SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Cross-Validated Greedy Stacking Ensemble with RFECV-Based Feature Selection for T2DM Prediction

Rizwan Akhtar[1], Muhammad Kalamuddin Ahamad[2]

[1]Department of Computer Application, Integral University, Uttar Pradesh, India. E-mail: rizwanakhtar360@gmail.com
[2]Department of Computer Application, Integral University, Uttar Pradesh, India. E-mail: mohdkalam@iul.ac.in

Abstract:
Type 2 Diabetes Mellitus (T2DM) is a significant health burden and causes a considerable burden of disease and mortality, especially in South Asians. This research introduces an Explainable Cross-Validated Greedy Stacking Ensemble (CV-GSE) technique for predicting T2DM on the Type-2 Diabetes Dataset Bangladesh (T2DDB), which contains a set of 1065 patient records containing demographic, anthropometric and biochemical features. The proposed framework first uses Recursive Feature Elimination with Cross-Validation (RFECV) to find an informative subset of predictors and then tests eight heterogeneous machine-learning classifiers via stratified 10-fold cross-validation with SMOTE only on the training folds. Out-of-fold (OOF) probability predictions are made for each candidate classifier and classifiers are added to the final model in a greedy forward fashion, based on OOF ROC-AUC. Then a tuned XGBoost meta-classifier is used to combine the selected base learners. A dual-layer explainability approach is also provided in which SHapley Additive exPlanations (SHAP) is used for feature and base-learner attribution and Local Interpretable Model-Agnostic Explanations (LIME) is used for instance-level explanations. In total, RFECV selected 6 informative predictors (Age, BMI, SBP, DPF, Insulin, and Glucose) with maximum cross-validated ROC-AUC of 0.9558. The stacking ensemble proposed in this study obtained 98.50% accuracy, 98.25% precision, 99.70% recall, 98.62% F1-score, and 96.85% ROC-AUC, which were better than the evaluated baseline classifiers. The outcomes show the capability of the proposed method to predict T2DM using feature selection, heterogeneous ensemble learning, OOF-based greedy learner selection, and Explainable Artificial Intelligence (XAI). But, before clinical applicability can be established, there needs to be external validation, with independent and diverse datasets.

Keywords: Type 2 Diabetes Mellitus (T2DM), Explainable Artificial Intelligence (XAI), RFECV, Cross Validation (CV), SHAP, LIME

1. Introduction

Diabetes mellitus (DM) is a serious issue for public health, because of its rising prevalence and associated long-term complications. It is a serious health problem because of its high socio-economic burden, long-term health consequences and growing prevalence at a rapid pace. Insulin resistance and reduced insulin secretion are the main characteristics of this disease. This can lead to the body experiencing: chronic hyperglycaemia (high blood sugar), which may lead to life-threatening issues such as cardiovascular disease, nephropathy, neuropathy, retinopathy and premature death [1,2]. The International Diabetes Federation (IDF) reports that diabetes prevalence is on a steady upward trend globally, indicating the need for dependable early prediction models to aid in early diagnosis, preventative treatment and effective disease management [3]. There are 2 kinds of DM. The aetiology of type 1 diabetes mellitus (T1DM) is the loss of pancreatic β-cells, which results in an incapacity to promptly lower blood glucose levels. Insulin resistance and insufficient insulin production are the pathophysiologies of T2DM, also known as non-insulin dependent diabetes [4].

Early identification of individuals at higher risk is crucial to reduce the morbidity and overall health burden of DM [5]. Older age, physical inactivity, obesity, impaired glucose regulation, family history of diabetes, dyslipidaemia, hypertension and cardiovascular disease are some of the demographic, metabolic, and clinical factors associated with an increased risk of diabetes. The recognition of these risk factors can help with early screening, prevention and timely disease management. In the last few years, with the development of Data Mining techniques, the prediction of diseases is becoming more accurate; the Data Mining algorithms are now capable of recognizing complex patterns in clinical data that might be hard to spot with traditional statistical techniques [6,7]. There are various supervised ML algorithms, such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), Naı̈ve Bayes (NB), and Artificial Neural Network (ANN) have been demonstrated to be effective as a means to predict diabetes [8,9]. However, the prediction ability of these approaches is highly reliant on features used as inputs that are of high quality and meaningful. The addition of extra, irrelevant, or less useful features can lead to higher computational

complexity, lower model interpretability, and impact the accuracy and generalizability of predictions [10-11].

The selection of features has become an important part of intelligent clinical decision-support systems [12]. The goal of feature selection is to select the most relevant predictors in the prediction of T2DM instead of including all the available ones [13]. Selecting a concise, informative feature subset can improve computational efficiency, strengthen model robustness, reduce overfitting risk, and enhance clinical interpretability by focusing on the most relevant predictors of T2DM [14]. RFECV is one of the most effective feature selection methods based on wrapper, which selects the best feature subset [15]. RFECV algorithm can be used to remove the least informative features while also testing the performance of the model using CV, which helps minimize feature selection bias and stabilizes the features selected to the model [16]. Moreover, XAI can quantify the contribution of each feature, which can be used to interpret predictive models. [17]. In particular, SHAP offer a theoretically sound approach to explain model predictions and evaluate the clinical significance of selected features [17,18]. Although the diabetes prediction field has been extensively studied, most studies have been limited to maximizing prediction accuracy from laboratory-based datasets, such as PIDD [19]. Fewer studies have systematically explored clinical datasets based on symptoms, like the Early-Stage Diabetes Risk Prediction Dataset (ESDRPD), which use clinical and demographic data that are observed rather than biochemical measurements, for prediction purposes [20,21]. These are the typical datasets used to predict T2DM. Such datasets are especially valuable for the development of non-invasive, low-cost screening systems which can be used in primary health care and resource-limited environments [19,20]. Furthermore, there are many published works that have been conducted based on the Split the data once into a train and test set, followed by feature scaling or feature selection prior to CV. This can result in information leakage and optimistic bias, which can lead to estimates of model performance. Those that are greater than the true capacity [21].

This study aims to create an explainable and CV-GSE model for the early prediction of T2DM. The proposed framework combines RFECV, various supervised machine learning (ML) classifiers, out-of-fold (OOF) prediction generation, greedy forward model selection, and stacking ensemble learning to enhance the reliability and predictive power of T2DM classification. The clinical features that are informative are identified using RFECV and the model performance is robustly estimated using in-fold SMOTE to handle class imbalance. Several heterogeneous classifiers are tested, and the OOF probability predictions of these classifiers are then fed into a greedy forward selection (GFS) algorithm to determine a set of base classifiers that complement each other in terms of predictive performance. Finally, the selected classifiers are combined using a stacking architecture and an XGBoost meta-learner to take advantage of the complementary predictive information between the selected classifiers. Finally, SHAP and LIME are added to give global and local explanations of the model predictions. The proposed framework is tested with individual baseline classifiers based on accuracy, precision, recall, F1 score and ROC-AUC to evaluate the effectiveness of predictions and generalization ability.

The key contributions of the research:

1. To find and confirm the most informative clinical parameters for the prediction of T2DM using RFECV approach.
2. To design and compare various supervised ML classifiers for prediction of T2DM using stratified 10-fold cross validation with in-fold SMOTE.
3. To make OOF predictions and use greedy forward model selection to select a small and complementary set of base classifiers for ensemble learning.
4. To build a CV-GSE with the selected base classifiers and a meta-learner XGBoost, and test its prediction accuracy against the individual classifiers and other existing methods.
5. To improve the interpretability of T2DM prediction by combining SHAP and LIME for global feature-level and local instance-level explanations of model predictions.

The proposed framework offers an explainable, CV-GSE solution for the prediction of T2DM. The framework starts with data preprocessing and then uses the RFECV algorithm to select a subset of clinical predictors that are informative and validated. The selected features are then fed into a variety of supervised ML classifiers for training and testing. The model is evaluated using 10-fold stratified CV, SMOTE is used for class imbalance in the training folds, but never oversampling in the validation folds.

OOF predictions are generated for each base learner to capture the complementary predictive strengths of the candidate classifiers. These OOF predictions are then passed to a GFS procedure, which adds a classifier only when it improves the selected evaluation criterion. This process removes unnecessary or redundant learners and yields a compact set of complementary models. The selected base learners are subsequently integrated into the proposed CV-GSE, where an XGBoost classifier serves as the meta-learner and learns from their predictions.

XAI techniques are embedded at a global and local level to enhance the transparency of the proposed framework. SHAP quantifies the contribution and importance of features and model predictions, while LIME is used to provide instance-level explanations for individual predictions of T2DM. Therefore, the

proposed framework aims to not only attain competitive prediction performance but also to gain interpretable insights into the factors that influence T2DM risk prediction.

