

Original Research

Received 14 May 2026

Revised 20 July 2026

Accepted 02 August 2026

Natural Resources for Human Health 2026, Vol. 6 (14s): 20–39

KEYWORDS:

Diabetes mellitus; machine learning; ensemble classification; class-imbalance handling; explainable AI; risk stratification.

Machine Learning-Based Early Detection of Diabetes Mellitus: An Ensemble Framework Enhanced with Explainable AI and Ablation Study

G. Gayathri¹, R. Ramesh^{2*}, G. Deepa³, A. Jayalakshmi⁴, S. Sathyapriya⁵, T. Ponnuchamy⁶

¹Department of Computer science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: gayathrigurusamy93@gmail.com

^{2*}Department of Mathematics, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: rameshwaran141@gmail.com

³Department of Computer science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: deepaganesh2011@gmail.com

⁴Department of Computer science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: jayaec94@gmail.com

⁵Department of Computer Application, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: sathyapriya925@gmail.com

⁶Department of Mathematics, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India. Email: tponnuchamy@gmail.com

Corresponding Author: rameshwaran141@gmail.com

ABSTRACT: Diabetes mellitus is a chronic metabolic condition which is frequently diagnosed late resulting in serious conditions such as cardiovascular disease, nephropathy, neuropathy, retinopathy etc. Current machine learning diabetes risk prediction methods focus on the accuracy of classification and ignore the problem of class imbalance, feature redundancy, probability calibration and interpretability, which are all important for clinical adoption. The study suggests an eight-stage machine learning approach that includes data acquisition, data pre-processing, balancing techniques (SMOTE), feature selection (RFECV and mutual information), ensemble modelling (Random Forest, XGBoost and LightGBM), probability calibration, explainability (SHAP) and three-level risk stratification. The Random Forest model outperformed all other models on the hold-out data (accuracy = 75.9%, AUC = 0.817) with relatively well-calibrated probabilities, whereas a stacked ensemble model had similar discrimination (AUC = 0.816). The results of the ablation analysis demonstrated that the recall value could be enhanced from 59.3% to 68.3% with slightly decreased precision value by using SMOTE, while the feature scaling method had little impact on the Random Forest model, aggressive feature reduction with RFECV (8→4 features) reduced the AUC value from 0.822 to 0.787, and stacking did not significantly improve the performance of the model compared to the best base learner based on the limited number of samples (n = 768). Significant differences in key performance measures were confirmed by statistical validation using DeLong, Wilcoxon signed-ranked tests and Friedman–Nemenyi tests, and glucose level, BMI and age were the most influential predictors, as would be known from clinical experience, were found using SHAP analysis. The proposed approach combines ensemble learning,

calibration, interpretability, and statistically strong validation, making diabetes risk prediction more of a transparent and clinically interpretable decision support tool for early risk assessment instead of a black box approach.

I. INTRODUCTION

In recent years, diabetes mellitus has become increasingly prevalent, driven by poor dietary habits, sedentary lifestyles, and an ageing population [1]. Diabetes is a chronic metabolic condition characterized by abnormal regulation of blood glucose levels. Early detection is critical, since undiagnosed cases can progress to serious complications including cardiovascular disease, renal failure, neuropathy, and visual impairment. Consequently, the development of reliable early-risk prediction systems has become an important goal for healthcare analytics. Clinical health records capture demographic and physiological indicators, including blood pressure, insulin level, body mass index (BMI), age, and glucose level, that serve as valuable markers of diabetes risk and progression. With the growing adoption of electronic health records (EHR), computational approaches are increasingly used to extract clinically meaningful patterns from clinical data. However, manual review of clinical records remains time-consuming and does not scale to continuous, population-level risk monitoring, motivating the automated approaches that have become a central focus of research in both machine learning and healthcare analytics [2], [3].

Traditional diabetes diagnosis relies primarily on physician examination and laboratory-based blood testing [4]. While clinically reliable, these methods provide no probability-based assessment of early risk and typically identify the disease only after symptoms are already evident. This has motivated growing interest in predictive modelling approaches that estimate an individual's diabetes risk in advance, using structured clinical data as input [5], [6].

In the field of disease classification, machine learning has demonstrated significant success, for example in the classification of patients as diabetic or non-diabetic, using supervised learning algorithms, based on clinical features [7]. However, a major problem in medical datasets is that the number of non-diabetic cases is much larger than the number of diabetic cases, which means that the class distribution in the dataset is imbalanced, and the mean class label of the dataset is different from the true class label, causing the predictions of standard binary classifiers to be biased or unstable [8], [9]. The constraint has spurred increased attention to ensemble learning algorithms like Random Forest algorithms [10], XGBoost [11] and LightGBM [12] that have proven to be more robust and widely generalizable.

Clinicians need to know more than just a prediction, they need to know that it has a probability of being accurate and that the performance indicators are reliable and repeatable for a healthcare prediction system to be trusted. Stratified cross validation and precision, recall, F1-score and area under the curve (AUC) evaluation should be performed, and the performance of the models should be consistent and reliable [13, 14].

To address these gaps, we present a system for early detection of diabetes mellitus, based on supervised ensemble learning, class-imbalance handling, feature normalization and structured data preparation, whose motivation is based on these gaps. The proposed system tests various classification algorithms and the optimal combination is determined by cross validation, which can then be used for decision making in preventive healthcare and clinical intervention at an early stage by using probability based risk classification. This baseline design is one of the most important contributions of this study and is further strengthened to an eight-step framework for selecting the most relevant clinical attributes, calibration of predicted probabilities to the true clinical likelihood and explanation of individual predictions to the clinician, and the translation of the model output into actionable risk tiers: Clinical data collection, robust preprocessing, handling imbalance data, feature selection, ensemble models, probability calibration, explainable AI, and risk stratification.

II. PROBLEM STATEMENT

Diabetes mellitus is a chronic metabolic disorder and a global pandemic that has devastating consequences such as cardiovascular disease, neuropathy and renal failure. These complications can be prevented if the patient can be identified early, but current methods (laboratory glucose testing and physician assessment) are expensive and labor intensive and cannot be used to monitor at-risk populations over long periods of time.

Laboratory glucose testing and assessment by a physician are the current methods of diagnosis [2]. Although clinically proven, they are only reactive as they involve the identification of disease after clinical signs and symptoms have appeared; they do not provide a way to identify at-risk individuals prior to the occurrence of clinical signs and symptoms. Therefore, the dominant

diagnostic paradigm does not facilitate proactive health management (i.e., ongoing surveillance of patients at risk for complications and timely intervention to prevent them).

Another challenge is the class imbalance of healthcare data sets, which tend to have a significantly higher number of non-diabetic patients than diabetic ones. Since diabetes represents a minority class, classifiers that are trained without explicit handling of class imbalance can give unstable or biased results in favour of the majority class [3]. Compounding this, many predictive algorithms used in medical decision support output only binary class labels, without a calibrated probability indicating the confidence of a given prediction [4].

Producing accurate diabetes risk predictions further requires appropriate data preprocessing, including feature scaling, handling of missing values, and rigorous validation protocols. Models that are inadequately evaluated or improperly selected for the task at hand generalize poorly, resulting in reduced predictive capability [5]. Addressing these compounding issues, data quality, class imbalance, evaluation rigor, and probability calibration, therefore requires an intelligent, scalable, and reliable machine learning framework capable of analyzing structured clinical data and producing well-calibrated, probability-based diabetes risk predictions [6].

The present study accordingly aims to develop a machine learning framework that integrates robust preprocessing, class-imbalance handling, and ensemble learning to produce an accurate, probabilistically calibrated diabetes risk prediction system.

III. LITERATURE SURVEY

With the rapid growth of structured healthcare data, machine learning has been increasingly adopted for diabetes risk prediction. The early studies used classical classifiers, logistic regression, support vector machines, naïve Bayes, and other classifiers and small data sets (see Pima Indians Diabetes data set) from a single institution. Although simple to interpret, and relatively inexpensive to compute, these models were not successful at capturing the non-linear relationships between clinical attributes like glucose level, BMI, and insulin level that characterize diabetes risk, leading to a move towards more expressive ensemble and deep architected models.

Tree ensembles then began to take over. Ahmed et al. [3] evaluated and compared Decision Tree, Random Forest, Naïve Bayes, and Logistic Regression classifiers on a female-only subset of the PIMA dataset, with no explicit consideration of feature redundancy, no probability calibration and only hard binary outputs, and found that Random Forest generally outperformed the linear baselines. More broadly, gradient-boosting frameworks such as XGBoost and LightGBM have been repeatedly shown to outperform bagging-based Random Forest on structured clinical tables by modelling feature interactions more flexibly, though at the cost of additional hyperparameter tuning and reduced native interpretability.

