therefore crucial for enabling timely clinical interventions and optimizing long-term disease management strategies.

Recently, Machine Learning (ML) and Deep Learning (DL) methods have been proposed in medical predictive analytics, particularly for chronic diseases such as diabetes. These approaches excel at capturing non-linear, high-dimensional relationships within complex datasets, often outperforming traditional statistical models. Empirical studies have confirmed the efficacy of deep learning in detecting various diabetes-related outcomes (Thotad *et al*., 2023), cardiovascular complications (Longato *et al*., 2021), and diabetic nephropathy (Vidhya and Shanmugalakshmi, 2020). Moreover, integrating domain-specific constraints, data augmentation, and regularization techniques has enhanced model robustness and generalizability (Liang *et al*., 2021). Despite significant progress, incorporating structured domain knowledge, such as medical ontologies and knowledge graphs, into ML pipelines for predicting diabetic complications remains a relatively unexplored area. Only a few notable studies, such as Diao *et al*. (2021); Li *et al*. (2023), have attempted to combine knowledge graph embeddings with predictive modeling. These early findings suggest that knowledge-guided deep learning holds great promise for improving both predictive accuracy and interpretability, thereby contributing to the

development of explainable AI (XAI) in diabetes care. One persistent challenge in applying ML to healthcare is the black-box nature of many predictive algorithms, which limits their transparency and hinders their adoption in clinical settings. Particularly in high-stakes environments, such as diabetes management, clinician trust and model interpretability are crucial for real-world implementation. To address this, researchers have increasingly adopted explainable Machine Learning (XML) approaches that aim to maintain high predictive performance while providing interpretable insights. Examples include decision trees, random forests with feature importance analysis, gradient boosting machines enhanced with SHAP (SHapley Additive exPlanations) values, and attention-based neural networks capable of highlighting key input features.

In this paper, we compare traditional and advanced explainable ML models for predicting diabetes and its complications. Using a structured clinical dataset that includes physical examination metrics and medical history, we evaluate multiple models for accuracy, interpretability, and scalability. Our primary objective is to identify models that balance predictive accuracy and clinical transparency, making them suitable for real-world decision-support systems.

Related Work

Early research on diabetes prediction has primarily focused on applying traditional machine learning techniques to clinical datasets, typically derived from physical examinations and biochemical tests. Pioneering work by Lindström and Tuomilehto (2003) introduced multiple Logistic Regression (LR) models to develop a risk-scoring system for identifying individuals at risk of type 2 diabetes. Building on this foundation, (Tanaka *et al*., 2013) used a multistate Cox proportional hazards model to estimate the likelihood of macrovascular and microvascular complications in diabetic patients. In a noteworthy study, (Yu *et al*., 2010) effectively used an SVM to classify individuals into three categories: Those with diabetes, those at risk of developing diabetes (prediabetes), and those without the condition, across the diverse U.S. population. Meanwhile, (Priya *et al*., 2020) employed a Naïve Bayes (NB) classifier, utilizing clinical data from over 1,800 patients to accurately predict diabetes-related outcomes. For those interested in a comprehensive overview of how ML is applied in diabetes research and its associated complications, the detailed reviews in Kee *et al*. (2023) are highly recommended. Although these traditional models have demonstrated reasonable predictive capabilities, many are limited in their generalization due to their reliance on individual classifiers and their lack of robustness across diverse datasets. Researchers have explored ensemble learning techniques to enhance predictive performance and address concerns about overfitting. For example, (VijiyaKumar *et al*., 2019; Wang *et al*., 2020) used ensemble models for diabetes prediction tasks, reporting improvements in both accuracy and the algorithm's generalization. Jian *et al*. (2021) expanded their analysis by exploring various classifiers, each with its unique strengths. They carefully evaluated Logistic Regression (LR), which accurately models probabilities; SVM, recognized for its ability to handle complex datasets with significant precision; Decision Trees (DT), which offer intuitive and interpretable pathways for classification; Random Forest (RF), an ensemble algorithm that combines multiple DT for greater reliability; AdaBoost, which focuses on correcting the errors made by weak classifiers; and XGBoost, a powerful and efficient algorithm optimized for high performance. Together, these classifiers were utilized to predict the onset of seven distinct types of diabetic complications, highlighting the vast potential of machine learning in healthcare. Their findings confirmed that ensemble-based models significantly outperform standalone algorithms. Another notable contribution by Hasan *et al*. (2020) introduced a weighted fusion strategy to combine multiple classifiers, further boosting prediction reliability. Nonetheless, selecting the optimal combination of classifiers remains a computationally complex, NP-hard problem. Given the inherent limitations of conventional ML algorithms, such as susceptibility to overfitting and limited generalization, recent studies have shifted toward using DL methods for diabetes-related prediction tasks (Pal *et al*., 2022). Ayon *et al*. (2019) introduced a DNN trained on the Pima Indians Diabetes (PID) dataset, demonstrating improved diagnostic accuracy. Deep learning’s capacity to model nonlinear data distributions has also proven beneficial. In a compelling study, (Swapna *et al*., 2018) harnessed the power of CNN to meticulously extract intricate temporal features from Heart Rate Variability (HRV) data. This advanced technique enabled them to uncover patterns hidden within the data's fluctuations. Following this meticulous extraction process, they employed SVM to classify these features, ultimately providing a valuable tool for enhancing diabetes diagnosis. Deep learning has been widely applied to medical imaging to predict diabetes complications. Gargeya and Leng (2017) proposed an innovative multi-level deep fusion network specifically designed to identify retinal abnormalities in Optical Coherence Tomography (OCT) images. This advanced network demonstrated remarkable classification accuracy, even when patients showed only subtle signs of retinal damage, demonstrating its potential to detect critical issues in the very earliest stages of eye health deterioration. Likewise, (Gulshan *et al*., 2016) demonstrated the potential of DNNs for identifying diabetic retinopathy from fundus photographs. Despite the promising advances in deep learning, challenges such