Organization of the Paper. The remainder of this paper is structured as follows:

1. **Section 2** reviews current data mining and machine-learning approaches for T2DM prediction and identifies the research gaps addressed in this study.
2. **Section 3** describes the proposed framework, including the T2DDB dataset, preprocessing, RFECV-based feature selection, heterogeneous base-classifier development and evaluation, OOF prediction generation, diversity-based learner selection, stacking ensemble construction, and stratified 10-fold cross-validation with in-fold SMOTE.
3. **Section 4** presents and discusses the experimental results, including RFECV-based feature selection, baseline and ensemble performance comparisons, OOF-based greedy learner selection, SHAP- and LIME-based explainability analyses, and comparison with existing studies.
4. **Section 5** summarizes the main findings and methodological contributions, discusses the limitations of the proposed framework, and outlines directions for future research.

2. Related Works

T2DM is a growing global health concern, leading to extensive research on ML and data-mining methods for early prediction and classification. Studies have examined demographic, anthropometric, clinical, and biochemical variables to identify individuals at high risk of diabetes using supervised learning algorithms. Commonly used classifiers include LR, DT, RF, SVM, k-NN, NB, GB, and ANN. Comparative studies show that ML models can achieve strong predictive performance; however, classifier effectiveness depends on the dataset, feature representation, preprocessing strategy, and evaluation metrics. This variability indicates that no single classifier is universally optimal for all T2DM datasets or clinical objectives [22]. Earlier work comparing seven ML classifiers on the PIDD found that SVM achieved strong discriminative performance, while RF and XGBoost also produced competitive results [23]. These findings support the shift from evaluating individual classifiers toward systematic ensemble learning. Ensemble methods have become important in diabetes prediction because they combine the strengths of multiple classifiers and capture complex, nonlinear relationships among clinical variables. Common ensemble approaches include RF, GB, AdaBoost, XGBoost, LightGBM, and Extra Trees (ET) [24]. Although ensembles can improve performance, combining many classifiers does not necessarily produce a better model [25-27]. When classifiers generate highly correlated predictions, they may become redundant and add little complementary information. Therefore, learner diversity is essential for constructing an effective ensemble. Stacking offers one way to address this issue by passing the predictions of multiple base classifiers to a meta-classifier, which learns how to combine their complementary predictive information [28]. However, effective stacking also requires systematic selection of a diverse and informative set of base learners. Feature selection is another key component of ML-based T2DM prediction, as irrelevant or redundant predictors can increase model complexity, reduce interpretability, and raise the risk of overfitting. Traditional feature-selection strategies include filter, wrapper, and embedded methods. Among these, recursive feature elimination (RFE) is a widely used wrapper-based approach that removes less informative predictors in an iterative manner. RFECV extends RFE by using cross-validation to identify the feature subset that delivers the best predictive performance [29]. The application of RFECV is especially relevant to clinical prediction because it can decrease the number of predictors without losing variables with the highest contribution to discrimination. In the current study, the two-stage feature-selection approach is used, selection of relevant feature by RFECV. The first seven predictors age, BMI, systolic blood pressure, diastolic blood pressure, diabetes pedigree function, insulin and glucose were then selected by RFECV to form the final six features: age, BMI, systolic blood pressure, diabetes pedigree function, insulin and glucose. Another challenge in the healthcare prediction is the class imbalance, which can lead to the majority class being given more weight in the ML algorithms, thereby affecting their ability to detect observations from the minority class. Oversampling techniques, especially the Synthetic Minority Over-sampling Technique (SMOTE), have thus been extensively adopted to enhance representation of the minority classes. The use of oversampling, however, needs to be carefully controlled, as it is possible to introduce information leakage between the training and validation data sets by applying SMOTE prior to partitioning data [30,31]. It's more appropriate to apply oversampling only to the training set of each CV fold [32]. In this current study, SMOTE is integrated in the ML pipeline and is used when performing the stratified 10-fold CV, thus the minority class is represented in the model training process through a synthetic representation and the original observations used for validation will be retained for the performance evaluation. The evaluation strategy is also crucial for the reliable estimation of predictive performance. Single train–test splits are often employed, but can lead to unstable estimates, especially in small datasets or if they contain class imbalance. Stratified k-fold CV is more robust method of evaluation as it repeats the partition of the data set with class distribution across folds [33]. However, if feature selection, preprocessing or other data-dependent operations are performed on the whole dataset prior to splitting the data into folds, CV alone is not sufficient to prevent information

leakage. This leakage can lead to unrealistic predictions of predictive performance [34]. In the current setting, 10-fold CV is performed to test the eight base classifiers and OOF predictions are created to construct the ensemble. These OOF predictions guarantee that each observation is assigned a prediction from a model that was not used during the fold of training that it belongs to. This serves as a proper ground for the assessment of prediction relationships between the candidate classifiers and building the stacking feature space. OOF predictions also allow for the exploration of model diversity prior to the construction of the ensemble. In the present study, instead of keeping all eight classifiers just because of their individual predictive power, the correlation between the OOF predictions of the classifiers is calculated and a predefined threshold on the correlation is used to determine redundant learners. The classifiers in the list are the eight classifiers that were tested as candidates: SVM-RBF, RF, DT, XGBoost, KNN, ET, NB, and LR. The SVM and RF were selected as complementary base learners in the diversity-based selection procedure. Finally, their OOF predictions were provided to a tuned Random Forest meta-classifier to build the final stacking ensemble [35]. This is unlike the traditional ensemble methods which simply merge models without taking into account the prediction redundancy. The goal is to build an ensemble that is compact and that contains classifiers that provide complementary predictive information, instead of very similar predictions. Interpretability is also becoming a more important aspect of clinical prediction-based data mining. While high predictive accuracy might be enough for some applications, in healthcare, the clinician and other stakeholders would like to know what variables affect the decision-making process of the model, and how patient-specific characteristics contribute to a particular prediction. The importance of XAI techniques, like SHAP and LIME, has therefore become significant in medical data mining research [36]. Unlike LIME, which can explain individual observations at the local level, SHAP can quantify the contribution of individual features to model predictions and offer global and model-level interpretations. The present study uses both to give complementary interpretations of the proposed framework. To explore the relative importance of features, SHAP was used to rank them, with glucose being the top feature and BMI, systolic blood pressure, age, diabetes pedigree function, and insulin following in importance, and LIME was used to visualize the contribution of individual feature values to the prediction for each individual [37]. Although significant advances have been made in data mining-based prediction of T2DM, there are still some methodological gaps. First, many studies have concentrated on finding the best-performing classifier, although there are indications that different classifiers can exhibit different performance with respect to different performance measures. Second, feature selection is typically not dealt with in the context of ensemble construction, and the quality of the selected representation of the features can directly affect the performance and interpretability of the downstream models. Third, when multiple classifiers are highly correlated, stacking can be redundant, and it is also necessary to select the models with diversity before stacking [38]. Fourth, class-imbalance handling needs to be part of the CV procedure to minimise the risk of information leakage. Lastly, while some diabetes prediction methods are emerging that integrate feature optimization, heterogeneous classifiers evaluation, OOF-based diversity assessment, stacking, and explainability, there is still a need for a single methodological pipeline that incorporates all of these components [39-41].

This study presents an ensemble Explainable Cross-Validated Greedy Stacking (CV-GSE) framework to overcome the drawbacks of the existing T2DM prediction methods, which combines XAI, heterogeneous supervised ML, stratified 10-fold CV, OOF prediction generation, greedy forward learner selection, and stacking ensemble learning with feature selection using RFECV. First, eight different classifiers are tested with stratified 10-fold CV that only applies SMOTE to the training folds. The resulting OOF probability predictions are then fed into a GFS algorithm to select a small set of complementary base learners. The selected learners are then stacked with XGBoost as the meta-learner. SHAP and LIME are integrated with the model to better explain the model at the global level and at the instance level, respectively, for better model transparency. The proposed framework is systematically compared with individual baseline classifier based on accuracy, precision, recall, F1-score and ROC-AUC. Therefore, the study combines feature selection, CV learner selection, ensemble learning, robust model evaluation and explainability into one data mining system for predicting T2DM.