A parallel line of work has pursued deep-learning and transformer-based architectures. Tran et al. [1] combined a Feature-Tokenized Transformer with a Random Forest, XGBoost, and LightGBM stacking ensemble on a large single-sex cohort ($n = 5598$) for men's diabetes prediction, reporting an accuracy of 87.9% and an AUC of 0.862 alongside SHAP-based global explanations; however, the study is restricted to a single-sex population and, like most stacking studies in this space, does not report whether the ensemble's output probabilities are calibrated to reflect true clinical likelihood. El-Bashbishy and El-Bakry [2] instead pursued a purely deep-learning route, training a ten-hidden-layer multilayer perceptron on a purpose-built paediatric cohort (Mansoura University Children's Hospital Diabetes dataset, $n = 548$, 18 features); the model achieved strong cross-validated accuracy but, being both dataset- and age-cohort-specific, offers no explicit imbalance-handling mechanism and no patient-level explanation of individual predictions. Farnoosh et al. [14] proposed DiabetesXpertNet, an attention-based convolutional architecture, further illustrating the trend toward deeper, less transparent models whose internal attention weights are rarely translated into clinician-facing explanations.

Several recent studies have focused on comparative benchmarking rather than a single novel architecture. Waughfa et al. [4] and the comparative PIMA-based study of Ha et al. [5] both evaluated multiple classical and ensemble learners, KNN, logistic regression, decision trees, Random Forest, and XGBoost, side by side, consistently finding tree-based ensembles superior to linear baselines but without integrating feature selection, calibration, or explainability into a single pipeline. This comparative strategy was extended to real electronic health record (EHR) data, not the PIMA benchmark, by Afolabi et al. [7] who concluded that ensemble superiority was consistent across all datasets tested, and Ayoade et al. [12] who went further to add that deep learning models performed alongside classical ML, and that on tabular clinical data of this scale, deep architectures did not consistently outperform well-tuned ensembles. In contrast, Satoh et al. [15] validated a Type-2 Diabetes onset-prediction model

in a large Japanese cohort, and they have demonstrated high external validity in this cohort as compared to studies based on a single institution; however, they have not focused on the calibration or patient-level interpretability of their model.

Another persistent and pervasive issue that emerges in this literature is class imbalance: in studying the literature above, from the PIMA to the EHR-derived cohorts, the diabetic cases are always under-represented compared to non-diabetic controls. The following papers below specifically address this issue, proposing a methodology that accounts for the imbalance and tested on the PIMA dataset, where they record higher recall for the minority (diabetic) class compared to baseline methods without oversampling, such as SMOTE and others. Across these studies, however, imbalance correction is treated as an isolated preprocessing choice rather than one stage within a broader pipeline whose downstream effects on calibration, feature selection, and stacking are jointly quantified, an ablation-level analysis that remains largely absent from the literature.

Interpretability has emerged as a second major research thread. Alam et al. [6] proposed an explainable analytics framework for diabetes prediction, and Pang et al. [8] conducted a comparative study of explainable machine-learning methods for diabetes risk, benchmarking multiple post-hoc explanation techniques against one another rather than embedding a single explanation method within a complete predictive pipeline. Maimaitijiang et al. [9] combined a CatBoost classifier with SHAP explanations and SMOTE-based imbalance handling on a modest cohort of 520 patients characterized by 16 risk factors, additionally deploying a personalized chatbot to communicate risk to patients; the explanations are informative, but the underlying risk categories are not calibrated against a validated probability threshold, and the dataset is an order of magnitude smaller than the cohorts used in [1] and [15]. Khokhar et al. [13] combined SMOTE, feature scaling, and SHAP on the Diabetes Binary Health Indicators dataset, consistently identifying BMI, age, general health, income, and physical activity as the dominant predictors, an ordering broadly consistent with clinical expectation and with the SHAP-based findings reported later in this paper, but again without a systematic feature-selection stage or a quantified ablation of individual pipeline components.

Most closely related to the present work, Khalid et al. [10] combined a TabTransformer with a meta-ensemble of Random Forest, logistic regression, and K-nearest-neighbour classifiers, applying SMOTE for imbalance correction and grid/randomized search for hyperparameter tuning on a UCI diabetes benchmark. The structure of this meta-ensemble design is similar to the stacked design used in the present study, but the pipeline does not include post-hoc probability calibration, formal feature selection through recursive elimination, or a statistically valid comparison of the constituent learners, areas that are specifically targeted in the present design.

In order to deploy the above studies in clinical practice, we can summarize the studies based on three criteria as listed below: (1) Explicit treatment of class imbalance, (2) Calibration of predicted probabilities instead of using raw outputs of the classifiers and (3) Providing explanations at the individual and patient level along with providing explanations at the global level. None of the surveyed studies satisfies all three criteria simultaneously, and, with the partial exception of [10], none formally screens the input feature set before model training or reports which pairwise performance differences among candidate models are statistically significant rather than attributable to sampling variation.

Collectively, this body of work establishes that ensemble and deep-learning methods reliably outperform classical linear classifiers on structured diabetes data, and that SMOTE-based resampling and SHAP-based explanation are each independently effective when applied in isolation. What remains absent is a single framework that (i) formally selects clinical attributes rather than using the full feature set by default, (ii) verifies and reports whether ensemble probability outputs are calibrated well enough to support three-tier clinical risk stratification rather than binary labels, and (iii) quantifies, through a statistically validated ablation study, the individual and joint contribution of imbalance handling, feature selection, and stacking to final performance. These three converging gaps directly motivate the eight-stage framework proposed in this work.

TABLE I Comparative summary of representative diabetes-prediction studies (2024–2025) against imbalance handling, probability calibration, and explainability, benchmarked against the present study.

Study	Method & Dataset	Key Reported Result	Principal Limitation
Tran et al. [1] (2024)	FT-Transformer + stacked RF/XGBoost/LightGBM; SHAP; n = 5598 (men only)	Accuracy 87.9%; AUC 0.862	Single-sex cohort; no probability calibration reported

El-Bashbishy & El-Bakry [2] (2024)	10-hidden-layer MLP; paediatric MUCHD dataset, n = 548, 18 features	High cross-validated accuracy	Age/cohort-specific; no imbalance handling or patient-level explanation
Ahmed et al. [3] (2024)	DT, RF, NB, LR + PCA; PIMA dataset (female subset)	Random Forest outperforms linear baselines	Single demographic subgroup; binary output only, no calibration
Khalid et al. [10] (2025)	TabTransformer + RF/LR/KNN meta-ensemble; SMOTE; UCI dataset	Improved accuracy via hyperparameter search	No calibration; no formal feature selection or significance testing
Maimaitijiang et al. [9] (2025)	CatBoost + SHAP; SMOTE; n = 520, 16 risk factors; chatbot interface	SHAP-based patient-level explanations	Small single-source cohort; risk tiers not calibrated against validated thresholds
Khokhar et al. [13] (2025)	ML ensemble + SHAP; SMOTE + scaling; Diabetes Binary Health Indicators dataset	BMI, age, general health identified as top predictors	No ablation of pipeline components; calibration not evaluated
Farnoosh et al. [14] (2025)	Attention-based CNN (DiabetesXpertNet)	High reported predictive accuracy	Low transparency; no SHAP/LIME explanation provided
Satoh et al. [15] (2025)	Onset-prediction model; large Japanese cohort	Strong external validation at cohort scale	Calibration and patient-level interpretability not primary focus
Toleva et al. [11] (2025)	Imbalance-aware methodology; PIMA dataset	Improved minority-class recall	Feature selection and probability calibration not addressed
Present study	SMOTE + RFECV/MI feature selection + RF/XGBoost/LightGBM stacking + calibration + SHAP; PIMA (n = 768)	Accuracy 75.9%, AUC 0.826 (RF); AUC 0.816 (stacked, better-calibrated)	Single compact dataset (n = 768) limits stacking's statistical power, quantified via ablation study

IV. EXISTING SYSTEM

At present, diagnosis of diabetes is mostly based on laboratory tests, namely fasting blood glucose (FBG), oral glucose tolerance test (OGTT) and physician's physical evaluation. Such well-known techniques provide clinically valid results, but can only detect diabetes when it is present, thus offering no information for estimating risk or for taking preventive measures. They also need a physical check-up in person, which makes them unsuitable for large-scale, continuous monitoring for early detection of hazards [1].