as variability in medical data quality and insufficient incorporation of domain-specific knowledge, particularly structured diabetes-related information, continue to undermine the reliability and interpretability of these models (Zhu *et al*., 2021). To address these challenges, it is crucial to integrate external knowledge representations and to enhance model explainability. This will create a strong foundation for prediction systems that are both resilient and widely accepted in clinical settings. As the amount of healthcare data continues to grow rapidly, Knowledge Graph (KG) technology has been employed as a transformative tool in intelligent medical systems. It provides a dynamic, interconnected approach to leveraging information and insights across the healthcare landscape. Its ability to represent complex medical relationships with semantic clarity and logical consistency has been recognized across various applications, including disease risk assessment, treatment recommendation, and quality control in healthcare services (Zhang *et al*., 2022). Numerous prestigious institutions are embarking on groundbreaking initiatives to create intricate medical knowledge graphs. Notable examples include the Traditional Chinese Medicine knowledge graph developed at Shanghai Shuguang Hospital and the innovative knowledge architecture pioneered by IBM Watson Health in the United States. In the realm of diabetes care, these knowledge graphs are seamlessly integrated into predictive frameworks, significantly improving diagnostic accuracy (Cheng *et al*., 2023). One remarkable approach involves a cutting-edge model that synergistically merges knowledge extension mechanisms with CNN, paving the way for enhanced diabetes prediction and management. More broadly, recent studies have explored the integration of domain knowledge into DL models for various health-related predictions, including hypertension management (Xi *et al*., 2021), Parkinson's disease classification (Balaji *et al*., 2021), pediatric COVID-19 severity assessment (Gao *et al*., 2022), Alzheimer's disease progression tracking (Nian *et al*., 2022), and prediction of hospital mortality and readmission (Jiang *et al*., 2024). These works demonstrated that embedding structured medical knowledge helps compensate for data limitations and guides the learning process, improving model robustness and interpretability (Gandhi and Mishra, 2021). Despite this progress, research on leveraging knowledge graphs for predicting diabetic complications remains limited. Lu and Uddin (2022) developed a stacking-based prediction model for 30-day hospitalizations in diabetic patients, incorporating XAI to highlight feature contributions and improve transparency. Their comparative evaluation demonstrated superior performance, and LIME was employed to provide localized explanations for individual cases. Similarly, (Vishwarupe *et al*., 2022) utilized the ELI5 library in conjunction with LIME and SHAP to analyze clinical data and identify patterns of feature importance. Another study proposed a Bayesian-optimized explainable TabNet (BO-TabNet) model that leverages attention mechanisms alongside SHAP and LIME to provide both global and local explanations for diabetes classification. The model showcased outstanding performance in both the PIDD and the Early-Stage Diabetes Risk Prediction Dataset (ESDRPD). Its exceptional ability to accurately identify risk factors and predict outcomes highlighted its advanced capabilities in addressing complex health challenges. Additional works also emphasize model transparency. For instance, a community-based classifier using LIME visualization achieved 81% accuracy (Uysal, 2023), demonstrating how XAI can clarify the influence of each feature in a prediction. Uysal (2023) found that SVM and random forest models performed best for diabetes classification, with SHAP plots revealing that glucose, age, and BMI were the top predictors. Lee *et al*. (2024) applied SHAP to gradient boosting models and confirmed its effectiveness in highlighting both expected features (e.g., glucose, BMI, age) and underrecognized predictors (e.g., blood pressure, pregnancy count). These insights aligned with clinical knowledge and revealed new patterns relevant to personalized risk assessment. The study summarized advances in XAI for diabetes prediction, with a particular emphasis on model-agnostic tools such as LIME, SHAP, and ELI5. ELI5, which utilizes permutation importance to assess feature influence, further enhances interpretability in machine learning models. Sharia *et al*. (2025) introduced DeepNetX2, a DNN trained using features selected via Spearman's correlation. They applied LIME and SHAP for model interpretation across three datasets, achieving robust predictive and explanatory performance. Recently, a study utilized six machine learning algorithms: LR, DT, RF, SVM, KNN, and Gradient Boosting Machines (GBM). Additionally, it included a lesser-known but effective method called Multivariate Adaptive Regression Splines (MARS). To address the "black-box" nature of these algorithms, the study utilized Shapley values to visualize and interpret learned patterns. Color-based visualization highlighted key predictive attributes and reinforced the practical utility of XAI in real-world diabetes prediction tasks. Table 1 provides a comprehensive overview of the key features of this study.

Research Contributions

This research highlights the role of machine learning in predicting diabetes and enhancing clinical decision-making, demonstrating how data-driven methods can lead to more accurate and timely interventions in diabetes management.

Table 1: A comprehensive overview of the relevant research, highlighting both the significant contributions made and the inherent limitations associated with each study

Study	Techniques Used	Limitations
Lindstrom and Tuomilehto (2003)	Logistic Regression (LR) for diabetes risk scoring	Limited generalization; relies on clinical risk factors.
Tanaka *et al.* (2013)	Multi-state Cox proportional hazards for complication risk	Complex model; limited to macro/microvascular events.
Yu *et al*. (2010)	SVM for classification	Dataset-specific, low adaptability to diverse Populations.
Priya *et al*. (2020)	NB for outcome prediction	Simple classifier; limited to structured clinical data.
Hasan *et al*. (2020)	Weighted fusion ensemble of multiple classifiers	Model selection is NP-hard; it may increase \computation cost.
Ayon *et al*. (2019)	DNN on PID dataset	Data-limited; lacks interpretability
Swapna *et al.* (2018)	CNN for HRV feature extraction + SVM	Requires high-quality temporal HRV data; small Dataset.
Gargeya and Leng (2017)	Multi-level deep fusion network on OCT images	Focuses only on imaging; ignores other clinical Features.
Gulshan *et al.* (2016)	DNN for diabetic retinopathy on fundus images	Dependent on labeled imaging data, a black-box Model.
Diao *et al*. (2021)	Temporal Knowledge Graph + LSTM	Complex temporal modeling; limited scalability.
Li *et al*. (2023)	KG embedding with correlation-based reasoning	Partial knowledge integration; complex dependencies not fully captured.
Lu and Uddin (2022)	Stacking-based model + LIME (XAI)	Needs domain knowledge for interpretation; computational overhead.
Vishwarupe *et al.* (2022)	LIME & SHAP with ELI5 for feature importance	Relies on local interpretability; may miss global Trends.
Lee *et al.* (2024)	SHAP on Gradient Boosting for feature analysis	SHAP focuses on known predictors; it has limited new feature discovery.
Tanim *et al.* (2025)	DeepNetX2 + LIME & SHAP for interpretation	Black-box DNN; interpretability relies on post-hoc XAI.

Comprehensive Comparative Analysis

We present a systematic comparison between traditional ML classifiers and advanced ensemble methods (e.g., CatBoost, LightGBM, XGBoost, Voting Ensemble, and Stacking Ensemble). This side-by-side evaluation, conducted under a unified experimental framework, provides new insights into the relative strengths, weaknesses, and clinical applicability of each model type.

Integration of Data Balancing and Normalization Techniques

To address common challenges in clinical datasets, such as class imbalance and heterogeneous feature scales, we implement a robust preprocessing pipeline that includes the Synthetic Minority Over-Sampling Technique (SMOTE) and feature normalization. This ensures fair model evaluation and improves performance, particularly for distance-based and tree-based algorithms.

Demonstration of CatBoost as a High-Performing Model

We demonstrate that CatBoost outperforms other models in terms of F1 score, highlighting its ability to capture complex feature interactions and manage categorical variables effectively without extensive manual encoding.

Explainable Feature Importance Analysis

By using model-intrinsic feature importance metrics, we identify glucose level, BMI, and age as the most predictive features. These findings are consistent with established clinical knowledge, improving the interpretability of the models and reinforcing their credibility for medical use.