Table1. Summarized related studies from literature

SN	Reference	Methodology	Research Focus	Key Findings	Limitations
1	Kaliappan et al. [42]	Filter/wrapper feature selection + multiple ML classifiers + stacking	Feature-selection and classifier performance across diabetes datasets	Demonstrated that feature selection substantially influences classification and stacking performance	Uses multiple datasets and does not specifically select complementary learners using OOF prediction correlation

2	Z. Zhang et al. [43]	Harmony Search + feature selection + stacking	Optimized feature and base-learner selection for diabetes prediction	Joint optimization of feature selection and base-learner combinations improved stacking performance on PIDD	Uses metaheuristic optimization rather than direct OOF prediction-correlation analysis
3	Mohsen et al. [44]	Multimodal deep learning model integrating ECG features	Integration of ECG with traditional models	Enhanced prediction accuracy with ECG data	Potential overfitting with complex models
4	Mohanty et al. [45]	SHAP-based feature selection + multiple ML/ensemble models	Feature importance and computationally efficient diabetes prediction	SHAP-based selection reduced the feature space while maintaining or improving predictive performance	Uses SHAP for feature selection rather than RFECV followed by diversity-based learner selection
5	Lee et al. [46]	AI-based prediction model	Aging population with comorbidities	Improved prediction accuracy in older adults	Requires validation across different age groups
6	Abnoosian et al. [47]	Ensemble machine learning (RF, SVM)	Diabetes disease prediction using model combination	Demonstrated the potential of ensemble learning for improving diabetes classification	Does not specifically address diversity-based selection of base learners according to prediction correlation
7	K. O. et al. [48]	Feature engineering + XGBoost, NGBoost, Bagging, LightGBM, AdaBoost + stacking + SHAP	Stacking and explainability for diabetes prediction	Demonstrated the effectiveness of combining multiple classifiers and using SHAP for interpretation	Does not explicitly use prediction-correlation thresholds to remove redundant base learners
8	V. Q. Tran et. Al. [49]	FT-Transformer + RF + XGBoost + LightGBM + stacking + SHAP	Explainable stacking for diabetes prediction	Stacking with a Random Forest meta-classifier produced strong predictive performance and SHAP-based interpretation	Population restricted to men; learner diversity is not explicitly evaluated using OOF prediction correlations
9	Khokhar et al. [50]	Machine learning with explainable AI (SHAP, LIME)	Model interpretability	High accuracy with strong interpretability	Potential bias in feature selection
10	I. A. Elgendy et al. [51]	Seven base ML models + stacking + multiple meta-models + SHAP	Ensemble prediction and multi-level explainability	Demonstrated improved prediction through stacking and provided explanations of model and	Uses several base learners without the specific compact, correlation-based learner-selection strategy proposed in your study

				feature contributions	

In conclusion, the literature shows a trend toward the development of integrated T2DM prediction models, emphasizing the importance of selecting relevant features, thorough validation, the use of ensemble learning, and interpretability. In this direction, the current research introduces a CV greedy stacking model that merges informative features and complementary classifiers to boost the prediction performance of T2DM while simultaneously offering interpretable insights into the factors affecting the classification.

3. Materials and Methods

3.1 Overview of Proposed Framework

The proposed architecture as shown in Fig. 1 describe the explainable CV-GSE framework for prediction of T2DM. The framework combines data preprocessing, feature selection, heterogeneous ML models, cross-validation, OOF prediction, greedy model selection, stacking ensemble learning, performance evaluation, and explainability in a single workflow. The overall architecture is composed of five layers: Data Layer, Feature Layer, Model Layer, Performance Evaluation Layer, and Explainability Layer. The Data Layer is the first layer of the framework and the Type 2 Diabetes Dataset Bangladesh (T2DDB) is used to develop and test the prediction models. The raw data is fed into the Feature Layer to be pre-processed and feature optimized. The Feature Layer first locates the target variable, and transforms it to a binary representation that is appropriate for the classification task T2DM. Any missing data is addressed with appropriate imputation techniques and categorical variables are coded into numerical values. Pre-processed data is then used to select an informative set of predictors using Recursive Feature Elimination with Cross-Validation (RFECV). RFECV recursively tests the importance of candidate features by a Random Forest estimator and cross-validation, thus shrinking the feature space and keeping the features that are useful for discriminative information for further modelling. The Model Layer assesses a wide variety of eight supervised ML classifiers, including SVM-RBF, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), XGBoost, Naı¨ve Bayes (NB), Extra Trees (ET), and KNN. Heterogeneous classifiers allow the framework to be able to learn different patterns and decision boundaries, instead of relying on a single modelling approach. Stratified 10-fold CV is used to evaluate the classifiers, which preserves a similar class distribution between the folds. To overcome the problem of class imbalance, SMOTE is performed on training data in each CV fold, which prevents information leakage from the validation data. During the CV process the cross-validated probability predictions of the candidate classifiers are also produced as OOF predictions. Each of these OOF predictions is a prediction for the individual training observation and can be used to inform the next model-selection step. The framework does not simply add all the candidate classifiers together but rather use a greedy forward selection algorithm to choose a small number of base classifiers. The classifiers are added in a progressive manner starting with an initial candidate model, based on the contribution to the cross-validated ROC-AUC. A candidate classifier is kept if adding it to the ensemble improves the ensemble's predictive accuracy, otherwise it is discarded. This process leads to an efficient ensemble of complementary predictive information, while not unnecessarily increasing the number of base learners. The base classifiers selected are then combined using the proposed CV-GSE. The selected heterogeneous classifiers in the first level of the stacking architecture make probability predictions. The predictions are then passed as meta-level inputs to an XGBoost meta-learner that learns to use the information produced by the base models. Therefore, it is a two-level learning strategy, where the first level is composed by heterogeneous classifiers and the second level is done by XGBoost. The Performance Evaluation Layer is used to systematically evaluate classifiers and the proposed CV-GSE. Various classification metrics are used to assess the models, these measures complement the overall predictive performance, positive class identification and discrimination capability. To assess the improvement and robustness of the stacking framework in predictive performance, the results of the proposed stacking framework are evaluated against the individual baseline classifiers.

Finally, the Explainability Layer incorporates SHAP and LIME to improve the interpretability of the ML predictions. SHAP is used to assess the contribution and relative importance of the selected clinical features, providing a global view of the factors influencing model predictions. LIME complements this analysis by generating local explanations for individual observations and showing how specific feature values contribute to each prediction. Together, these methods provide both global and instance-level insight into the model's decision-making process.

3.2 Dataset Description:

The study was based on a Type-2 Diabetes Dataset Bangladesh (T2DDB) containing 1,065 patients and 10 variables: nine predictor variables and 1 binary outcome variable. Table 2 shows the dataset used for T2DM classification, which includes demographic, anthropometric, physiological, and biochemical data. The target variable "Type-2 Diabetic" was converted to a binary outcome with the diabetic and non-diabetic groups. The data set had 840 cases of T2DM (78.87%) and 225 cases of non-T2DM (21.13%), which was a

significant class imbalance that was mitigated during model training using SMOTE in the cross-validation pipeline. Predictor variables are shown in table 2.