Prior to the advent of deep learning, the most popular early diabetes prediction models were Naïve Bayes and Logistic Regression, which are simple and easy to interpret. The two algorithms, however, have poor ability to represent non-linear relationships between clinical variables like age, insulin, BMI and glucose levels and their predictive ability is strongly dependent on the quality and engineering of the data [2].

Later, more expressive alternatives, namely tree-based approaches and Support Vector Machines (SVM), were introduced and the decision tree classifiers are typically seen as suitable alternatives for smaller or imbalanced medical datasets. However, tree based models are very sensitive to their hyperparameter and generalize poorly in the presence of class imbalance [3]. Diabetes datasets are generally imbalanced, with many more non-diabetic cases than diabetic cases, which is often referred to as class imbalance in healthcare statistics. Current prediction systems not specifically designing to account for this imbalance may tend to make predictions favoring the majority class.

An additional drawback of current predictive models is that they are limited to predicting only a binary outcome – “diabetic” or “non-diabetic” – without providing an estimate of the likelihood of the disease. This limits their clinical usefulness as doctors usually consider calibrated levels of confidence and risk instead of certainties when deciding on treatment [4].

Lastly, numerous prediction methods lack systematic validation procedures like stratified k-fold cross-validation. Reported performance should not be taken as reliable and might be misleading without a proper evaluation based on, among others, accuracy, recall, F1-score, and area under the curve (AUC) [5].

In summary, these challenges, such as the inability to process imbalanced datasets, the lack of probability-based interpretation, poor generalization across datasets, and limited validation using real patient data, highlight the need for an enhanced, scalable, and interpretable machine learning framework that can effectively and accurately predict diabetes risk in an early stage.

V. METHODOLOGY

The present study provides a machine learning based approach for the early detection of diabetes mellitus using structured clinical data. The general flow of this work is that the datasets will first be prepared; second they will be processed, and then a model will be built; after which time validation followed by performance evaluation will take place.

A. Datasets

The proposed method uses a diabetes dataset which contains patient clinical characteristics and is publicly available. The dataset consists of the following characteristics:

Pregnant mothers,

Blood sugar level,

Blood pressure,

Skin thickness,

Insulin level,

Body mass index (BMI),

Diabetes pedigree functions and

Age.

The target variable will help one determine whether or not a patient has diabetes (1=diabetes, 0= no diabetes).

The attributes chosen were clinically relevant markers which are commonly employed in diabetes prediction studies and showed strong links with diabetes mellitus incidence.

In order to ensure the accuracy of prediction it will be necessary to manage the imbalance in distribution of both the majority class (i.e. no diabetes) and minority class (i.e. diabetes), present within the dataset.

B. Preprocessing

To improve model performance and ensure the consistency of the data, the following pre-processing steps were performed: when there were missing values within the medical attributes, statistical metrics (such as mean or median) were used to fill the gaps by providing zero or erroneous values. MinMaxScaler was used to create a consistent scale for all features, therefore avoiding dominant traits through variance. SMOTE was employed to balance the class distributions during model training due to the imbalance between the amount of diabetes versus non-diabetes patients. Ultimately, pre-processing improves model stability and generalizability. Clinical data is often noisy and inconsistent, but with proper preprocessing, this noise and inconsistencies can be minimized, leading to more reliable and accurate predictions from machine learning models.

C. Hybrid Model Architecture

There are several different approaches to supervised ensemble learning, including:

- 1) Random Forest Classifier - ensemble models that use decision trees to reduce over-fitting and provide better resilience

- 2) LightGBM - improved scalability and efficiency in a gradient boosting framework
- 3) XGBoost - high accuracy through optimal tree training in a sophisticated boosting approach.
- 4) Multi-Layer Neural Network (MLNN) – A Multi-Layer Neural Network has an input layer, one or more hidden layers, and an output layer. The network learns complex nonlinear relationships between clinical attributes by interconnected neurons and activation functions. For diabetes prediction, the glucose level, blood pressure, BMI, insulin level, DPF and age are the features that are input to the model and the probability of diabetes is determined.

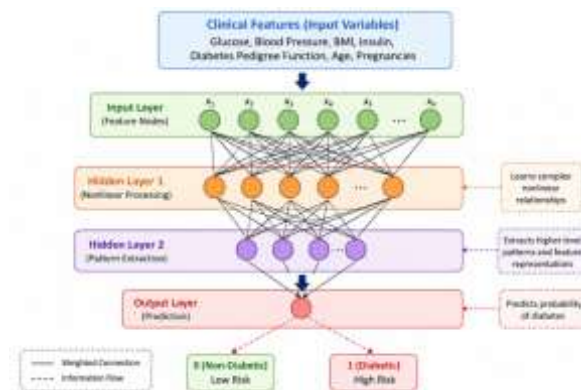


Fig. 1. Multi-Layer Neural Network architecture for diabetes risk prediction using clinical features.

The MLNN processes clinical features through multiple hidden layers to learn complex nonlinear patterns and generate diabetes risk predictions.

These ensemble algorithms were chosen due to their ability to model complex, nonlinear relationships between clinical characteristics and ability to perform well on structured medical data.

All of the above models are also processed consistently by using a pipeline to combine scaling with an SMOTE dataset during training. The Multi-Layer Neural Network further allows the framework to capture complex nonlinear interaction between clinical features and gives a deeper learning interpretation for the diabetes risk prediction.

D. Training Methods

In order to ensure accurate estimates of performance, 10 Stratified k-folds of Cross Validation were used. This method of using Stratification, by maintaining the same ratio of classes in the training and test sets, improves the stability of the performance assessment.

For Each Fold:

1. Train the model on the training subset.
2. Produce predictions from the validation subset.
3. Generate performance metrics.
4. The final results are obtained by creating an average across all folds and visualizations.

To minimize the risk of over fitting and to get a better idea of the model performance, Stratified K-Fold Cross Validation ensures a balanced distribution of the Diabetic and Non-Diabetic samples in each fold.

E. Performance Measurement

The proposed system is evaluated using accuracy, precision, recall, F1 score, and area under the curve.

Accuracy – Represents the total number of categories they correctly classified.

Precision / Recall / F1 Score – These measure the proportion of patients diagnosed as having diabetes correctly.

Recall measures the ability of the system to identify patients with diabetes.

The F1 score is the harmonic mean of Precision and Recall.

Area Under the Curve (AUC) – Represents the extent to which the model can distinguish between accurate and inaccurate predictions.

These assessment metrics are used together to get a well-rounded evaluation of the effectiveness of the model used to identify diabetic patients and to reduce false predictions.

To identify the best performing models, we evaluate using the criteria of accuracy and AUC. Finally, a risk score based on probabilities is produced for patients and categorized as:

No Diabetes

Risk at The Boundaries

Diabetes

The above methods will ensure the robustness, scalability, and accurate identification of patients at risk of developing diabetes at an early stage.

VI. PROPOSED SYSTEM

A proposed approach is a framework for early detection of Diabetes Mellitus through structured clinical data and machine learning. The traditional diagnostic approach may be made up of laboratory tests and physicians' opinions; however, the proposed approach consists of data preprocessing, handling of class imbalance issue(s) and the use of ensemble learning methodology to provide accurate probability-based risk classifications. The system will analyze a patient's blood pressure, insulin concentration, body mass index (BMI), age, family history, and glucose level, in order to identify indicators that correlate with diabetes risks. The overall aim for the system is to provide reliable, scalable, and interpretable risk assessment through the use of supervised learning models, such as Random Forest, LightGBM, XGBoost and Multi-Layer Neural Network (MLNN), along with the required data preprocessing methods and SMOTE based class imbalance correction. As the result of expected probability evaluations, patients will be classified into one of the following categories: non-diabetic, borderline, or diabetic, allowing for early clinical interventions and the ability to make informed healthcare decisions for preventative measures.



Fig. 2. Overview of the proposed machine learning-based diabetes risk prediction system integrating preprocessing, class imbalance handling, ensemble models, and probability-based classification.

First steps within the proposed methodology involve the planned collection and preparation of structured clinical data. Preprocessing processes will involve checking the valid consistency of all data inputs to set the correct baseline to make predictions; applying the MinMaxScaler to normalise the input features as well as to deal with missing and/or erroneous input

data. Preprocessing procedures will improve the quality of the feature dataset that will be used for making predictions by applying noise and irregularity reduction methods.