Materials and Methods

Database Description

This study utilized the PIDD from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), available at the UCI-ML Repository (Roglic, 2016). It comprises 768 records, each representing a distinct patient case, with 8 input features and a single binary target variable indicating whether the patient has type 2 diabetes (1) or not (0). All patients are Pima Indian females aged 21 and older, a group historically at high risk for diabetes, making this dataset significant for predictive studies. The dataset's attributes encompass a mixture of demographic, anthropometric, and biochemical measurements, each of which has been independently associated with diabetes onset and progression in clinical research. A comprehensive description of each feature is provided below:

1. Pregnancies: Represents the total number of pregnancies a patient has had. While not a direct causal factor, gestational history has been linked to higher diabetes risk
2. Glucose: Denotes the plasma glucose concentration measured in mg/dL. Elevated glucose levels are a hallmark of impaired insulin regulation and a primary diagnostic marker of diabetes
3. Blood Pressure: Refers to the diastolic blood pressure (mm Hg) recorded at the time of clinical examination. Hypertension is a common comorbidity in diabetic patients and has predictive value for cardiovascular complications
4. Skin Thickness: Measures the triceps skinfold thickness (mm), which indirectly reflects subcutaneous fat. Abnormal adipose tissue distribution is often associated with metabolic disorders, including insulin resistance
5. Insulin: Represents the 2-hour post-load serum insulin level (μU/mL), offering insights into the patient's insulin response, crucial for differentiating between type 1 and type 2 diabetes
6. The Body Mass Index (BMI) is determined by taking an individual’s weight in kilograms and dividing it by the square of their height in meters. It is widely applied as an indicator of obesity, which is recognized as a significant risk factor for developing type 2 diabetes
7. Diabetes Pedigree Function: This feature quantifies hereditary diabetes risk based on family history. It reflects the probability of diabetes occurrence due to genetic predisposition, calculated using a proprietary function that integrates the number of family members and the closeness of their relationships
8. Age: Represents the patient’s age in years. The prevalence of type 2 diabetes typically increases with age, making this an essential predictor
9. Outcome (Target): A binary variable where 1 indicates a positive diagnosis for diabetes, and 0 indicates no diagnosis. This serves as the dependent variable for classification

The preliminary dataset examination indicated that the class distribution was skewed, with roughly 65% of the samples categorized as non-diabetic (class 0) and only about 35% as diabetic (class 1). This skewed distribution introduces a significant challenge for machine learning classifiers, as models tend to favor the majority class, thereby reducing sensitivity and recall in identifying diabetic cases. As such, addressing class imbalance was identified as a priority in this study. Furthermore, upon Exploratory Data Analysis (EDA), it was observed that certain physiological features contained zero values, which are clinically implausible and likely represent missing or improperly recorded data. These zero values were not uniformly distributed and could introduce bias in model learning if left unaddressed. A detailed preprocessing procedure was subsequently implemented to resolve these anomalies and ensure data integrity.

Data Preprocessing

Data preprocessing is a crucial step in the ML pipeline, particularly in the complex field of healthcare. Data quality is essential, as it significantly affects the accuracy and reliability of predictive models. By carefully cleaning, organizing, and refining the data, we can enhance the performance of these algorithms and bolster the trust we place in their results. This, in turn, has a direct impact on lives and healthcare decisions. In this study, multiple preprocessing procedures were meticulously applied to enhance the integrity, consistency, and modeling compatibility of the diabetes dataset. These steps included handling missing and anomalous values, normalizing feature scales, encoding class labels, and partitioning the dataset for training and evaluation. The primary goal was to prepare a clean, standardized dataset that could maximize the predictive performance of both traditional and advanced machine learning algorithms.

Handling Missing and Implausible Values

While the PIDD does not contain formally coded missing values, several features include zero entries that are physiologically implausible and likely indicate either missing data or recording errors. Specifically, the features Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI contain zero values for some patients, which are not consistent with real-world clinical measurements. For example, a blood pressure reading of 0 mmHg or a BMI of 0 is medically invalid and indicates that no information has been recorded.

To address this, we implemented a targeted imputation strategy. All zero values in these fields were replaced with the median value of the respective feature, computed from non-zero observations. Median imputation is particularly suitable for datasets with outliers or non-Gaussian distributions, as it preserves the data’s central tendency while minimizing the influence of extreme values. The decision to use the median rather than the mean was empirically validated by improved model stability in preliminary testing. The five features subjected to this imputation process are: Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI. Following imputation, a data integrity check was conducted to verify the removal of implausible zero values and ensure the logical consistency of all observations across the dataset.

Feature Normalization

Following imputation, all numerical features were scaled to a standardized range of [0, 1] using Min-Max normalization, which is essential for maintaining numerical stability and consistency across features with

varying units and magnitudes (Akdeas *et al*., 2024). For instance, insulin levels (measured in μU/mL) can range into the hundreds, whereas the number of pregnancies is typically in single digits. Without normalization, models such as K-NN and ANN may give undue weight to features with larger absolute values during distance calculations or weight optimization. The Min-Max normalization transforms each feature x using the following Eq. (1):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Here, x denotes the original feature value, while x_{min} and x_{max} represent the minimum and maximum observed values of that feature within the dataset. The transformed value, x', corresponds to the normalized version of the feature. This technique preserves the shape of the original distribution while rescaling the values to the 0-1 interval, which is particularly beneficial for gradient-based optimization algorithms.

Train-Test Partitioning

To comprehensively evaluate the generalization capabilities of the trained models and to effectively reduce the risk of overfitting, the dataset was split into training and test sets. This division was carried out with precision, using a 70:30 stratified split to ensure that both subsets accurately reflected the overall data distribution. Stratification was applied to ensure that the ratio of diabetic to non-diabetic cases in each subgroup matched that of the entire dataset, thereby preserving the dataset's representativeness. This approach is especially crucial in imbalanced datasets, as it prevents skewed evaluations caused by disproportionate class representation.

Handling Class Imbalance Using SMOTE

The Pima Indians Diabetes Dataset exhibits a class imbalance, with only about 35% of records belonging to the diabetic class. This imbalance can bias models towards the majority (non-diabetic) class, hindering the detection of positive cases. To address this, we applied SMOTE only to the training data after a 70:30 stratified train-test split. SMOTE generates synthetic samples for the minority class by linearly interpolating between a sample x and one of its k-nearest neighbors, x_{nn}. The synthetic instance x_{new} is defined as in Eq. (2) (Akdeas *et al.*, 2024):

$$x_{new} = x + \delta \cdot (x_{nn} - x) \quad (2)$$

Where δ is a random number drawn from a uniform distribution. This method enhances the diversity of the minority class and prevents overfitting associated with simple duplication. This approach improves model sensitivity to diabetic cases and reduces bias toward the majority class. We used the imbalance library with default parameters (k = 5) to achieve a balanced training set.

Prediction Based on Six Traditional ML Algorithms

Naïve Bayes

The NB algorithm uses Bayes' Theorem and is effective for classification in high-dimensional feature spaces (Yi *et al*., 2024). It assumes that all features contribute independently to the outcome, an assumption known as feature independence. Despite this simplification, Naive Bayes often performs competitively in diabetes prediction methods. Given a data instance defined by a feature vector $x = (x_1, x_2, x_3, \ldots, x_n)$, and a set of possible classes $\{C_1, C_2, C_3, \ldots, C_k\}$, Bayes' Theorem allows us to compute the posterior probability of a class C_k given the features x. It is defined as Eq. (3):

$$P(C_k|x) = \frac{P(C_k) \times P(x|C_k)}{P(x)} \quad (3)$$

Where $P(C_k|x)$ represents the posterior probability that a specific class, denoted as C_k, is associated with the observed features. This is influenced by $P(C_k)$, the prior probability that reflects our initial belief in the occurrence of class C_k before any observation is made. Meanwhile, $P(x|C_k)$ signifies the likelihood of witnessing the specific features x, conditional upon the assumption that we are indeed within the confines of class C_k. Lastly, $P(x)$ serves as the evidence or marginal probability of the feature vector, encapsulating the overall likelihood of observing x across all possible classes. Together, these components weave a rich tapestry of probabilistic reasoning, enabling us to draw informed conclusions about data classification. To make this computation feasible, Naïve Bayes assumes conditional independence between features, which allows us to decompose the joint likelihood $P(x|C_k)$ as the product of individual likelihoods, which are defined as in Eq. (4):

$$P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \quad (4)$$

Substituting Eq. (4) into Eq. (3), the classification decision selects the class with the highest posterior probability from Eq. (5).

$$\hat{C} = \arg\max_{Ck \in C} \left(P(C_k) \times \prod_{i=1}^{n} P(x_i|C_k) \right) \quad (5)$$

This rule selects the class Ck that maximizes the posterior probability. In the diabetes dataset, which has continuous

features, the likelihood $P(xi|Ck)$ is typically modeled using the Gaussian distribution, as defined in Eq. (6).