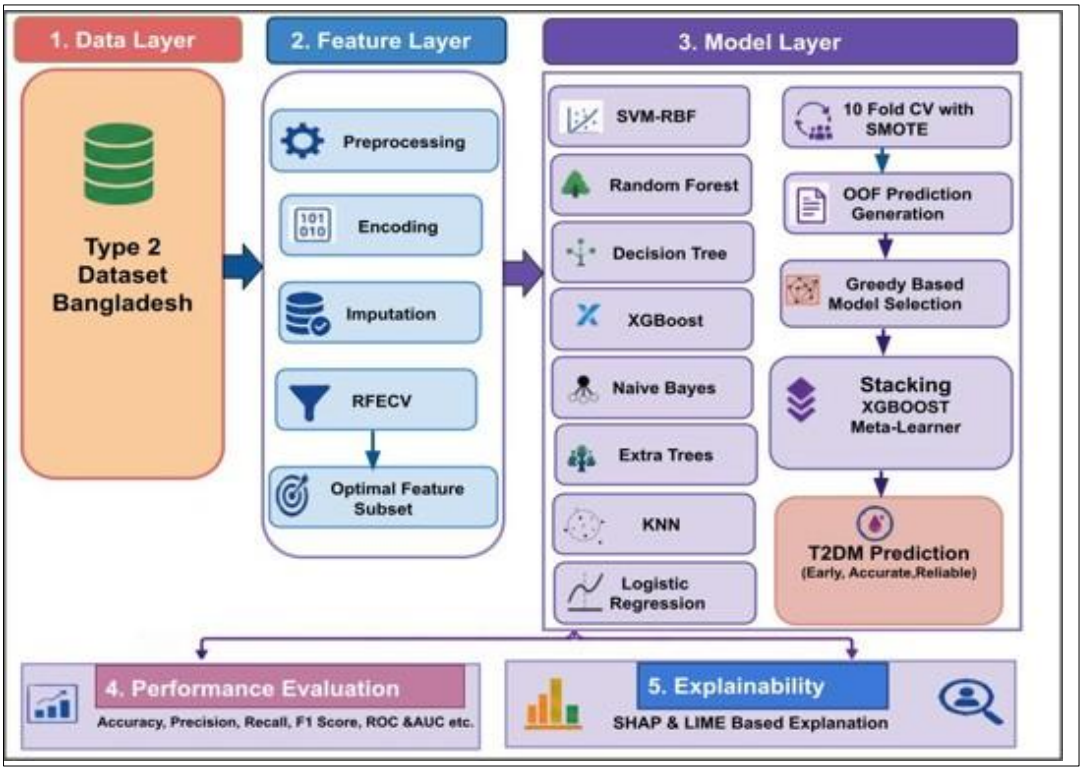

Figure 1: CV-GSE framework for Early T2DM Prediction

Table 2. Description of the Type-2 Diabetes Dataset Bangladesh

No	Variable	Type	Description
1	No. of Pregnancy	Numerical	No of pregnancies reported by the individual.
2	Age	Numerical	Age of the individual in years.
3	BMI	Numerical	BMI representing the individual's weight relative to height.
4	BP (Systolic)	Numerical	Systolic blood pressure measurement.
5	BP (Diastolic)	Numerical	Diastolic blood pressure measurement.
6	Diabetes Pedigree Function	Numerical	A measure representing the estimated hereditary predisposition to diabetes.
7	Insulin	Numerical	Insulin measurement recorded for the individual.
8	Skin Thickness (mm)	Numerical	Skin-fold thickness measurement in millimetres.
9	Glucose	Numerical	Blood glucose measurement used as an important metabolic indicator for T2DM.
10	Type-2 Diabetic	Class	Diabetes status (Positive = 1, Negative = 0)

Preprocessing involved identifying and binarizing the target variable, imputing missing numerical values using the median, and encoding categorical variables where applicable. RFECV was then applied to select the optimal feature subset for training and evaluating the eight ML classifiers.

3.3 Dataset Preprocessing:

The dataset was systematically pre-processed before model development to ensure consistency and suitability for binary T2DM classification. The "Type-2 Diabetic" variable was designated as the outcome variable, while the remaining variables were retained as predictor features. The outcome was recoded into two classes: Class 0 for non-T2DM and Class 1 for T2DM. Missing values were handled according to variable type: numerical missing values were imputed using the median, whereas missing categorical values were labelled as "Unknown" and subsequently encoded numerically using label encoding. After preprocessing, the analytical dataset comprised 1,065 observations and 10 variables, including nine predictors and one binary outcome variable.

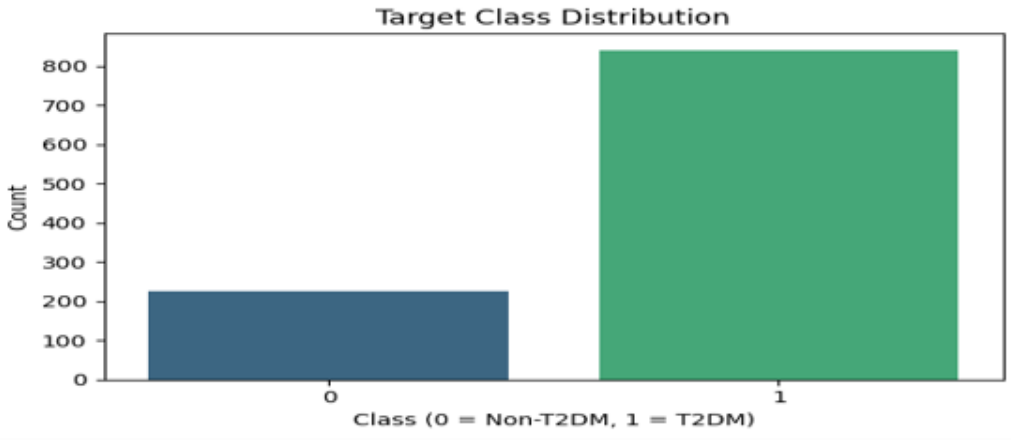

Figure 2: Distribution of T2DM and Non-T2DM Classes

Fig 2 illustrates the target-class distribution after data preprocessing. Class 0 denotes non-T2DM cases (n = 225), whereas Class 1 denotes T2DM cases (n = 840), resulting in a total analytical sample of 1,065 observations.

3.4 Recursive Feature Elimination with Cross-Validation (RFECV)

To reduce the number of predictors used in the prediction of T2DM, a parsimonious subset was identified using feature selection by RFECV. The model performance was then assessed using 10-fold optimization criterion of ROC-AUC, with least informative, with the least informative features being eliminated recursively from the model via RFECV. This data-driven method was used to eliminate the redundant features and ensure the inclusion of the most relevant features for the predictive performance.

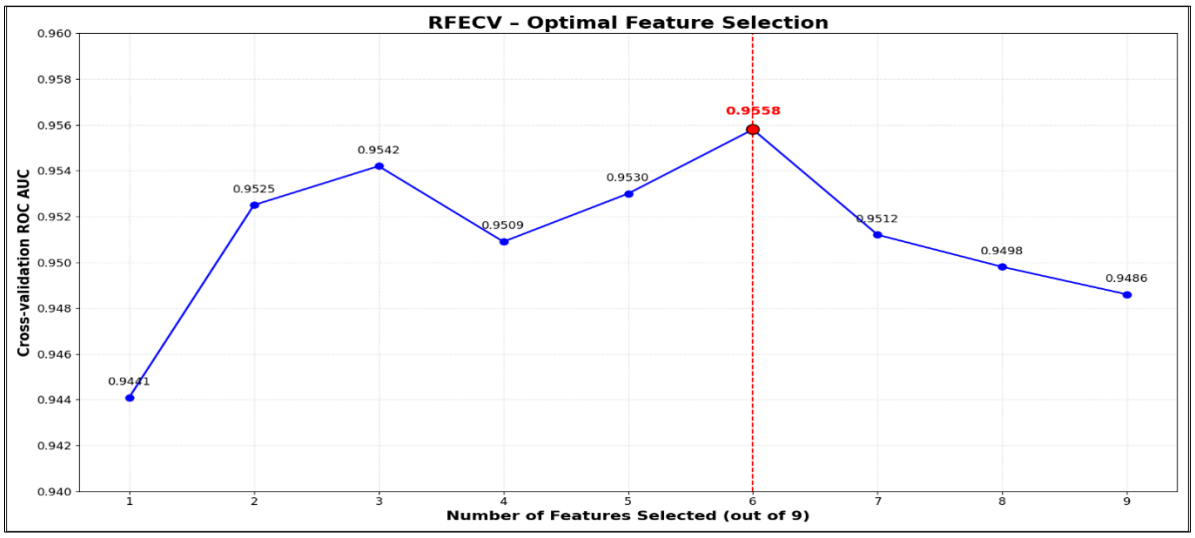

Figure 3: RFECV-based optimal feature selection

The best feature subset was selected by RFECV using the CV predictive performance. The best cross-validated ROC-AUC was obtained using 6 features with value of 0.9558 as indicated in Fig 3. Age, BMI, BP (Systolic), Diabetes Pedigree Function, Insulin, and Glucose were selected by RFECV and BP (Diastolic) was eliminated by recursive feature elimination. Therefore, the final six-predictor feature set was utilized for further prediction of T2DM and construction of the proposed stacking framework.

$$S^* = \underset{S \subseteq F}{arg\,max} \left[\frac{1}{K} \sum_{k=1}^{K} AUC_k(S) \right] \quad (1)$$

where Frepresents the complete feature set, Srepresents a candidate feature subset, $K = 10$represents the number of cross-validation folds, and $AUC_k(S)$denotes the ROC-AUC obtained using feature subset Son the k-th fold. S^*represents the optimal feature subset.

3.5 Machine Learning Models

In order to get various predictive patterns for T2DM classification, eight different supervised ML classifiers were selected as candidate base classifiers, namely Support Vector Machine with RBF kernel (SVM-RBF), Random Forest (RF), Decision Tree (DT), XGBoost, Naïve Bayes (NB), Extra Trees (ET), KNN and Logistic Regression (LR). These classifiers belong to various paradigms of learning: linear classification, kernel-based learning, single-tree learning, bagging based ensembles and boosting based ensembles. Heterogeneous classifiers allow exploring various decision-making patterns and mitigate reliance on a

single modelling approach. Stratified 10-fold CV was used to evaluate each classifier, in which the observations were split into ten folds, with the class distribution of T2DM and non-T2DM observations approximately preserved in each fold. The data set is imbalanced, so the synthetic minority over-sampling technique (SMOTE) is added to the training stage of each fold in the CV. Therefore, synthetic minority class observations were only created from the training data, and the corresponding validation fold was not touched, which minimized the risk of information leakage. To address the issue of feature scale sensitivity, feature standardization was added to the modelling pipeline for classifiers that are sensitive to feature scale.