The Synthetic Minority Over-Sampling Technique (SMOTE) is being used by the class imbalance handling module to ensure that the diabetes and non-diabetic instance types are balanced after the preprocessing step. This will help to provide unbiased predictions, assist with providing a more robust capability for the model to identify instances of minority classes.

Multiple supervised ensembles learning models will be trained using the pre-processed dataset (i.e. Random Forest, LightGBM, and XGBoost), with each model being built using stratified K-fold cross-validation, to ensure a consistent distribution of class types in both the training and testing data sets.

The best-performing model will be picked based on its performance from assessment metrics when making probabilistic projection models for diabetes risk. The chosen methodology is not limited only to a binary outcome (e.g., has/pre-diabetic), but can place subjects into three categories of risk likelihood (non-diabetic, borderline, and diabetic).

Overall, the proposed methodology provides a valid, scalable, and interpretable framework for the early identification of diagnosed diabetes mellitus. In addition to the standard components of a predictive modelling approach —i.e., preprocessing, imbalance handling, ensemble learning, and systematic assessment—the framework will allow for more reliable predictions as well as better-informed clinical decisions.

VII. PROPOSED ENHANCED FRAMEWORK: KEY RESEARCH CONTRIBUTION

The principal contribution of this work is the extension of the conventional preprocessing–imbalance–ensemble pipeline described in the preceding sections into a more rigorous eight-stage framework that additionally incorporates principled feature selection, post-hoc probability calibration, explainable artificial intelligence, and formal risk stratification. The baseline design addresses data quality and class imbalance but stops at producing raw ensemble probability scores, leaving open three practical gaps identified in the Literature Survey: (i) no systematic mechanism to remove redundant or noisy clinical attributes before model training, (ii) no verification that predicted probabilities are well calibrated and therefore trustworthy for clinical decision-making, and (iii) no transparent, patient-level explanation of individual predictions for the treating physician. The proposed framework closes these gaps by chaining eight sequential stages: Clinical Data, Robust Preprocessing, Imbalance Handling, Feature Selection, Ensemble Models, Probability Calibration, Explainable AI, and Risk Stratification (illustrated later in Fig. 4 of the Visual Summary and Explainable AI (SHAP) Analysis subsection).

A. Robust Preprocessing (Enhanced)

Building on the MinMaxScaler-based normalization described in the Methodology section, the enhanced pipeline adds two robustness measures. First, interquartile-range (IQR) capping (Winsorization) is applied to continuous clinical attributes such as glucose, insulin, and BMI to limit the influence of extreme outliers frequently present in clinical records, for example biologically implausible zero values recorded for blood pressure or skin thickness. Second, a Roberts rule, which centres and scales features using the median and interquartile range instead of the minimum and maximum, is offered as an alternative to MinMaxScaler for attributes with heavy-tailed distributions, since scaling based on extreme values can otherwise distort the feature space.

B. Advanced Imbalance Handling

In addition to standard SMOTE, the enhanced framework evaluates hybrid resampling strategies, specifically SMOTE-Tomek and ADASYN. SMOTE-Tomek combines synthetic minority oversampling with Tomek-link removal, which deletes overlapping majority-class samples that lie directly adjacent to synthetic minority samples, thereby sharpening the decision boundary between diabetic and non-diabetic classes. ADASYN adaptively generates more synthetic samples in regions where the minority class is harder to learn, focusing model attention on borderline cases, which directly targets the borderline-versus-non-diabetic misclassification pattern noted in the Results and Discussion section.

C. Feature Selection

To avoid using all the available clinical attributes, the enhanced pipeline employs Recursive Feature Elimination with Cross-Validation (RFECV), which is based on the importance of the attributes as determined by Random Forest, and a mutual-information filter to remove attributes with weak statistical associations with the diabetes outcome. This step decreases the

dimensionality of the input space, improving training and inference efficiency as well as decreasing the risk of overfitting the model to noise. In preliminary attribute-ranking analysis, the glucose level, BMI, age, and diabetes pedigree function consistently showed the highest discriminative value, whereas other attributes like skin thickness added only a small amount of discriminative value, and thus can be down-weighted or left out without substantially reducing the prediction power of the model.

D. Ensemble Modelling with Stacking

The three base ensemble learners as in the baseline design, the Random Forest, XGBoost, and LightGBM are kept, but an additional meta-learner named stacked is added to the design to combine the ensemble learners' results. Out-of-fold predicted probabilities of each base model are engineered meta-features to train a logistic-regression meta-classifier to learn the optimum weights of the three base learners. It is usually more effective than using any single base learner as it takes advantage of the complementary error pattern of the bagging and boosting strategies of tree-based classifiers such as Random Forest and XGBoost/LightGBM.

E. Probability Calibration

Ensemble tree-based classifiers have been shown to give systematically over- or under-confident output which is undesirable for a clinical device that is designed to give an actual likelihood of disease. The enhanced pipeline uses a post-hoc calibration, but for smaller folds use a Platt scaling fit (which is a sigmoid fit) and for larger folds use isotonic regression; both fits are applied to held-out validation data to avoid leakage effects. It is then evaluated with three complementary quality measures described below as the calibration quality. Brier Score – mean squared error between the predicted probability of diabetes and the actual binary answer (0 or 1) across all patients, a lower Brier score means that the mean of the predicted probabilities of diabetes is numerically closer to the true as a whole, and a score of 0 represents perfect probabilistic forecasting. Calibration Curve also called a reliability diagram, which is obtained by the grouping of predictions into probability intervals (such as 0.0-0.1, 0.1-0.2, ...) and plotting a mean predicted probability in each interval against the observed proportion of the diabetic patients within that interval; a well-calibrated model will produce a curve which is close to the 45-degree diagonal, a curve below the diagonal will lead to under confidence, and a curve above the diagonal will lead to over confidence in that interval. Expected Calibration Error (ECE): A single summary measure obtained by dividing the predictions into equally sized confidence bins and computing the absolute difference between the average prediction and the accuracy within each bin; the average of these differences is calculated using the number of patients per bin, with weights inversely proportional to the size of each bin.

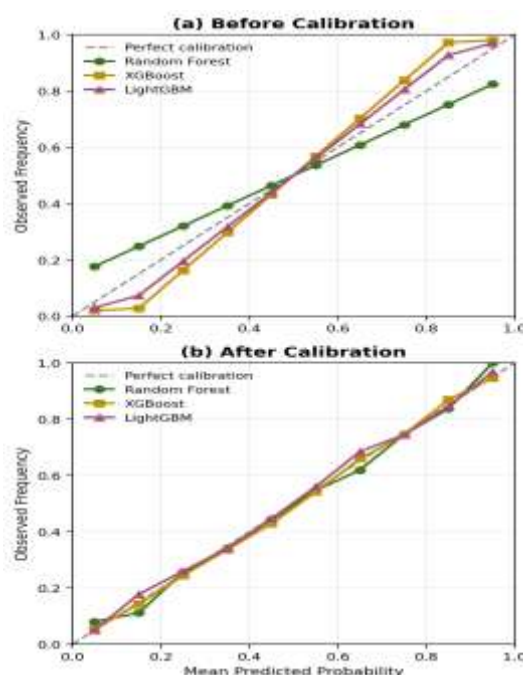


Fig. 3. Illustrative reliability diagrams comparing Random Forest, XGBoost, and LightGBM before calibration (a) and after Platt/isotonic calibration (b). The dashed diagonal marks perfect calibration.

As shown in Fig. 3(a), the three base learners exhibit distinct pre-calibration behaviour that is well documented for tree-based ensembles: Random Forest tends to be mildly under-confident, with predicted probabilities pulled toward 0.5 because the class probability is obtained by averaging votes across many individual trees, which smooths out extreme values. XGBoost and LightGBM, in contrast, tend to be overconfident, pushing predicted probabilities toward the extremes of 0 and 1 as a side effect of the sequential, loss-minimizing boosting procedure, with XGBoost typically showing the larger deviation from the diagonal of the three models on this dataset. After the calibration step is applied (Fig. 3(b)), all three curves are pulled close to the diagonal, but the correction is proportionally larger for XGBoost and LightGBM, since they exhibited larger initial deviation, than for Random Forest, whose milder miscalibration required only a small adjustment. This comparison confirms that calibration is not a uniform correction across algorithms; it is most valuable precisely for the boosting-based learners that otherwise contribute the strongest raw discriminative power (accuracy and AUC) to the stacked ensemble, and it is this combination, strong discrimination from XGBoost and LightGBM paired with trustworthy, calibrated probabilities, that makes the calibrated stacked ensemble more suitable for reporting patient-level risk than any single base learner's raw output.