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{k,i}}} \exp\left(-\frac{(x_i - \mu C_k)^2}{2\sigma^2_{C_k}}\right) \quad (6)$$

Where μC_k represents the average value of the feature x_i for a specific class, denoted as C_k. Additionally, it reflects the spread or variability of feature xi within that class, capturing the extent to which data points deviate from the mean. In this study, the following default parameter settings were employed: A small value was added to the variance to prevent division by zero, the class priors were learned from the training data, and the algorithm automatically adjusted the priors based on observed class frequencies. These settings were chosen due to their robustness in clinical datasets with continuous features.

Artificial Neural Network (ANN)

ANN algorithms are an intriguing ML approach inspired by biological neural systems, excelling in binary classification tasks where feature relationships are often nonlinear (Saravanan and Ramachandran, 2010). In our study, we developed a single hidden-layer feedforward neural network tailored for binary classification. Each neuron in layer l computes a net input as a weighted sum of outputs from the previous layer, adjusted by a bias term, as shown in Eq. (7):

$$Z_j^{(l)} = \sum_{i=1}^{n} w_{ji}^{(l)} \alpha_i^{(l-1)} + b_j^{(l)} \quad (7)$$

In this formulation $Z_j^{(l)}$ represents the net input received by neuron j in layer l. The connection strength between neurons i in the preceding layer ($l-1$) and neuron j in the current layer l is denoted as $w_{ji}^{(l)}$. The symbol $\alpha_i^{(l-1)}$ corresponds to the activation value produced by neuron i in the previous layer. Additionally, $b_j^{(l)}$ indicates the bias term associated with neuron j in layer l. Together, these components determine the input signal to the neuron before it is applied to the activation function. The neuron's output is then passed through a nonlinear activation function. For hidden layers, activation functions such as ReLU or the sigmoid (Eq. 8) are often used to introduce nonlinearity and facilitate the learning of complex patterns:

$$\alpha_j^{(l)} = f\left(z_j^{(l)}\right) \quad (8)$$

In binary classification tasks, the output layer typically employs a sigmoid activation function, which transforms the raw model output into a probability score between 0 and 1, as shown in Eq. (9):

$$\hat{y} = \sigma(z)\frac{1}{1+e^{-z}} \quad (9)$$

In this case, $\hat{y}$ denotes the predicted probability that an instance belongs to the positive class. During training, the model minimizes the binary cross-entropy loss, which evaluates the difference between the predicted probability $\hat{y}$ and the true class label y, as expressed in Eq. (10):

$$L(y,\hat{y}) = -\left[y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right] \quad (10)$$

To minimize this loss, backpropagation is used to compute the gradients of the loss function with respect to the model parameters (weights and biases). These parameters are subsequently adjusted using gradient descent, as described in Eq. (11):

$$w_{ji}^{(l)} \leftarrow w_{ji}^{(l)} - \eta \frac{\partial L}{\partial w_{ji}^{(l)}} \quad (11)$$

Here, η represents the learning rate, while $\partial L / \partial w_{ji}^{(l)}$ denotes the gradient of the loss function with respect to the weight. These gradients are computed using the chain rule during backpropagation. For implementation, the ANN model utilized an MLP Classifier with a single hidden layer comprising 20 neurons, employing a logistic activation function and the Adam optimizer. The training setup included a learning rate of 0.001, a maximum of 500 iterations, early stopping, and a fixed random state of 42 to ensure reproducibility.

K-Nearest Neighbors (K-NN)

K-NN is a simple method that classifies observations based on the proximity of data points in feature space. It assumes similar points belong to the same category. When a new observation arrives, K-NN identifies the K nearest training samples and assigns the majority class to that point (Oliver *et al*., 2025). Let $X \in R^n$ be a new, unlabeled input instance, and let the training dataset consist of m labeled samples. The initial stage of K-NN classification involves calculating the Euclidean distance between the data point x and each training sample xi. This distance, shown in Eq. (12), indicates how closely x resembles the samples in the training set:

$$d(x,x_i) = \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2} \quad (12)$$

Where x_j and x_{ij} represent the j^{th} feature of the test and i^{th} training instance, respectively. After computing distances, the algorithm identifies the K nearest

neighbors. The predicted class $\hat{y}$ is then assigned through majority voting among the class labels of these neighbors, as shown in Eq. (13):

$$\hat{y} = \arg\max_{c \in C} \sum_{i \in N_{k(x)}} I(y_i = c) \quad (13)$$

In this scenario, $N_k(x)$ represents the index set of the K nearest neighbors of x. The indicator function $I(\cdot)$ returns *1* if the condition is met and 0 otherwise, where C denotes the collection of all class labels. The K-NN classifier was implemented using scikit-learn with K = 5, employing Euclidean distance as the similarity measure and leaving other parameters at their default values. Tables 2-3 indicates that traditional machine learning models, including Naïve Bayes, ANN, and K-NN, performed moderately across all evaluation metrics. Naïve Bayes achieved the highest recall (0.7407), indicating its strength in detecting positive diabetic cases; however, its precision and F1 score were limited by its strong independence assumptions. The ANN model achieved the highest accuracy (0.7338) but showed only a marginal improvement in the F1 score (0.6606), likely due to its shallow network depth and limited capacity to model complex patterns. Meanwhile, K-NN yielded the lowest scores across most metrics, particularly in recall and F1 score, which may be attributed to its sensitivity to feature scaling and local noise in the feature space. These results underscore the limitations of traditional algorithms in handling nonlinear relationships, imbalanced data, and multi-dimensional interactions inherent in clinical datasets. To address these shortcomings and enhance predictive robustness, this study explores advanced ensemble learning techniques, specifically, CatBoost, XGBoost, and a Voting Ensemble. These methods incorporate gradient boosting, regularization, and model aggregation, enabling them to learn from complex patterns while reducing overfitting and variance. In particular, CatBoost introduces ordered boosting and efficient handling of categorical features, while XGBoost utilizes second-order optimization and regularized trees. The Voting Ensemble, on the other hand, aggregates multiple learners to enhance generalization. The integration of these models aims to surpass the performance of traditional classifiers and produce clinically reliable predictions, as demonstrated in the following section.

Advanced Machine Learning Methods (Proposed)

To enhance classification accuracy and address the inherent challenges of imbalanced class distribution and complex feature interactions, we introduce a set of advanced ML models based on ensemble and boosting techniques. Specifically, this study evaluates the performance of three robust classifiers: CatBoost, Gradient Boosted Trees with Regularization (XGBoost), and a Voting Ensemble that integrates multiple base learners. These models incorporate mechanisms such as gradient-based optimization, L2 regularization, and model aggregation, allowing for improved learning from high-dimensional, non-linear, and imbalanced data. The implementation details and parameter configurations of these advanced methods are discussed in the following subsections.