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \qquad 0 \leq \lambda \leq 1 \qquad (2)$$

where x_i represents a minority-class sample, x_{nn}represents one of its nearest minority-class neighbours, and $\lambda$is a random interpolation factor between 0 and 1. Each candidate classifier was evaluated based on accuracy, precision, recall, F1-Score and ROC-AUC. CV method was employed to create OOF probability predictions for each candidate classifier in addition to performing the performance evaluation. The following observations are not used to train the corresponding model and are instead used as input for the proposed model-selection and stacking stages, which are referred to as OOF predictions. The resulting OOF predictions are then used as a common ground to find the base learners that are useful for providing predictive information to the final ensemble.

3.6 Cross Validated Greedy Stacking Ensemble

The proposed framework does not directly stack all the candidate classifiers, but rather uses an OOF greedy forward selection approach to build a small stacking ensemble. Stratified 10-fold CV was used to create probability predictions for each base learner, which were then used as meta-level information to progressively assess the contribution of each base learner. ROC-AUC was used as the main selection criterion since it is a threshold-independent measure of the discriminatory ability of the models. For every iteration, one unselected classifier was temporarily included in the current model subset, and the performance of the updated model subset was assessed using cross-validated ROC-AUC. The greedy forward selection strategy was evaluated by cross-validated ROC-AUC and at each step the classifier with the largest improvement was added to the model as shown in fig 4. RF had the best ROC-AUC score of 0.9584 in the first attempt and was chosen as the first base learner. LR further boosted the ROC-AUC to 0.9613 and Extra Trees to 0.9710. For the fourth iteration, the ROC-AUC of 0.9722 was the highest observed with the inclusion of DT. The classifiers did not yield any further improvements at the fifth iteration, and so the process of selecting the classifiers was stopped. The final selected subset consisted of RF, LR, ET and DT. The final combination achieved an absolute ROC-AUC score improvement of 0.0138 over the initial selected learner (from 0.9584 to 0.9722). The results show that the greedy selection strategy was able to find a set of complementary learners that was compact, but not too large, and that did not include unnecessary classifiers. This is a progressive improvement which means that the classifiers chosen added some predictive power to the ensemble, but the other classifiers did not add enough predictive power to the ensemble to warrant their inclusion. The probability predictions of the four base learners were then concatenated together to form the meta-feature matrix.

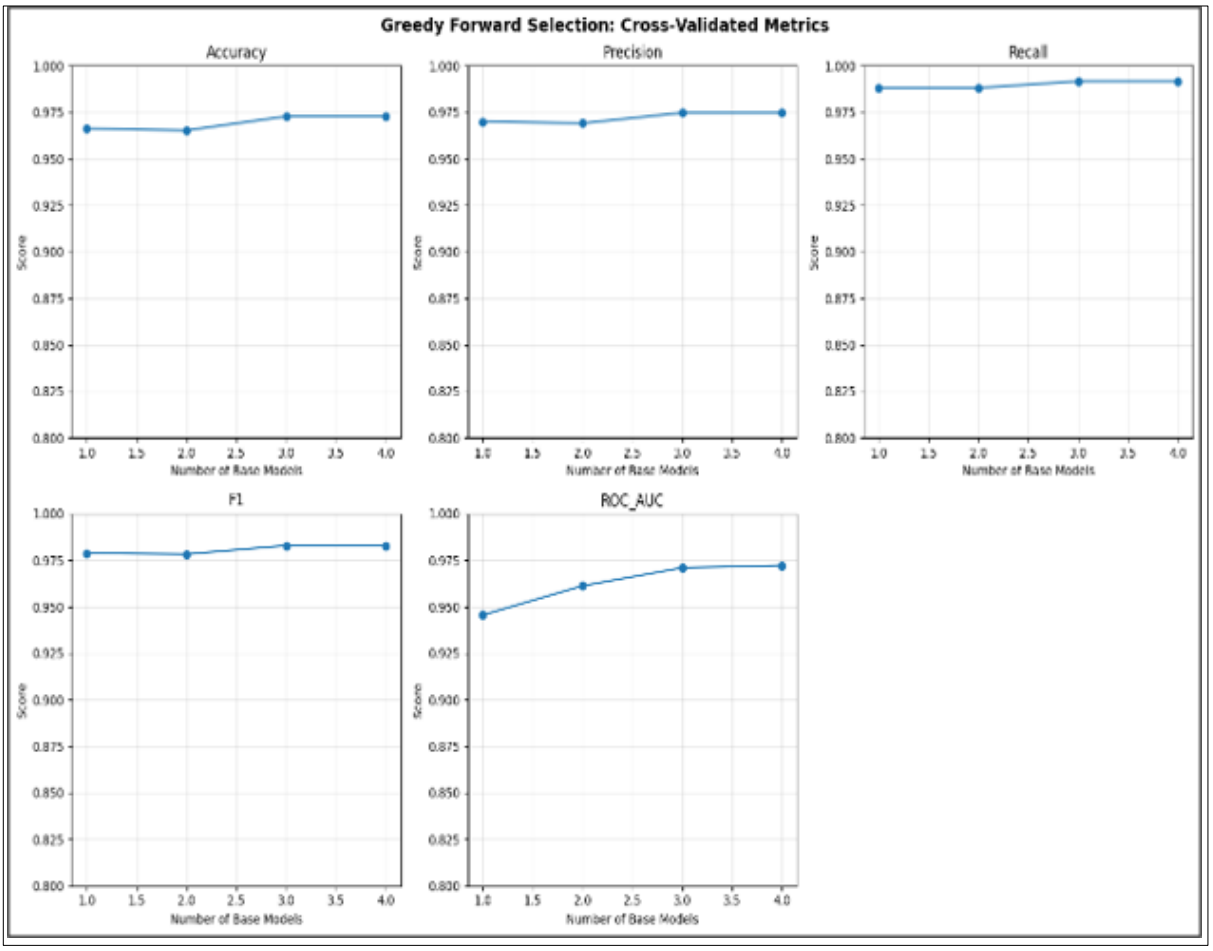

Figure 4. Greedy forward selection of base classifiers based on ROC-AUC.

These meta-features were passed on to a meta-learner (XGBoost) that learns the relationship between the predictions of the selected classifiers and the T2DM outcome.

$$\hat{y} = g_{XGB}(p_{RF}, p_{LR}, p_{ET}, p_{DT}) \quad (3)$$

where g_{XGB}represents the XGBoost meta-classifier and $y$represents the final predicted T2DM outcome. This architecture is a two-level stacking structure, in which the heterogeneous classifiers are used to make the first-level prediction and XGBoost is used to make the second-level prediction. The proposed ensemble is called Cross-Validated Greedy Stacking Ensemble (CV-GSE). The proposed CV-GSE framework is summarized in Algorithm 1 and comprises feature selection, training the base-learners using cross-validated data, generating OOF predictions, selecting the most useful learners by a greedy approach, stacking, evaluation and explainability.

Algorithm 1: Cross-Validated Greedy Stacking Ensemble (CV-GSE)

Input: Type 2 Diabetes Dataset $D = X, Y$

1. Load and preprocess the T2DM dataset to obtain feature matrix $X$and target vector Y.
$$D = \{(x_i, y_i)\}_{i=1}^{N},\ y_i \in \{0, 1\}$$
2. Apply RFECV to identify the optimal feature subset based on maximum cross-validated ROC-AUC.
$$S^* = arg\ \max_{S \subseteq S_0} ROC\ -\ AUC_{CV}(S)$$
3. Define eight candidate classifiers:
$$\mathcal{L} = \{SVM\ -\ RBF, RF, DT, XGBoost, NB, ET, KNN, LR\}$$
4. Train each classifier using stratified 10-fold cross-validation, applying SMOTE only to the training folds and standardization within the pipeline.
$$D = \bigcup_{k=1}^{10} D_k$$
5. Generate out-of-fold (OOF) probability predictions for each classifier.
$$P_j^{OOF} = [p_{1j}, p_{2j}, \ldots, p_{Nj}]$$
6. Select base learners greedily according to the highest improvement in ROC-AUC.
$$L_j^* = arg\ \max_{L_j \in \mathcal{L} \setminus L_s} [AUC(L_s \cup \{L_j\}) - AUC(L_s)]$$

7. Stop the selection process when no additional learner improves ROC-AUC.
$$\Delta AUC = AUC_{new} - AUC_{current} \leq 0$$
8. Construct the meta-feature matrix using OOF predictions of the selected learners.
$$Z = [P^{OOF}_{L_1}, P^{OOF}_{L_2}, \dots, P^{OOF}_{L_m}]$$
9. Train XGBoost as the meta-classifier using the generated meta-features.
$$M^* = arg \max_{M \in XGBoost} AUC_{CV}(M; Z; Y)$$
10. Generate the final T2DM prediction using the trained stacking model.
$$\hat{Y} = M^*(Z)$$
11. Evaluate the proposed framework using Accuracy, Precision, Recall, F1-score, and ROC-AUC.
12. Apply SHAP and LIME to explain the predictions of the final model.