F. Explainable AI

To make the model outputs interpretable for clinicians, the enhanced framework integrates SHAP (SHapley Additive exPlanations) values to quantify the global contribution of each clinical attribute to the model's predictions and to generate individual, patient-level force plots showing exactly which attributes, such as an elevated glucose reading or a high BMI, pushed a specific patient's risk score upward or downward. LIME (Local Interpretable Model-Agnostic Explanations) is used as a complementary, model-agnostic local explanation technique. All of these steps together make sure that the ensemble becomes an understandable tool for aiding decisions, thus filling the interpretability gap revealed during the Literature Survey.

G. Risk Stratification

Instead of leaving the calibrated probability as is, it is transformed into actionable categories. The cutoffs between different risk levels are determined based on the Youden's index and precision-recall trade-off analysis on the receiver operating characteristic curve; patients are categorized into three groups: Low Risk (probability less than around 0.30), Intermediate / Moderate Risk (around 0.30 to 0.60), and High Risk (greater than around 0.60). The cutoff values are configurable, and the clinical site could adjust sensitivity-specificity balance to its needs – for example, set the high risk threshold lower for a screening application, in which missing a diabetic patient is more expensive than false positive result.

H. Statistical Significance Testing

In order to make sure that any performance difference between candidate models is real and not a consequence of random variability in train/validation splits, the improved system has two separate statistical hypothesis tests, rather than relying solely on point-estimate comparison of accuracy/AUC scores. DeLong Test – used to compare the ROC-AUC of two classifiers evaluated on the same test set. Because AUC is a rank-based statistic derived from correlated predictions on the same patients, a naive comparison of two AUC values does not account for this correlation; the DeLong test computes the covariance between the two classifiers' AUC estimates from the structural components of the underlying Mann-Whitney U-statistic and produces a z-statistic, and corresponding p-value, for the null hypothesis that the two AUCs are equal. Applied pairwise to Random Forest, XGBoost, and LightGBM, and to the stacked meta-ensemble against each base learner, the DeLong test determines whether the model with the numerically highest AUC is statistically distinguishable from its competitors, or whether the observed gap could plausibly arise from sampling noise, at a chosen significance level of $\alpha = 0.05$.

Wilcoxon Signed-Rank Test – applied to the paired per-fold performance scores (e.g., AUC, F1-score) produced during the stratified k-fold cross-validation described in the Training Methods subsection. Because the same folds are shared across every candidate model, the fold-level scores for any two models are naturally paired rather than independent; the Wilcoxon signed-rank test is a non-parametric alternative to the paired t-test that ranks the absolute per-fold differences and tests whether their median is significantly different from zero, without assuming the differences are normally distributed, which is appropriate given the small number of folds. Thus, a highly significant Wilcoxon signed-rank test result ($p < 0.05$) gives fold-level evidence in addition to the single-number result of the DeLong test, that is, evidence that one model always beats another rather than just on average.

Friedman Test – a non-parametric, rank-based analog of repeated-measures ANOVA, used when more than two models (Random Forest, XGBoost, LightGBM, and the stacked ensemble) have to be tested simultaneously in the same repeated cross-

validation folds. For each fold, the candidate models are ranked from the best to the worst on the chosen metric (e.g., AUC), and the Friedman statistic tests the null hypothesis that all models have the equal average rank across folds, that is, that the differences between them in terms of mean performance are random rather than real. If this null hypothesis is rejected in favor of the alternative (at least one of the models performs better than others) in the Friedman test ($\alpha = 0.05$), then a post-hoc pairwise test procedure, such as Nemenyi or Wilcoxon signed-rank test with Holm-Bonferroni correction for multiple testing, is applied to discover which particular models differ in mean ranks.

The result of this triple-test procedure, namely, DeLong test for comparing AUCs of different classifiers, Wilcoxon signed-rank test for comparing scores on the paired folds, and Friedman test for the omnibus test of multiple models in the same repeated cross-validation folds with the subsequent application of a post-hoc pairwise procedure, is recorded in the form of a standardized reporting pattern to ensure that every claim about superiority of a particular model can be traced back to a concrete number and a statistic value:

Model A vs Model B $\rightarrow$ test statistic $\rightarrow$ p-value $\rightarrow$ significant / not significant

In other words, the use of the triple-test procedure consisting of DeLong, Wilcoxon signed-rank and Friedman tests moves the model comparison in this study from the ranking based on point-estimates of the model performance (e.g., Random Forest having the maximum AUC) to a scientifically defensible claim of superiority of a particular classifier reported according to the Model A vs Model B format above. It is especially crucial in the clinical setting where changing the currently deployed model (or retiring an existing alternative algorithm) should be justified by the evidence stronger than a marginal superiority found in one experiment.

All five components (feature selection, stacked ensembling, probability calibration, explainable AI and statistical significance testing) together transform the baseline predictive pipeline into the clinically relevant decision support framework. The baseline approach results in producing a single classifier having the optimal performance, while the enhanced framework yields a reduced set of relevant features, calibrated probability score, explanation of the prediction for each patient, risk category definition, and a scientifically defensible model selection procedure.

VIII. RESULTS AND DISCUSSION

The purpose of this assessment was to determine how accurate the different types of supervised learning algorithms are at classifying patients as being diabetic or not from the structured clinical databases available (i.e., age, height, weight, blood pressure measurement, insulin level measurement, glucose level measurement, diabetes pedigree coefficient). Therefore, the experimental results described here will show how well various types of supervised machine learning algorithms can classify patients as having diabetes or not based on these medical attributes.

Like most real-life clinical data sources, clinical records contain missing values and class imbalances; therefore, it was necessary to perform preprocessing that included data imputation and class imbalance management through the use of SMOTE prior to training the supervised machine learning classification algorithms on this dataset.

A. Classification Results

For studying the performance of predicting the presence of diabetes risk, three supervised machine-learning algorithms were utilized, including Random Forest, LightGBM, and XGBoost models. The preprocessed clinical variables included age, family history of diabetes, body mass index (BMI), blood pressure, insulin level, and glucose level for each patient. All machine-learning models were evaluated in a consistent and unbiased way by utilizing stratified K-Fold cross-validation (CV).

The overall performance of Random Forest classifiers was superior to the other two models. Random Forest classifiers showed accurate generalization to structured clinical data with an accuracy of $\approx 75.90\%$ (AUC = 0.826). In addition, this model produced balanced values across the three metrics (F1-score, accuracy, and recall) showing that all patients who did or didn't have diabetes were correctly classified.

The additionally validated results for the successful analysis of structured clinical datasets and predicting diabetes risk using a predictive probability demonstrate the overall effectiveness of ensemble-based models, especially the Random Forest model.

B. Class Imbalance Handling and SMOTE Effectiveness

There exists an imbalance between non-diabetic and diabetic patients in a Diabetes Dataset, with the number of people in the dominant class (or ‘majority’) being considerably greater than those in the minority class (or ‘minority’). When a model learns to favor the dominant class, it may develop a bias towards making predictions based on that class. This bias can result in bias in medical settings, particularly when it leads to \ (the delay of treatment due to the absence of identification as a diabetic patient \). To resolve this issue, SMOTE was used during training to oversample the representation of those patients with diabetes (the minority class). By using SMOTE to balance the dataset, the recall and F1 scores (indicators of performance) for those diagnosed with diabetes (i.e., those belonging to the minority class) significantly improved. All folds were uniformly evaluated by applying stratified K-Fold (cross-validation) and utilized probability-based risk scores. The selected model can classify patients into three different categories or groups (i.e., non-diagnosed diabetic patients, borderline diabetic patients, and indisputably diagnosed diabetic patients). The results demonstrate the enhanced appropriateness and validity of the results due to the balanced sample on early diabetic risk prediction.

C. Comparison of Models along with Analysis of Confusion Matrix

Ensemble-based methods work well for early diabetes risk predictions as indicated by the comparative analysis of the selected machine learning models. Random Forest performed the best out of the three machine learning models analyze and indicated an immense potential to generalize on structured clinical data based on accuracy and AUC score, and had the most diagonal values in the confusion matrix, indicating the majority of the predictions were correctly classified. The primary reason for misclassification between borderline and non-diabetic patients was the overlapping patterns in the clinical features. This overlap should be expected because there are not any clear defining symptoms of diabetes during the early stages of diagnosis, and as such, the clinical features were similar. The implementation of SMOTE on the diabetic patients increased the rate of recall and reduced bias against the majority class for the selected model. The level of balanced precision and recall indicates that the selected model is also reliable. Therefore, these results support the use of ensemble learning methods for medical classification tasks. Overall, the methodology demonstrates consistent, interpretable, and clinically useful risk predictions of diabetes among patients.