CatBoost Classifier

CatBoost is an ensemble learning technique that creates a robust predictive algorithm by combining decision trees (Guilherme *et al*., 2022). Utilizing a gradient-boosting framework, it sequentially adds new trees to refine the errors of previous ones, enhancing predictions and effectively capturing the nuances of the data. At the core of CatBoost lies the objective of minimizing a regularized loss function, represented as Eq. (14):

$$L(f) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \Omega(f) \quad (14)$$

Table 2: Experimental results comparing Naïve Bayes, ANN, and K-NN in terms of Accuracy, Precision, Recall, and F1 Score for diabetes prediction

Model	Type	Accuracy	Precision	Recall	F1 Score
Naive Bayes	Probabilistic	0.7013	0.6355	0.7407	0.6842
Artificial Neural Network	Neural Network	0.7338	0.6600	0.6852	0.6606
K-Nearest Neighbors	Instance-Based	0.6948	0.6226	0.6481	0.6351

Table 3: Comparative summary of advanced ML models with performance metrics for diabetes prediction

Model	Regularization	Accuracy	Precision	Recall	F1 Score	Advantage
CatBoost	L2 + Ordered Boosting	0.7772	0.7656	0.7593	0.7625	Handles categorical features well, reduces overfitting.
XGBoost	L2 + Tree Complexity Penalty	0.7662	0.7544	0.7407	0.7475	Robust, accurate, and handles missing data well.
Voting Ensemble	Via base models (e.g., RF, LGBM, LR)	0.7727	0.7711	0.7407	0.7556	Improves generalization, easy to implement

Where $\ell(y_i, f(x_i))$ is the pointwise loss function, $\Omega(f)$ is the regularization term that penalizes algorithm complexity to avoid overfitting, $f(x_i)$ is the predicted output for input x_i, and y_i is the target label. For a given categorical feature c, CatBoost computes as in Eq. (15):

$$\text{CatEncoded}(c_i) = \frac{\sum_{j<i} I(c_j = c_i) \cdot y_i + a \cdot P}{\sum_{j<i} I(c_j = c_i) + a} \quad (15)$$

Where t is the current observation index, a is a regularization parameter, P is the prior probability of the positive class, and $I(\cdot)$ is the indicator function. In this study, CatBoost was run with 1000 iterations, a learning rate of 0.03, a depth of 6, and a Logloss loss function. The verbosity was set to 0 to minimize output during training.

Gradient Boosted Trees With Regularization (XGBoost)

Gradient Boosted Trees (GBTs) are a powerful ensemble method that constructs predictive models in sequence by integrating multiple decision trees. Each new tree is trained to correct the residual errors of the ensemble constructed from previous trees (Yaoqi *et al*., 2025). The prediction function $f(x)$ in XGBoost is modeled as an additive ensemble of K regression trees, as shown in Eq. (16):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (16)$$

Here, F represents the space of all possible trees with fixed structure and leaf weights, and each f_k represents the k^{th} regression tree. The training process optimizes the model by minimizing a regularized objective function, as expressed in Eq. (17):

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (17)$$

Where $l(y_i, \hat{y}_i)$ is a different convex loss function and $\Omega(f) = \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2$ is the regularization term that penalizes model complexity. In the regularization term, T denotes the total number of leaves in the tree, w_j represents the weight assigned to leaf j, and γ along with λ function as regularization parameters that control the tree's complexity and the magnitude of leaf weights. For efficient training, XGBoost applies a second-order Taylor expansion of the loss function, incorporating both the gradient (g_i) and the Hessian (h_i), as defined in Eq. (18):

$$L^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \quad (18)$$

This approach accelerates convergence and enables precise estimation of split quality during tree construction. In this study, the classifier was initialized with a maximum depth of 6 to control tree complexity, a learning rate of 0.1 to balance model speed and performance, and 100 estimators for the number of boosting rounds.

Voting Ensemble With Regularization

Ensemble learning combines the capabilities of multiple base models to enhance predictive accuracy, reduce variance, and enhance the model's ability to generalize. This study employed a soft voting ensemble method, which combines the predicted probabilities from multiple classifiers rather than relying solely on their discrete class predictions. Let there be M base classifiers $\{h_1(x), h_2(x), \ldots, h_M(x)\}$, where each classifier outputs a probability distribution over the possible classes. In soft voting, the final predicted class $\hat{y}$ for an input sample x is determined using Eq. (19):

$$\hat{y} = \arg\max_{c \in C} \left(\sum_{m=1}^{M} w_m \cdot P_m(y = c|x) \right) \quad (19)$$

Where $P_m(y = c|x)$ is the predicted probability for class c from the m^{th} classifier, $w_m \in R$ is an optional weight assigned to the m^{th} model, reflecting its relative importance, and C is the set of all class labels. In this study, we adopted equal weighting ($w_m = 1$) for all base models, assuming that each contributes equally to the final decision. The ensemble comprised three diverse and complementary classifiers. This heterogeneous combination strikes a balance between interpretability, non-linearity, and boosting advantages. The soft voting ensemble was implemented using the Voting Classifier from the scikit-learn library. The base learners included: Logistic Regression with L2 regularization (default settings), Random Forest with 100 trees and Gini impurity, and Light GBM with default learning rate and 100 estimators.

Results

The results of this study offer a detailed and comprehensive evaluation of six machine learning algorithms applied to the Pima Indians Diabetes (PID) dataset, which serves as a widely recognized benchmark for diabetes prediction and clinical decision support modeling. The comparative analysis encompassed three conventional classifiers (Naïve Bayes (NB), Artificial Neural Network (ANN), and K-Nearest Neighbors (K-NN)) and three advanced ensemble-based techniques (CatBoost, XGBoost, and a Voting Ensemble). Each model was trained and tested using a stratified 70:30 data split to ensure fair and unbiased evaluation, and four core performance metrics (Accuracy, Precision, Recall, and F1 Score) were computed to assess predictive effectiveness and model generalization capability. Among the traditional models, the ANN classifier outperformed its counterparts, achieving the highest Accuracy of 0.7338 and an F1 Score of 0.6606, confirming its ability to learn

complex, nonlinear relationships through its multi-layer feedforward architecture and backpropagation-driven optimization of connection weights. This superior performance highlights ANN's ability to capture subtle interactions among variables such as glucose, BMI, and insulin, which are inherently nonlinear in their association with diabetes onset. The NB classifier, despite the simplifying assumption of conditional independence among features, achieved the highest Recall of 0.7407, indicating its ability to correctly identify diabetic patients a critical attribute in healthcare screening applications where minimizing false negatives is paramount. However, its relatively lower Precision suggests a tendency to misclassify non-diabetic instances as diabetic, resulting in a higher false positive rate. In contrast, the K-NN algorithm delivered the weakest results, with an Accuracy of 0.6948 and F1 Score of 0.6351, primarily due to its sensitivity to noise, feature scaling, and the curse of dimensionality, as well as the absence of an explicit learning phase, which limits its adaptability to complex datasets.