Output: Optimized features, selected base learners, trained CV-GSE model, performance metrics, and explainability results.

Algorithm 1 summarizes the sequential operation of the proposed CV-GSE framework. The steps of the procedure include pre-processing, feature selection using RFECV with SMOTE limited to training folds, and then stratified 10-fold cross-validation of candidate classifiers. The greedy forward selection procedure uses OOF probability predictions for all candidate learners. Each iteration, the candidate that makes the best change in the chosen ROC-AUC value is added to the ensemble. When no learner can get any better on the criterion, selection ends. The predictions of the selected learners are then used in the construction of the meta-feature matrix that is fed into the XGBoost meta-classifier. The final model is tested using accuracy, precision, recall, F1-score, and ROC-AUC, and then SHAP and LIME analysis.

Lastly, the proposed CV-GSE was assessed and compared to the individual baseline classifiers within the same framework of cross validation based on performance metrics. This evaluation is used to consistently determine if the chosen mix of diverse learners results in better predictive accuracy and a smaller ensemble size.

3.7 Performance Evaluation

The predictive performance of each technique was evaluated before and after Stratified 10-Fold CV using five widely accepted evaluation metrics.

1. Accuracy

The percentage of correctly categorised samples is known as accuracy.
$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

2. Precision

The percentage of accurately anticipated diabetic cases among all expected diabetic cases is known as precision.
$$\text{Precision} = TP/(TP + FP) \quad (5)$$

3. Recall (Sensitivity)

Recall assesses the classifier's accuracy in identifying patients with diabetes.
$$\text{Recall} = TP/(TP + FN) \quad (6)$$

4. F1-Score

The harmonic mean of recall and precision is known as the F1-score.
$$F1 = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (7)$$

5. Area Under the ROC Curve (ROC-AUC)

ROC-AUC measures a classifier's ability to distinguish between diabetic and non-diabetic cases across classification thresholds. Higher ROC-AUC values indicate stronger class separability.

3.8 Explainability Layer

SHAP and LIME were used as complementary explainability techniques to improve the interpretability of the proposed T2DM prediction framework. To quantify the importance of the six RFECV-selected predictors, their contribution to model predictions was calculated using SHAP as shown in algo 8.
$$\boldsymbol{\phi}_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (8)$$

where Fdenotes the complete feature set, jdenotes the feature being explained, Srepresents a subset of features excluding j, f(S)denotes the model output using feature subset S, and $\boldsymbol{\phi}_j$represents the SHAP contribution of feature j.

It is used to quantify the contribution of the six RFECV-selected predictors to model predictions, and to obtain a global assessment of feature importance and the direction and magnitude of their effects. Furthermore, SHAP analysis was performed on the stacking model to understand the individual contribution of the selected base learners to the final prediction of the stacking model. LIME generated local, instance-level explanations showing how individual predictor values contributed to each T2DM prediction shown in algo 9.
$$\xi(x) = \arg\min_{g \in G} [L(f, g, \pi_x) + \Omega(g)] \quad (9)$$

where frepresents the original prediction model, grepresents an interpretable local surrogate model, $L(f, g, \pi_x)$measures the local approximation error, $\pi_x$represents the locality weighting around instance x, and $\Omega(g)$controls the complexity of the surrogate model.
Together, SHAP and LIME provided complementary global and local interpretability, improving transparency in the ensemble's decision-making process.

4. Results and Discussions

4.1 Comparative Performance of Base Classifiers and the Proposed Ensemble

The predictive performance of the eight baseline classifiers and the proposed stacking ensemble is presented in Table 3. The classifiers that performed best individually were: Random Forest (96.06%), XGBoost (96.02%) and Extra Trees (95.68%). The accuracy of Decision Tree and SVM (RBF) is 94.46% and 91.08% respectively while KNN, LR and NB obtained comparatively lower accuracy of 87.80%, 83.46% and 75.39% respectively.

Table 3: CV Performance of the Base Classifiers and Proposed CV-GSE

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
SVM (RBF)	91.08	97.05	91.55	94.17	94.06
Random Forest	96.06	96.66	98.45	97.53	95.84
Decision Tree	94.46	96.70	96.31	96.48	92.92
XGBoost	96.02	96.67	98.93	97.77	94.48
Naïve Bayes	75.39	95.22	72.50	82.11	89.08
Extra Trees	95.68	97.17	97.38	97.27	96.00
KNN	87.80	95.86	88.33	91.93	92.15
Logistic Regression	83.46	97.13	81.55	88.54	94.24
Proposed Cross Validated Greedy Stacking Ensemble (CV-GSE)	**98.50**	**98.25**	**99.70**	**98.62**	**96.85**

The proposed stacking ensemble achieved the highest accuracy (98.50%), precision (98.25%), recall (99.70%), F1-score (98.62%) and ROC-AUC (96.85%) among all eight classifiers. The proposed ensemble achieved 2.44 percentage points and 0.85 percentage point improvement in accuracy and F1-score respectively over the best individual classifier, namely, RF and XGBoost. Moreover, the ensemble outperformed the best individual classifier, Extra Trees, by 0.85 percentage points with ROC-AUC score of 96.00%. The high recall of 99.70% shows that the proposed ensemble successfully identifies the T2DM-positive cases while the F1-score of 98.62% is a good balance of precision and recall. Overall, the results showed that the stacking strategy of heterogeneous classifiers is better than any individual classifiers in predicting the results.