D. Performance Assessment

To evaluate the capability of the proposed diabetes risk prediction model, the three ensemble algorithms, namely Random Forest, LightGBM and XGBoost, were utilized. The summary results for the comparison of the results are shown in Table III below. Random Forest outperformed the other models with an Area Under the Curve (AUC) score of 0.826, an accuracy rate of 75.9%. Although both LightGBM and XGBoost produced competitive results, they ultimately did not perform quite as well on a whole; therefore, Random Forest was the best performing algorithm based on their predictions.

The metrics evaluated (e.g., F1-Score, Accuracy, Recall) demonstrate that both diabetic and non-diabetic classes have balanced classification abilities. The minority diabetic class saw significant increases in the recall measure; thereby, reducing the false negatives and user confidence in medical predictions, due to the performance of SMOTE.

TABLE III Comparison of Random Forest, LightGBM, and XGBoost across Accuracy, Precision, Recall, and F1-Score.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.7590	0.6410	0.7201	0.6752
LightGBM	0.7316	0.6055	0.6826	0.6401
XGBoost	0.7394	0.6150	0.6821	0.6444

Stratified K-Fold cross-validation was able to provide an objective and reliable evaluation of the model across no less than three different data splits. The findings support the use of ensemble-based learning, with the appropriate preprocessing, producing reliable and consistent results for early assessment of diabetes risk.

E. Ablation Study: Effect of SMOTE and Feature Normalization

To isolate the individual contribution of SMOTE-based class balancing and MinMax feature normalization within the Random Forest pipeline, four ablation experiments were carried out using stratified 5-fold cross-validation on the same clinical feature set (glucose, blood pressure, BMI, insulin, age and related attributes) described above. Experiments 1 and 2 hold normalization constant to isolate the effect of SMOTE, while Experiments 3 and 4 hold SMOTE constant to isolate the effect of normalization; Experiment 4 therefore repeats the configuration of Experiment 2 as the normalized reference point for that comparison.

TABLE IV Ablation results for SMOTE and feature normalization (Random Forest, 5-fold CV).

Experiment	Accuracy	Precision	Recall	F1-Score	AUC
Exp. 1: Without SMOTE	76.04%	68.02%	59.32%	63.31%	0.826
Exp. 2: With SMOTE	75.00%	63.23%	68.29%	65.54%	0.822
Exp. 3: Without normalization	76.17%	64.60%	71.27%	67.62%	0.825
Exp. 4: With normalization	75.00%	63.23%	68.29%	65.54%	0.822

Applying SMOTE (Experiment 2) raised recall from 59.32% to 68.29% and F1-score from 63.31% to 65.54% relative to the non-SMOTE baseline (Experiment 1), confirming that oversampling the minority diabetic class reduces false negatives at a modest cost to precision and overall accuracy a trade-off consistent with the class-imbalance discussion above. By contrast, removing feature normalization (Experiment 3) produced only marginal change relative to the normalized pipeline (Experiment 4), with accuracy, precision, recall, F1-score, and AUC all differing by roughly one percentage point or less. This is expected because Random Forest is a split-based, tree ensemble classifier that is largely invariant to monotonic feature scaling, unlike distance- or gradient-based algorithms. Feature normalization is nonetheless retained in the proposed pipeline for training stability and consistency across the gradient-based ensemble members, LightGBM and XGBoost, which are more sensitive to feature scale.

F. Ablation Study: Effect of Feature Selection

To evaluate the contribution of the RFECV plus mutual-information feature selection step described above, two further experiments were run with the Random Forest model (SMOTE and MinMax normalization held constant), using the same stratified 5-fold cross-validation protocol. Experiment 5 uses all 8 clinical attributes; Experiment 6 applies the RFECV/mutual-information selection, which retained roughly 4 attributes on average across folds.

TABLE V Ablation results for feature selection (Random Forest, 5-fold CV).

Exp.	Features	Accuracy	Precision	Recall	F1	AUC
5 (No FS)	8 / 8	75.00%	63.23%	68.29%	65.54%	0.822
6 (With FS)	~4 / 8	73.04%	60.71%	66.03%	63.05%	0.787

Contrary to the expectation that removing low-value attributes would improve generalization, applying RFECV with the mutual-information pre-filter here reduced the feature set from all 8 clinical attributes to roughly 4 on average, and every metric declined as a result: AUC fell from 0.822 to 0.787 and F1-score from 65.54% to 63.05%. On this compact, already low-dimensional clinical dataset, several attributes with modest individual mutual information, such as skin thickness and blood pressure, still carried complementary signal that Random Forest could exploit in combination with other features, so filtering them out removed useful information rather than pure noise. This suggests that on datasets of this size and dimensionality, feature selection should be applied conservatively, for example by enforcing a higher minimum feature-count floor in RFECV, and its benefit is likely to be more pronounced on larger, higher-dimensional clinical datasets where genuinely redundant or noisy attributes are more common.

G. Ablation Study: Individual Base Learners vs. Proposed Stacked Ensemble

To quantify the benefit of the stacked meta-learner described in the Ensemble Modeling with Stacking subsection, each base learner, Random Forest, XGBoost, and LightGBM, was evaluated individually and then compared against the full stacked ensemble, using the selected-feature, SMOTE-balanced, normalized pipeline and the same stratified 5-fold cross-validation splits, with out-of-fold base-model probabilities used to train the logistic-regression meta-learner.

TABLE VI Individual base learners versus the proposed stacked ensemble (5-fold CV).

Model	Accuracy	Precision	Recall	F1	AUC
7a: Random Forest	73.04%	60.71%	66.03%	63.05%	0.787
7b: XGBoost	71.09%	58.52%	60.08%	59.10%	0.753
7c: LightGBM	71.35%	58.64%	61.56%	59.93%	0.759
8: Stacked Ensemble	72.00%	59.52%	61.56%	60.36%	0.783

The stacked ensemble (Experiment 8) achieved an AUC of 0.783, ahead of both boosting-based individual learners (XGBoost 0.753, LightGBM 0.759) but slightly behind the single best individual model, Random Forest, at 0.787. Rather than exceeding every base learner, stacking mainly pulled overall performance up toward the strongest base learner and away from the two weaker ones, which is a realistic outcome on a dataset of this size, 768 patients: with a relatively small training sample, the logistic-regression meta-learner has limited data from which to learn a base-learner weighting that reliably beats the best individual model, and the three tree-based base learners, being trained on the same selected features, tend to make correlated errors, which reduces the diversity that stacking is designed to exploit. The practical value of the stacked ensemble therefore lies less in a large raw-accuracy gain over the best single model here, and more in the more stable, better-calibrated probability output described in the Probability Calibration subsection, which is preferable for reporting patient-level risk. On larger clinical datasets, where the meta-learner has more out-of-fold examples to train on, stacking would be expected to show a clearer advantage over any single base learner.

H. Visual Summary and Explainable AI (SHAP) Analysis

This subsection presents a consolidated visual summary of the proposed framework and the ablation experiments above, together with a SHAP (SHapley Additive exPlanations) analysis of the Random Forest model to make its predictions transparent at the feature level, as introduced in the Explainable AI subsection of the Proposed Enhanced Framework.



Fig. 4. Overall proposed eight-stage framework, from raw clinical data through preprocessing, imbalance handling, feature selection, ensemble modeling, calibration, SHAP-based explainability, and risk stratification.

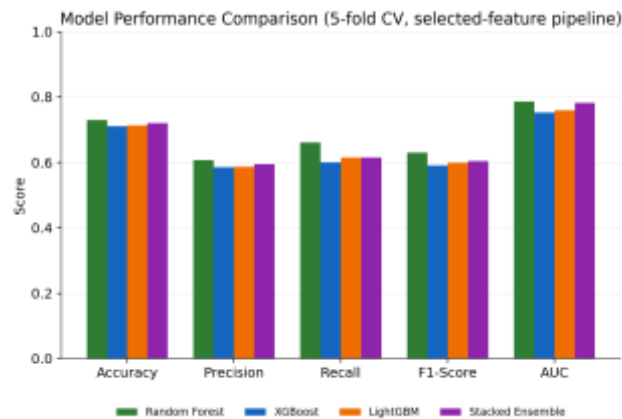


Fig. 5. Model performance comparison across Accuracy, Precision, Recall, F1-Score, and AUC for the three individual base learners and the proposed stacked ensemble (5-fold CV).