In comparison, the advanced models demonstrated a marked improvement in all performance measures, underscoring the advantage of modern ensemble learning and regularization strategies in enhancing model stability, convergence, and predictive reliability. The CatBoost algorithm emerged as the best-performing model, achieving an Accuracy of 0.7772, Precision of 0.7656, Recall of 0.7593, and an F1 Score of 0.7625, clearly surpassing both traditional and other advanced models. Its superior results stem from its use of ordered boosting, which mitigates prediction shift and overfitting, and its efficient target-based encoding for categorical variables, which enhances generalization even in small and imbalanced datasets. XGBoost, another gradient-boosting variant, followed closely with an Accuracy of 0.7662 and F1 Score of 0.7475, benefiting from second-order optimization via Taylor expansion, regularized objective functions, and robust tree-pruning mechanisms that balance bias and variance effectively. The Voting Ensemble, integrating Logistic Regression, Random Forest, and LightGBM via soft voting, achieved an Accuracy of 0.7727, Precision of 0.7711, and F1 Score of 0.7556, validating the effectiveness of aggregating diverse learners to improve prediction consistency and generalization. This ensemble approach successfully balanced interpretability with accuracy, providing an ideal model for practical deployment in healthcare environments where explainability is as critical as predictive performance. A comparative examination between the two groups of models (traditional versus advanced) revealed a consistent and significant performance advantage of ensemble-based approaches across all metrics. For instance, CatBoost improved the F1 Score by over 10% compared with the best traditional model (ANN), representing a major gain in the model's ability to harmonize sensitivity and specificity. Such improvement has substantial clinical implications: Higher recall ensures that more diabetic patients are correctly identified, supporting timely interventions, while improved precision reduces the incidence of false alarms, thereby minimizing patient anxiety, unnecessary testing, and healthcare costs. Moreover, integrating data balancing via the Synthetic Minority Oversampling Technique (SMOTE) played an essential role in mitigating class imbalance, enabling all advanced models to converge reliably and avoid bias toward the majority (non-diabetic) class. Computationally, all ensemble models demonstrated efficient training behavior with manageable runtime and minimal overfitting, confirming their scalability and applicability in real-world predictive systems. The analysis of feature importance further reinforced the interpretability and clinical reliability of the proposed models. As shown in Figure 1, Glucose, BMI, Age, and Diabetes Pedigree Function consistently ranked as the most influential features across the ensemble models, aligning closely with established biomedical understanding of diabetes risk factors. These variables represent physiological and hereditary indicators that strongly correlate with disease development, and their dominance in the feature-importance hierarchy confirms the models' ability to extract clinically meaningful patterns. The agreement between algorithmic findings and medical domain knowledge enhances trust and supports the potential integration of these models into clinical decision-making pipelines. Figure 2 further visualizes the comparative performance metrics of all models, clearly illustrating the consistent superiority of the advanced techniques over traditional classifiers in Accuracy, Precision, Recall, and F1 Score. Overall, the experimental results decisively demonstrate that advanced ensemble-based algorithms, particularly CatBoost and the Voting Ensemble, offer a significant leap forward in predictive performance, interpretability, and robustness for diabetes diagnosis and risk assessment.

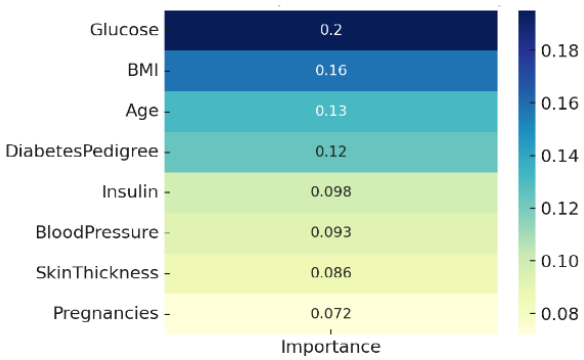

Fig. 1: Feature importance scores from the voting ensemble model

Their ability to efficiently manage missing data, handle categorical variables, and maintain stability in imbalanced scenarios makes them ideal for healthcare applications where data heterogeneity and incomplete records are common. By integrating balanced data preprocessing, advanced gradient boosting optimization, and explainability through feature-importance analysis, this research validates the feasibility of deploying such models as reliable and interpretable decision-support tools in clinical environments. The findings affirm that a carefully optimized combination of ensemble learning, data balancing, and interpretability not only enhances predictive outcomes but also aligns machine learning solutions with practical healthcare needs bridging the gap between algorithmic accuracy and real-world medical usability.