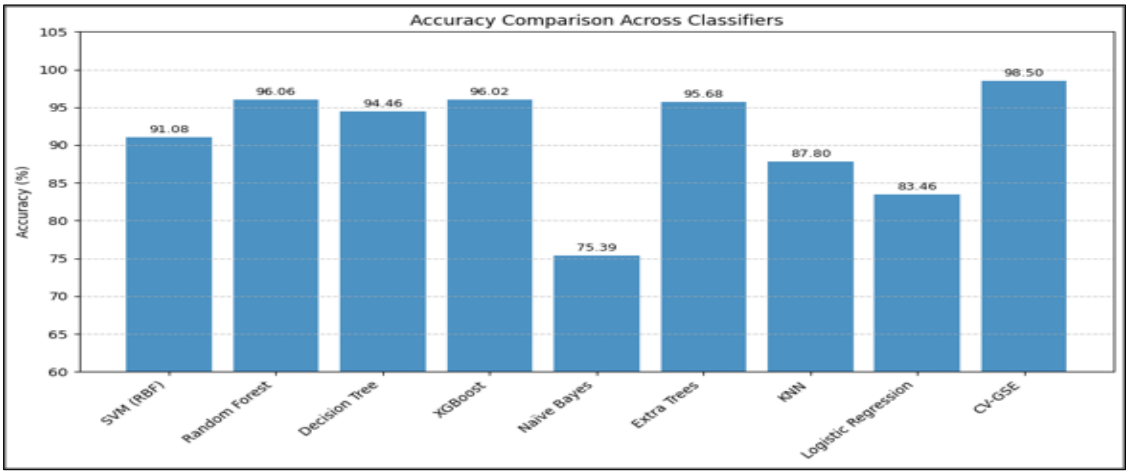

Figure 5(a): Accuracy of Base Classifiers and the Proposed Ensemble

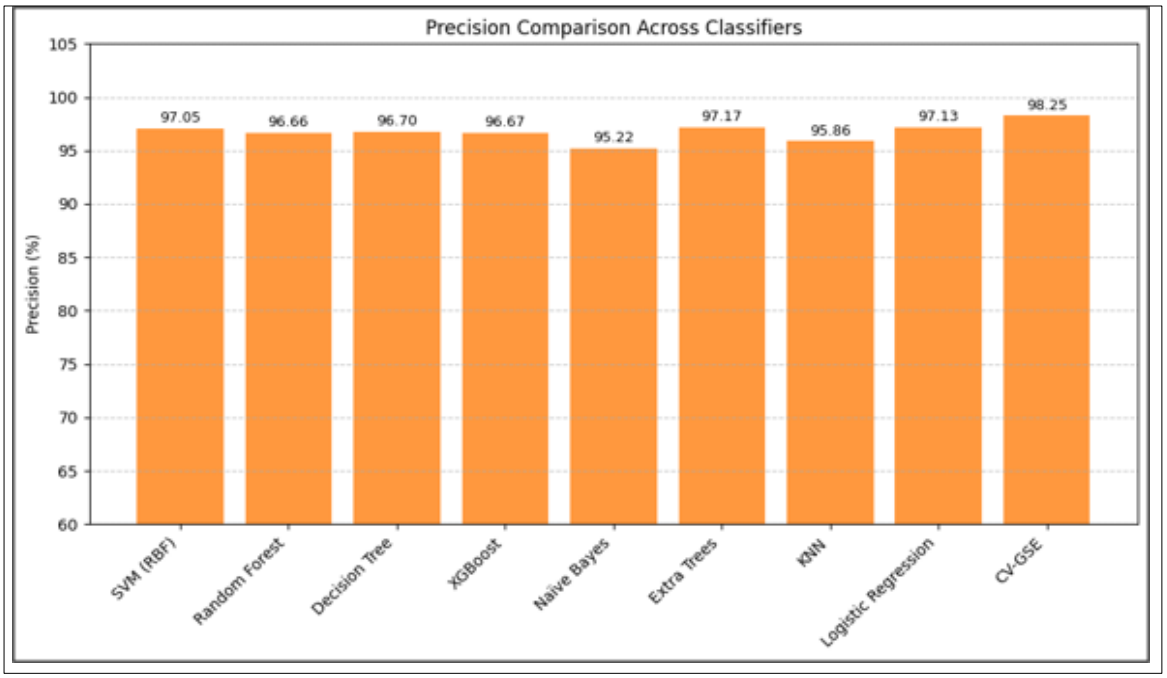

Figure 5(b): Precision of Base Classifiers and the Proposed Ensemble

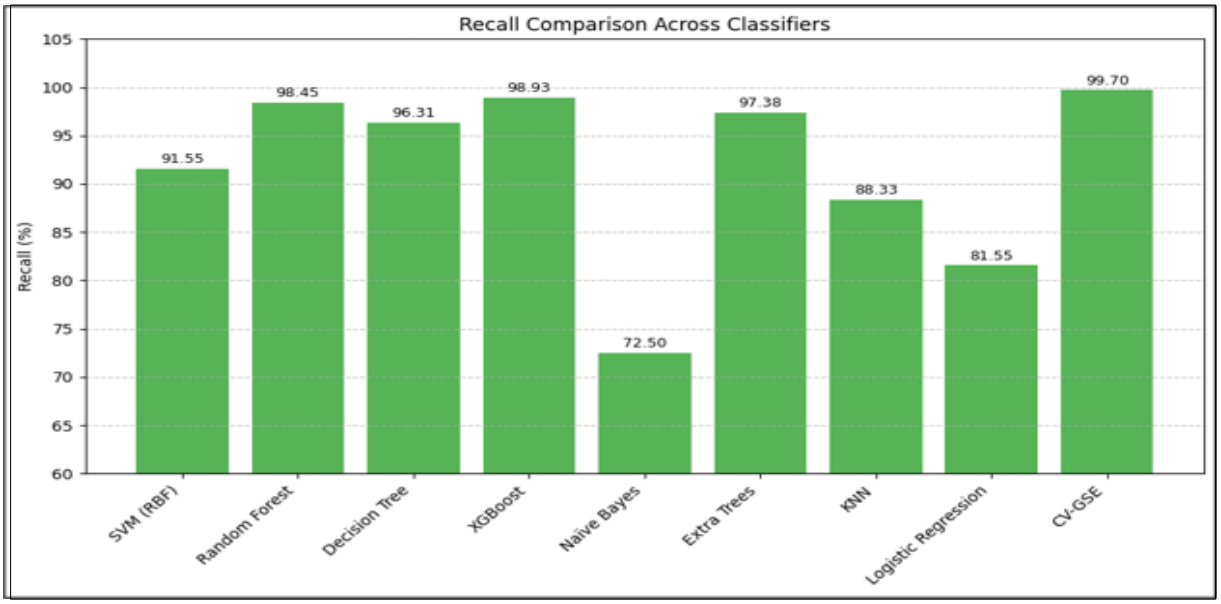

Figure 5(c): Recall of Base Classifiers and the Proposed Ensemble

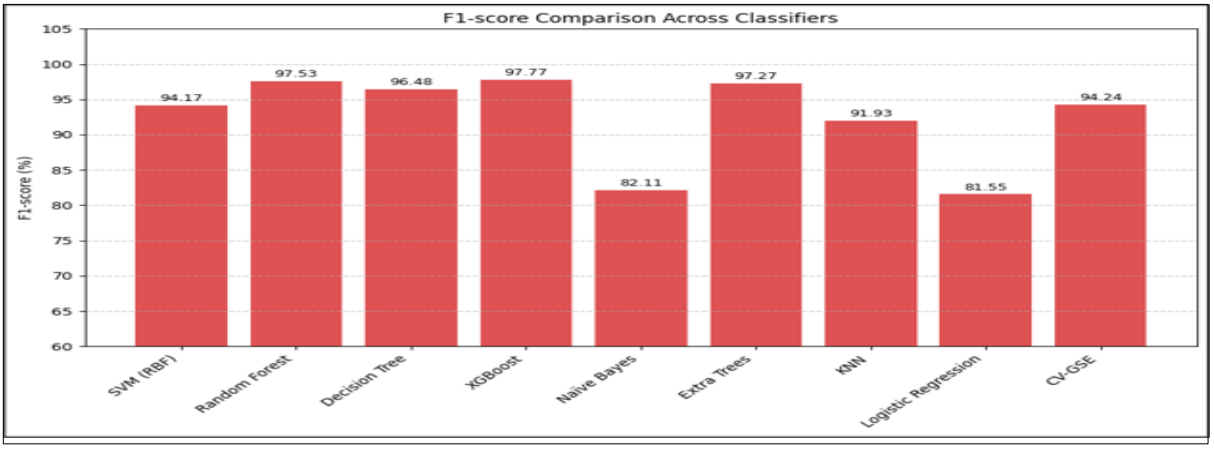

Figure 5(d): F1 Score of Base Classifiers and the Proposed Ensemble

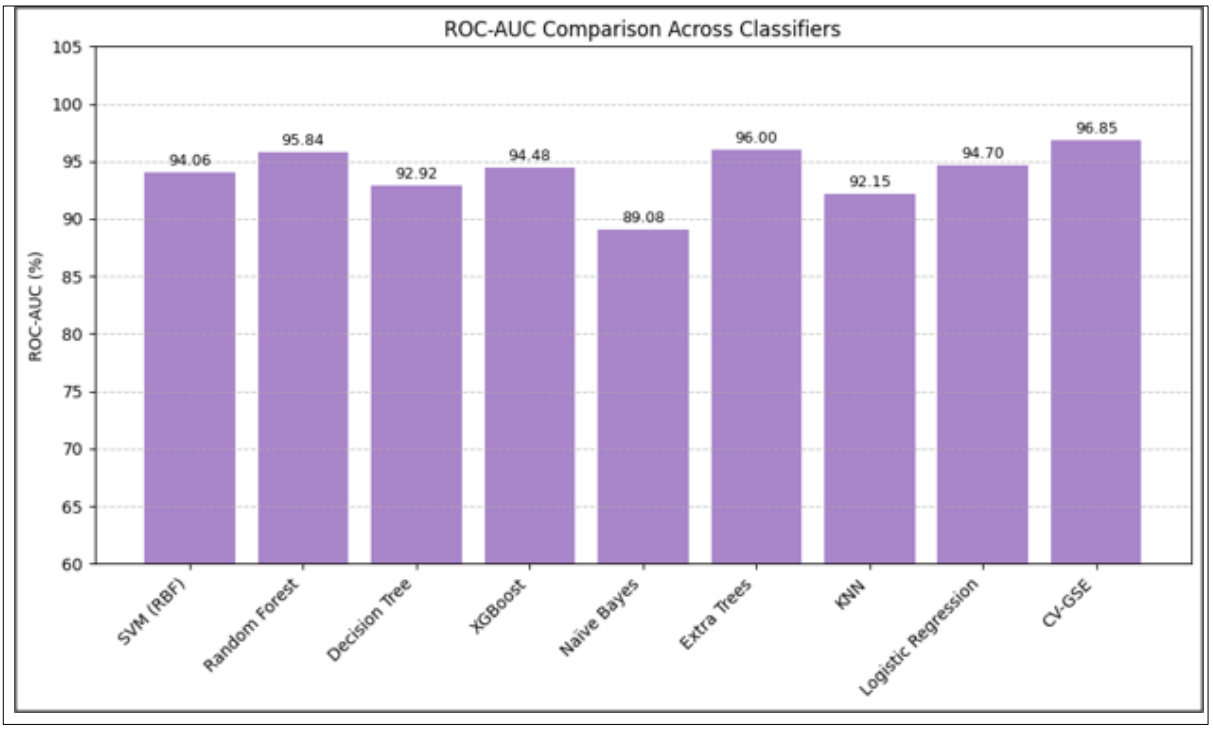

Figure 5(e): ROC-AUC of Base Classifiers and the Proposed Ensemble

Figures 5(a)–5(d) summarize the metric-level differences, while Figure 5(e) highlights the superior ROC-AUC of the proposed ensemble relative to the individual classifiers. Overall, these results indicate that CV-GSE improves the predictive and discriminative performance of T2DM classification by systematically integrating complementary predictions from heterogeneous base learners.