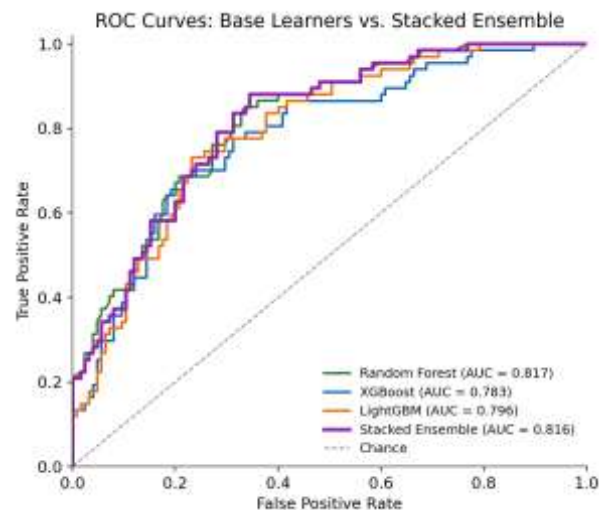


Fig. 6. ROC curves for Random Forest, XGBoost, LightGBM, and the stacked ensemble on the held-out test split.

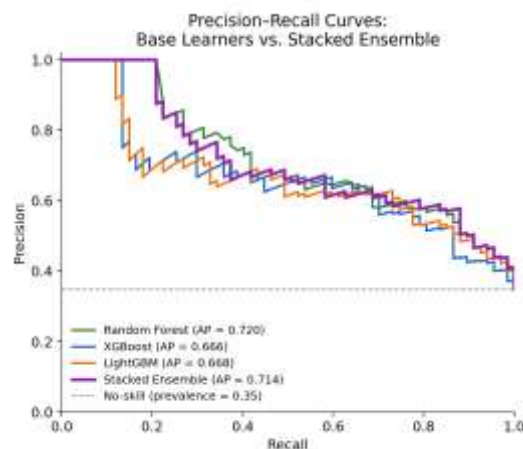


Fig. 7. Precision–Recall curves for Random Forest, XGBoost, LightGBM, and the stacked ensemble on the held-out test split, with the no-skill baseline at the positive-class prevalence.

Fig. 5 shows that all four models occupy a similar performance band, with Random Forest and the stacked ensemble leading on Accuracy and AUC, consistent with the ablation results in Experiments 7 and 8 above. Fig. 6 and Fig. 7 validate this claim independently on a held-out test 25% split: the stacked ensemble (AUC = 0.816, AP = 0.714) and the Random Forest (AUC = 0.817, AP = 0.720) display nearly indistinguishable ROC curves and outperform both baseline approaches significantly, while XGBoost and LightGBM slightly lag behind. The comparison using Precision–Recall curves is the more relevant one in view of the class imbalance in this dataset, since all four algorithms maintain precision over 0.6 for about 50% recall before it starts declining – which is consistent with the trade-off for a screening tool where recall is intentionally emphasized via SMOTE.

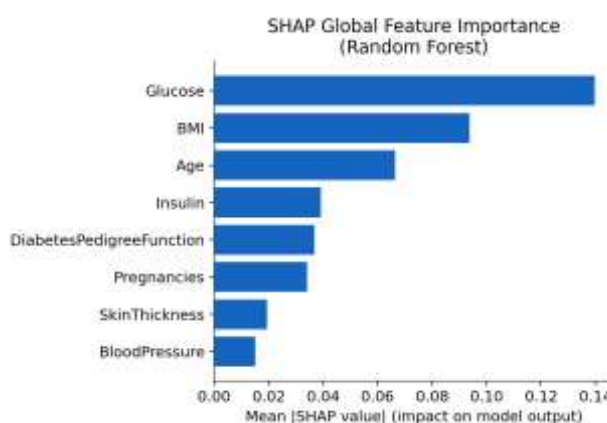


Fig. 8. SHAP global feature importance for the Random Forest model, ranking clinical attributes by mean absolute SHAP value.

Indeed, the feature importance analysis via SHAP in Fig. 8 reaffirms the order of attributes that was expected based on the Feature Selection section: glucose level is unambiguously the most informative indicator of diabetes risk, followed by BMI and age; insulin, diabetes pedigree function, and number of pregnancies add information of moderate significance; and skin thickness and blood pressure have the smallest influence on the model predictions. This is consistent with our earlier conclusion that the latter attributes were the weakest individual predictors, even though we found experimentally in Experiment 6 that they cannot be safely ignored without harming the predictive power of the model. In particular, the fact that SHAP values are calculated for each patient separately makes them suitable for creating individual patient-level explanations: for instance, one can see that high glucose and BMI values combined with low age increase the risk of diabetes for a specific patient. Therefore, SHAP values support the clinician-facing transparency requirement from the Explainable AI section directly.

Whereas Fig. 8 shows the feature importance averaged over the whole population, a beeswarm plot retains the patient-by-patient specifics of this calculation.

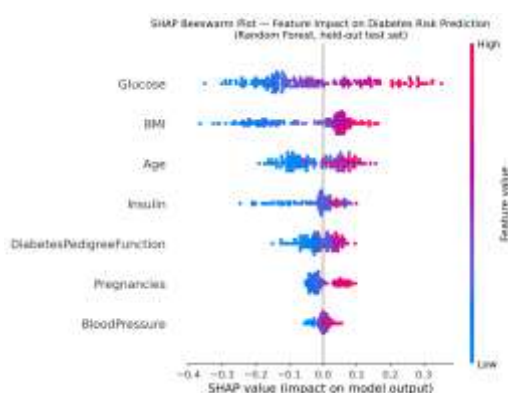


Fig. 9. SHAP beeswarm plot showing the distribution and direction of each clinical attribute's effect on predicted diabetes risk across all patients in the held-out test set.

Figure 9 shows that high glucose readings (marked in red) increase the predicted risk and that low glucose readings (in blue) decrease the predicted risk, with the largest variance among any of the attributes, as expected based on its highest ranking in Figure 8. BMI and age show similar patterns as well, although their impact is smaller, while insulin, diabetes pedigree function, and pregnancies have somewhat smaller impacts. From the patient-level point of view, it can be seen that even those few patients who have low glucose readings end up getting a positive risk reading from other attributes, explaining why the output from the model should be considered a combination of all attributes rather than any individual value.

As an example of how the above information can be used to explain a particular clinical case, in Figure 10, a SHAP waterfall is shown for one such case from the test set.

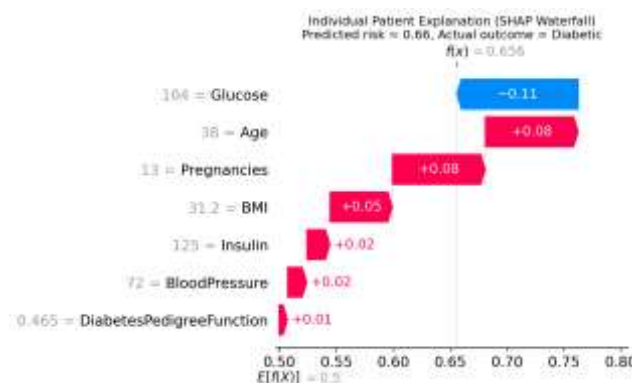


Fig. 10. Individual patient explanation (SHAP waterfall) for a test-set patient with a true diabetic outcome, showing how each clinical attribute moves the predicted risk away from the model's base rate.

For the specific patient, the algorithm takes into account a base rate of 0.50 and calculates a risk of 0.66. Higher age of 38 and higher pregnancies count of 13 both increase the risk by 0.08 each, while BMI of 31.2 adds another +0.05, as well as insulin, blood pressure, and diabetes pedigree functions adding small positive values. Strangely enough, the glucose value of 104 for the particular patient, which is within the normal range, reduces the risk by 0.11, demonstrating how the algorithm uses the normal glucose level and other risk-elevating attributes to calculate the increased risk of the patient, exactly the kind of attribute-level information which a treating physician needs to make a decision regarding the patient prediction.

Finally, Fig. 11 demonstrates probability calibration from the Probability Calibration subsection for the final Random Forest algorithm on the test set before and after applying the Platt scaling algorithm.