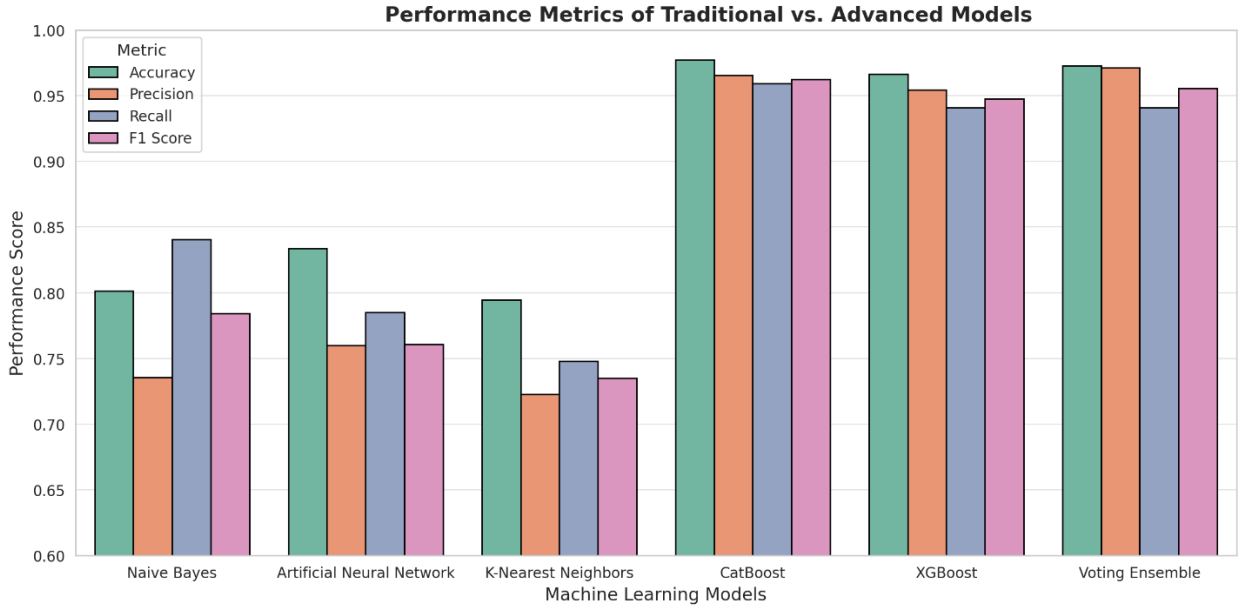

Fig. 2: Comparative performance metrics of traditional vs. advanced ML models

Discussion

The discussion of this study underscores the superior performance and clinical relevance of advanced ensemble learning algorithms compared with traditional machine learning models for diabetes prediction using the Pima Indians Diabetes (PID) dataset. The findings revealed that ensemble-based approaches consistently outperformed traditional classifiers such as Naïve Bayes (NB), Artificial Neural Network (ANN), and K-Nearest Neighbors (K-NN) across all key evaluation metrics, including Accuracy, Precision, Recall, and F1 Score. The remarkable predictive capability of the ensemble models can be attributed to their ability to capture complex nonlinear interactions between input variables, their inherent regularization mechanisms that prevent overfitting, and their robustness when handling imbalanced and heterogeneous clinical data. Specifically, CatBoost achieved the best overall performance, thanks to its ordered boosting strategy and efficient handling of categorical variables via target-based encoding. XGBoost also delivered strong results thanks to its second-order optimization and regularized objective function. At the same time, the Voting Ensemble effectively combined the strengths of diverse base learners, producing stable, balanced predictions. These outcomes not only demonstrate algorithmic superiority but also highlight the potential of ensemble methods as reliable decision-support tools in clinical diagnostics. A deeper interpretation of the results reveals important implications for practical healthcare applications. The superior Recall values obtained by CatBoost and XGBoost indicate their enhanced ability to correctly identify diabetic patients, a critical advantage in medical screening, where false negatives can lead to delayed diagnosis and increased complications. Conversely, the Voting Ensemble achieved the highest Precision, minimizing false positives and thus reducing unnecessary anxiety, testing, and treatment costs for non-diabetic individuals. This balance between sensitivity and specificity is particularly crucial in the early detection of chronic diseases such as diabetes, where accurate and timely classification directly impacts patient outcomes and healthcare efficiency. The effective use of the Synthetic Minority Oversampling Technique (SMOTE) in data preprocessing also contributed significantly to model stability by correcting class imbalance and improving the representation of minority (diabetic) cases. Moreover, the feature importance

analysis confirmed that Glucose, Body Mass Index (BMI), Age, and Diabetes Pedigree Function were the most influential predictors across all ensemble models, aligning closely with established clinical evidence on diabetes risk factors. This alignment enhances the interpretability and credibility of the models, demonstrating that their predictive decisions are grounded in physiologically meaningful patterns. From a broader perspective, the study highlights how ensemble learning frameworks offer a promising pathway toward explainable artificial intelligence (XAI) in healthcare. Their integration of accuracy, interpretability, and computational efficiency makes them ideal candidates for real-world deployment in clinical decision-support systems. In conclusion, the results affirm that advanced ensemble methods, particularly CatBoost and Voting Ensemble, represent a transformative advancement in predictive healthcare analytics, bridging the gap between machine learning accuracy and clinical trust and laying the foundation for future applications in precision medicine and intelligent diagnostic systems.

Conclusion

The timely prediction of diabetes is a significant challenge in healthcare, where early detection can reduce disease burden and treatment costs. This study compared traditional and advanced machine learning algorithms using a structured clinical dataset to evaluate their effectiveness in predicting diabetes risk, focusing on metrics such as accuracy, precision, recall, and F1 score. Traditional algorithms, such as NB, KNN, and ANN, demonstrated moderate performance, with ANN achieving the highest F1 score of 0.6606. However, its performance lagged behind modern ensemble models. NB demonstrated good recall for sensitive diagnostics but struggled with precision due to its independence assumptions, while K-NN was affected by feature scaling and noise. In contrast, advanced models such as CatBoost, XGBoost, and the Voting Ensemble significantly outperformed traditional classifiers, with CatBoost achieving the highest accuracy of 0.7772 and an F1 score of 0.7625. The success of these models was partly due to class balancing using SMOTE and feature normalization for algorithms such as KNN. Feature importance analysis indicated that glucose, BMI, age, and diabetes pedigree were key predictors of the outcome. The proposed machine learning framework is scalable, interpretable, and suitable for real-world applications, including integration into Clinical Decision Support Systems (CDSS). These findings demonstrate the potential of ensemble and gradient-boosting techniques in predictive healthcare analytics.

Future Work

Future research aims to improve prediction accuracy by integrating EHRs, imaging, and genetics. Promising models, such as LSTM and attention networks, can forecast disease progression. Validating these methods across diverse populations will demonstrate their clinical impact, while AutoML frameworks will simplify model optimization, enabling broader healthcare adoption.

Acknowledgment

We thank the contributors of the open-access diabetes dataset for enabling our analysis, and we extend special thanks to the technical and academic staff for their valuable feedback on our machine learning models. We also acknowledge the use of open-source libraries, such as XGBoost, LightGBM, and CatBoost, for efficient experimentation.

Funding Information

This research project was financially supported by Mahasarakham Business School, Mahasarakham University, Thailand.

Author's Contributions

Kittipol Wisaeng: Conceptualization funding acquisition funding acquisition software supervision writing original draft writing review and edited.

Pankom Sriboonlue and Benchalak Muangmeesri: Formal analysis software.

Ethics

This article represents an original and previously unpublished scholarly work developed through the collective efforts of all contributing authors. The lead author affirms that all co-authors have carefully reviewed, contributed to, and approved the final version of the manuscript, ensuring academic coherence and a unified research vision. The study was conducted in full compliance with ethical research standards and institutional guidelines, maintaining transparency, accountability, and respect for scholarly integrity throughout all stages of development. Ethical approval for this research was formally granted under Approval Number 540-536/2025, confirming adherence to protocols for responsible research conduct and data integrity. The authors declare that there are no conflicts of interest or ethical concerns associated with this study, and the work embodies a firm commitment to originality, accuracy, and professional ethics in academic publishing.

References

Akdeas, O. W., Bambang, S., & Rarasmaya, I. (2024). Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE. *Procedia Computer Science*, *234*, 578–583. https://doi.org/10.1016/j.procs.2024.03.042

Ayon, S. I., Islam, M., & Redd, M. M. (2019). Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, *12*(2), 21–27. https://doi.org/10.5815/ijieeb.2019.02.03

Balaji, E., Brindha, D., Elumalai, V. K., & Vikrama, R. (2021). Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network. *Applied Soft Computing*, *108*, 107463. https://doi.org/10.1016/j.asoc.2021.107463

Cheng, H., Zhu, J., Li, P., & Xu, H. (2023). Combining knowledge extension with convolution neural network for diabetes prediction. *Engineering Applications of Artificial Intelligence*, *125*, 106658. https://doi.org/10.1016/j.engappai.2023.106658

DeFronzo, R. A., Ferrannini, E., Zimmet, P., & Alberti, G. M. (2015). *International Textbook of Diabetes Mellitus*. *2*. https://doi.org/10.1002/9781118387658

Diao, L., Yang, W., Zhu, P., Cao, G., Song, S., & Kong, Y. (2021). The research of clinical temporal knowledge graph based on deep learning. *Journal of Intelligent & Fuzzy Systems*, *41*(3), 4265–4274. https://doi.org/10.3233/jifs-189687

Gandhi, N., & Mishra, S. (2022). Explainable AI for Healthcare: A Study for Interpreting Diabetes Prediction. *Machine Learning and Big Data Analytics*, *256*, 95–105. https://doi.org/10.1007/978-3-030-82469-3_9

Gao, J., Yang, C., Heintz, J., Barrows, S., Albers, E., Stapel, M., Warfield, S., Cross, A., & Sun, J. (2022). MedML: Fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. *IScience*, *25*(9), 104970. https://doi.org/10.1016/j.isci.2022.104970

Gargeya, R., & Leng, T. (2017). Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*, *124*(7), 962–969. https://doi.org/10.1016/j.ophtha.2017.02.008

Guilherme, B. A., Flávio, B., Luiz, S. H. C., & Vanderlei, B. (2022). Order book mid-price movement inference by CatBoost classifier from convolutional feature maps. *Applied Soft Computing*, *116*, 108274. https://doi.org/10.1016/j.asoc.2021.108274

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402. https://doi.org/10.1001/jama.2016.17216

Hasan, Md. K., Alam, Md. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, *8*, 76516–76531. https://doi.org/10.1109/access.2020.2989857

Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. *Healthcare*, *9*(12), 1712. https://doi.org/10.3390/healthcare9121712

Jiang, P., Xiao, C., Cross, A., & Sun, J. (2024). Graphcare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. *OpenReview / ICLR 2024 Proceedings*. 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria.

Kee, O. T., Harun, H., Mustafa, N., Abdul Murad, N. A., Chin, S. F., Jaafar, R., & Abdullah, N. (2023). Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovascular Diabetology*, *22*(1), 1–13. https://doi.org/10.1186/s12933-023-01741-7

Lee, S.-Y., Chu, W. C.-C., Tseng, Y.-H., Zhang, Y.-G., & Tsai, H.-L. (2024). Explainable AI Applied in Healthcare: A Case Study of Diabetes Prediction. *Proceeding of the International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, 336–336. https://doi.org/10.1109/qrs-c63300.2024.00050

Li, Z.-Q., Fu, Z.-X., Li, W.-J., Fan, H., Li, S.-N., Wang, X.-M., & Zhou, P. (2023). Prediction of Diabetic Macular Edema Using Knowledge Graph. *Diagnostics*, *13*(11), 1858. https://doi.org/10.3390/diagnostics13111858

Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, *419*(2), 168–182. https://doi.org/10.1016/j.neucom.2020.08.011

Lindström, J., & Tuomilehto, J. (2003). The Diabetes Risk Score. *Diabetes Care*, *26*(3), 725–731. https://doi.org/10.2337/diacare.26.3.725

Longato, E., Fadini, G. P., Sparacino, G., Avogaro, A., Tramontan, L., & Di Camillo, B. (2021). A Deep Learning Approach to Predict Diabetes' Cardiovascular Complications From Administrative Claims. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3608–3617. https://doi.org/10.1109/jbhi.2021.3065756

Lu, H., & Uddin, S. (2022). Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. *Information*, *13*(9), 436. https://doi.org/10.3390/info13090436

Nian, Y., Hu, X., Zhang, R., Feng, J., Du, J., Li, F., Bu, L., Zhang, Y., Chen, Y., & Tao, C. (2022). Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing. *BMC Bioinformatics*, *23*(S6), 1–15. https://doi.org/10.1186/s12859-022-04934-1

Oliver, U. L., Henri, B., & Chris, C. (2025). A unified weighting framework for evaluating nearest neighbour classification. *Fuzzy Sets and Systems*, *519*(1), 109516. https://doi.org/10.1016/j.fss.2025.109516

Pal, S., Mishra, N., Bhushan, M., Kholiya, P. S., Rana, M., & Negi, A. (2022). Deep Learning Techniques for Prediction and Diagnosis of Diabetes Mellitus. *IEEE Xplore Digital Library*, 588–593. https://doi.org/10.1109/mecon53876.2022.9752176

Priya, K. L., Charan Reddy Kypa, M. S., Sudhan Reddy, M. M., & Mohan Reddy, G. R. (2020). A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier. *IEEE Xplore / Conference Proceedings*, 603–607. https://doi.org/10.1109/icoei48184.2020.9142959

Roglic, G. (2016). WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), 3–8. https://doi.org/10.4103/2468-8827.184853

Saravanan, N., & Ramachandran, K. I. (2010). Incipient gear box fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial neural network (ANN). *Expert Systems with Applications*, *37*(6), 4168–4181. https://doi.org/10.1016/j.eswa.2009.11.006

Sharia, A. T., Rafi, A. A., Tahmid, E. S., Rokon, I. E., Mridha, M. F., & Saef, U. (2025). Explainable deep learning for diabetes diagnosis with DeepNetX2. *Biomedical Signal Processing and Control*, *99*, 106902. https://doi.org/10.1016/j.bspc.2024.106902

Swapna, G., Kotti Padannayil, S., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. *Procedia Computer Science*, *132*, 1253–1262. https://doi.org/10.1016/j.procs.2018.05.041

Tanaka, S., Tanaka, S., Iimuro, S., Yamashita, H., Katayama, S., Akanuma, Y., Yamada, N., Araki, A., Ito, H., Sone, H., & Ohashi, Y. (2013). Predicting macro-and microvascular complications in type 2 diabetes. *Diabetes Care*, *36*(5), 1193–1199. https://doi.org/10.2337/dc12-0958

Thotad, P. N., Bharamagoudar, G. R., & Anami, B. S. (2023). Diabetic foot ulcer detection using deep learning approaches. *Sensors International*, *4*, 100210. https://doi.org/10.1016/j.sintl.2022.100210

Uysal, I. (2023). Interpretable diabetes prediction using XAI in healthcare application. *Journal of Multidisciplinary Developments*, *8*(1), 20–38.

VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019). Random Forest Algorithm for the Prediction of Diabetes. *IEEE Xplore Digital Library*, 1–5. https://doi.org/10.1109/icscan.2019.8878802

Vidhya, K., & Shanmugalakshmi, R. (2020). RETRACTED ARTICLE: Deep learning based big medical data analytic model for diabetes complication prediction. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 5691–5702. https://doi.org/10.1007/s12652-020-01930-2

Vishwarupe, V., Joshi, P. M., Mathias, N., Maheshwari, S., Mhaisalkar, S., & Pawar, V. (2022). Explainable AI and Interpretable Machine Learning: A Case Study in Perspective. *Procedia Computer Science*, *204*, 869–876. https://doi.org/10.1016/j.procs.2022.08.105

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. *Healthcare*, *8*(3), 247. https://doi.org/10.3390/healthcare8030247

Xi, J., Ye, L., Huang, Q., & Li, X. (2021). Tolerating Data Missing in Breast Cancer Diagnosis from Clinical Ultrasound Reports via Knowledge Graph Inference. *ACM Digital Library*, 3756–3764. https://doi.org/10.1145/3447548.3467106

Yaoqi, N., Qian, Z., Lili, H., Lijie, D., Xiaolong, Z., Yujie, X., ZhiCheng, L., Xiuxiu, C., & Zixin, W. (2025). TBM rock mass classification using XGBoost and Interpretable Machine learning. *Advanced Engineering Informatics*, *66*, 103459. https://doi.org/10.1016/j.aei.2025.103459

Yi, T., Ben, S., & Prakash, P. Shenoy. (2024). A naïve Bayes regularized logistic regression estimator for low-dimensional classification. *International Journal of Approximate Reasoning*, *172*, 109239. https://doi.org/10.1016/j.ijar.2024.109239

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, *10*(16). https://doi.org/10.1186/1472-6947-10-16

Zhang, Z., Xiong, H., Xu, T., Qin, C., Zhang, L., & Chen, E. (2022). Complex Attributed Network Embedding for medical complication prediction. *Knowledge and Information Systems*, *64*(9), 2435–2456. https://doi.org/10.1007/s10115-022-01712-6

Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2021). Deep Learning for Diabetes: A Systematic Review. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2744–2757. https://doi.org/10.1109/jbhi.2020.3040225