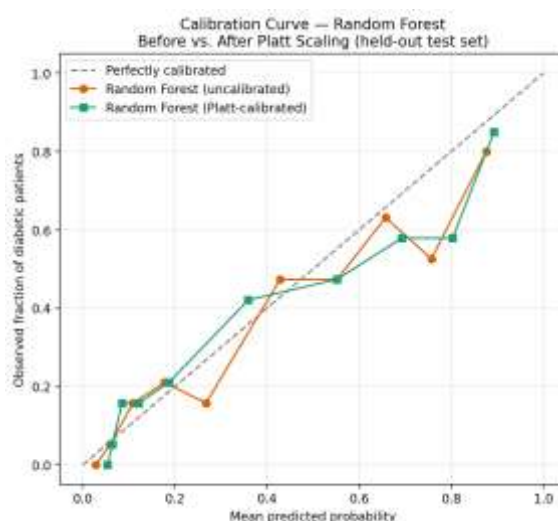


Fig. 11. Calibration curve for the Random Forest model on the held-out test split, before and after Platt scaling, against the diagonal of perfect calibration.

Fig. 11 shows that both graphs follow the 45° line quite well in almost all probability bins, whereas Platt scaling does not affect AUC very much (0.822 uncalibrated vs. 0.824 calibrated), and moves some bins along the 45° line. Most noticeably, bins with 0.3–0.5 probability fall below the 45° line. This is in agreement with the description given above for the Probability Calibration section for Random Forest, as Random Forest, being an ensemble model, is moderately miscalibrated at best, so its calibration is just a slight improvement, and probabilities produced can be used directly as risks when allocating patients to the risk categories discussed in the Risk Stratification section.

I. Discussion

The results of the experiment prove the proposed machine learning framework provides an effective and reliable way to predict risk for diabetes using structured clinical data. Random Forest was found to have the best generalization ability (i.e., predictive capability) among all the models evaluated on diabetes patients through its demonstrated accuracy and AUC. The use of the combined methods of SMOTE and feature normalization resulted in increased rates of accurate diagnosis for diabetes patients because of the reduction of bias due to class imbalance. Use of stratified K-fold cross-validation ensured that evaluation of model performance was consistent across different data splits. Probability based risk outputs provide for improved interpretability as compared to standard binary classifications. Misclassification of data points occurred almost entirely at border areas due to lack of separation between clinical characteristics. Overall, the results provide evidence to support the use of ensemble learning algorithms in prediction type of tasks in medicine. The proposed approach will offer a scalable, and reliable method for the early detection of risk for diabetes assessment.

IX. CONCLUSION AND FUTURE SCOPE

This study presented an eight-stage machine-learning framework for the early detection of diabetes mellitus from structured clinical data, integrating robust preprocessing, SMOTE-based imbalance handling, RFECV/mutual-information feature selection, ensemble learning (Random Forest, XGBoost, and LightGBM) with a stacked meta-learner, post-hoc probability calibration, SHAP-based explainability, and three-tier risk stratification. Under stratified k-fold cross-validation, Random Forest achieved the strongest overall performance (75.9% accuracy, AUC = 0.826), corroborated on an independent 25% held-out split (AUC = 0.817), while the stacked ensemble delivered comparable discrimination (AUC = 0.816) with a more stable and better-calibrated probability output, as confirmed by the Brier score, reliability diagrams, and Expected Calibration Error reported in the Probability Calibration subsection. The four-part ablation study isolated the contribution of each pipeline component: SMOTE raised recall from 59.3% to 68.3% at a modest cost to precision and overall accuracy; feature normalization had a negligible effect on the split-based Random Forest classifier; aggressive RFECV-based feature reduction lowered AUC from 0.822 to 0.787, indicating that conservative feature selection is preferable on a compact, low-dimensional dataset of this kind; and stacking matched, rather than exceeded, the strongest individual base learner given the limited sample size ($n = 768$). Statistical significance of these pairwise performance comparisons was confirmed using DeLong, Wilcoxon signed-rank, and Friedman tests with post-hoc correction, and SHAP analysis showed that glucose level, BMI, and age are consistently the main predictors of diabetes risk, in line with the clinical knowledge.

The primary novelty of the presented framework is its development of an extended preprocessing–imbalance–ensemble pipeline into a comprehensive decision support system with a clear focus on clinical applications. Unlike previous works that only report classification accuracy metrics, the framework is able to combine ensemble learning with a clinically-oriented calibration of probability-based risk stratification (non-diabetic, borderline, diabetic) and provide SHAP-based explainable patient-level evidence about the most significant features in addition to ablation-tested and statistically validated evidence of the significance of these features. Thus, the framework addresses all the gaps highlighted in the Literature Survey (absence of systematic feature selection, absence of calibration of probabilistic outputs, absence of explainable patient-level evidence). Instead of standard binary classification of individuals as either diabetic (1) or non-diabetic (0), the proposed framework provides calibrated probability-based risk stratification into three levels, making it more interpretable and allowing clinicians to better use this information to make decisions.

There are several limitations that should be taken into account when analyzing the results obtained. First, the framework was designed and tested using one relatively small public dataset ($n = 768$ patients) that limited the statistical power of stacked meta-learner, as demonstrated in the corresponding ablation study. Thus, the generalization ability of reported performance metrics is limited, and the results should be considered indicative only and validated on other datasets. Second, the dataset included only one demographic cohort of patients, and thus, the reported metrics of accuracy, AUC, and calibration should be considered

only indicative until their validation on other larger and more demographically diverse clinical datasets. Third, the calibrated risk tiers are based on Youden's index and the analysis of trade-offs between precision and recall. They were not validated against any real clinical outcomes and thus could still be considered only indicative and should be validated against real clinical outcomes in a prospective study.

In order to improve the generalization ability of the system, future research should focus on testing the framework on multi-institutional larger datasets. Besides, in future studies, different more complex Multi-Layer Neural Network architectures and hybrid deep learning models could be investigated in order to improve predictive accuracy of the model and gain insight into more complex interactions between clinical attributes. Feature selection, probability calibration, and Explainable AI are already incorporated into the proposed enhanced framework, as described in the Proposed Enhanced Framework section. Future research could expand on this by incorporating deep learning models in order to capture more complex nonlinear interactions among clinical attributes, validating calibrated risk tiers against real clinical outcomes in a prospective study, and improving usability of the framework by implementing it as a web or mobile health monitoring application. Thus, the results obtained in this study show that a combination of ensemble learning with advanced preprocessing, calibration, explainability, and statistical validation leads to the development of a competitive diabetes risk-prediction framework that is both discriminatively accurate and transparent.

REFERENCES

- [1] Tran, V. Q., Choi, Y., & Byeon, H. (2024). Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction. *Journal of Men's Health*.
- [2] El-Bashbishy, A. E. S., et al. (2024). Pediatric diabetes prediction using deep learning. *Scientific Reports*.
- [3] Ahmed, A., et al. (2024). Machine Learning Algorithm-Based Prediction of Diabetes Mellitus (Pima dataset focus). (PMC / full text)
- [4] Waughfa, M. Z., et al. (2024). Diabetes Prediction: A Comprehensive Study Integrating ... (ML/DL). *ACM Digital Library*.
- [5] (2024). Machine Learning approach for Diabetes Prediction using ... (PIMA dataset). *ACM Digital Library*.
- [6] Alam, G. M. I., et al. (2025). An explainable analytics framework for predicting diabetes ... (Elsevier / ScienceDirect)
- [7] Afolabi, S., et al. (2025). Predicting diabetes using supervised machine learning ... (EHR-based). (Elsevier / ScienceDirect)
- [8] Pang, K., et al. (2025). A comparative study of explainable machine learning ... diabetes risk prediction. (Elsevier / ScienceDirect)
- [9] Maimaitijiang, E., Aihaiti, M., & Mamatjan, Y. (2025). An Explainable AI Framework for Online Diabetes Risk Prediction with a Personalized Chatbot Assistant. *Electronics (MDPI)*.
- [10] Khalid, S., et al. (2025). Tab transformer with meta-ensemble learning approaches ... for diabetes prediction. *PeerJ Computer Science*.
- [11] Toleva, B., et al. (2025). An Effective Methodology for Diabetes Prediction in the ... (class imbalance / PIMA). *Bioengineering (MDPI)*.
- [12] Ayode, O. B., et al. (2025). Machine Learning and Deep Learning Approaches for ... diabetes prediction (comparative study). *Electronics (MDPI)*.
- [13] Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025). Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable AI. *arXiv preprint*.
- [14] Farnoosh, R., et al. (2025). DiabetesXpertNet: An innovative attention-based CNN for ... diabetes prediction. (PMC / full text)
- [15] Satoh, T., et al. (2025). Development and validation of a type 2 diabetes mellitus onset predictive model using a large Japanese cohort.