4.2 OOF-Based Greedy Base Learner Selection

A greedy selection approach using OOF was used to find a small and complementary subset of the 8 candidate classifiers. Stratified 10-fold CV was used for OOF probability predictions for each classifier. Classifiers were successively added to an empty ensemble, based on the contribution to the cross-validated ROC-AUC. The unselected classifier with the largest increase in the ROC-AUC was kept at each step. The process was stopped as soon as there was no more improvement in the classifiers. This strategy reduced learner redundancy while retaining complementary models with useful predictive information. The proposed CV-GSE was then constructed using the selected base learners.

4.3 Stacking Ensemble Performance

The OOF probability predictions of the selected base learners were fused to create the meta-feature representation. These meta-features were then given as input to an XGBoost meta-learner to make the final prediction of the T2DM. The results of the proposed CV-GSE and the eight base classifiers are shown in Table 3. The ensemble model outperformed the individual models with accuracy of 98.50%, precision of 98.25%, recall of 99.70%, F1-score of 98.62% and ROC-AUC of 96.85%. The proposed ensemble improved the best individual accuracy of RF (96.06%) by 2.44 percentage points. Likewise, it outperformed the best individual ROC-AUC (96.00%) of Extra Trees by 0.85 percentage points. The high recall rate of 99.70% highlights the high ability of the ensemble to detect T2DM-positive patients, and the F1 score of 98.62% shows a good balance between precision and recall. The results show the benefit of the proposed stacking strategy for integrating complementary OOF predictions.

4.4 Global Feature Importance (SHAP) The proposed CV-GSE framework was interpreted using SHAP analysis at two complementary levels. For the feature level, the six features selected by the RFECV were analyzed using XGBoost. The most important predictor as shown in Fig 6 is Glucose (5.887), followed by BMI (0.737), BP(Systolic) (0.637), Age (0.462), DiabetesPedigreeFunction (0.369), and Insulin (0.242) in terms of the mean absolute SHAP value. The contribution of Glucose was significantly higher, suggesting its major role in predicting T2DM, and BMI and SBP also gave useful predictive information.

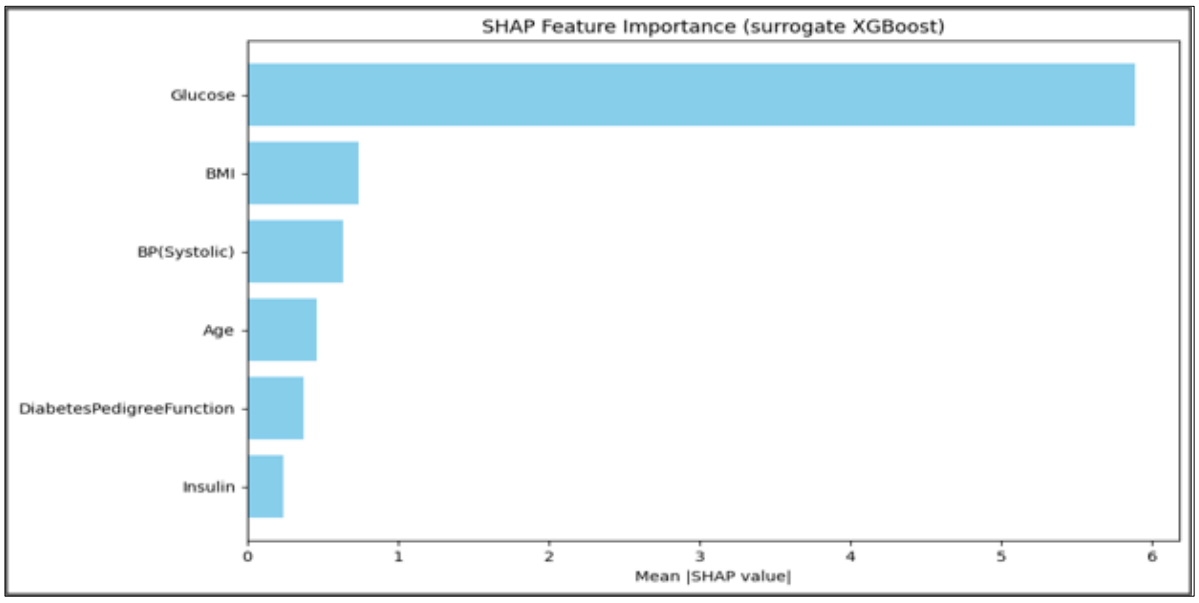

Figure 6: SHAP Feature Importance (Surrogate XGBoost)

The selected base learners' meta-level predictions were fed into the model level and SHAP analysis was applied. The ranges of SHAP contribution for Random Forest and Extra Trees are comparatively broad, as seen in Fig 7, suggesting that they have a greater impact on the final stacking decisions.

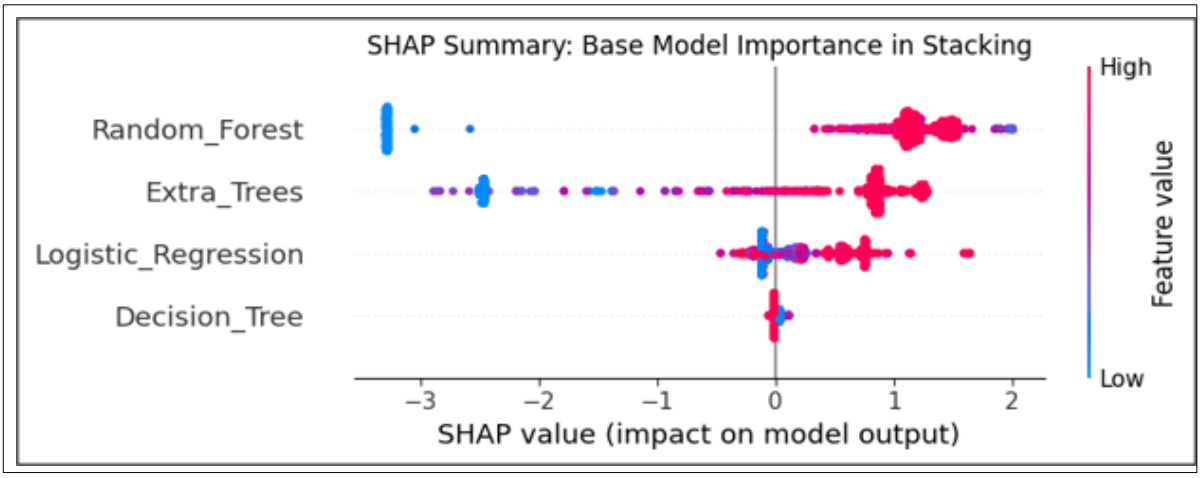

Figure 7: SHAP Importance of Base Models in the Proposed Stacking Ensemble

Logistic Regression also has significant positive and negative contributions, while Decision Tree has a relatively small contribution range. The positive and negative SHAP values show that the predictions made by the selected base learners can have a different effect on the individual ensemble decisions. In general, the SHAP analysis results indicate that the proposed stacking framework is able to achieve high predictive performance and provide information about the relative contribution of the clinical characteristics and the base-learner predictions.

4.5 Local Explainability Analysis Using LIME

The proposed stacking ensemble was applied to three representative cases and LIME was used to investigate patient-specific model reasoning. The local explanations derived from the above are shown in Fig. 8(a)–(c). The results of the analysis show that the individual clinical features contribute, and contribute differently, in various patient instances, thus complementing the global trends that are revealed by the SHAP analysis. The highest positive contribution was found for Glucose (0.2655), followed by SBP > 130 mmHg (0.0350), Age (44-53) (0.0158), and BMI ≤ 22.65 (0.0091) for Instance 0. In contrast, Insulin ≤ 0.00 (−0.0473) and Diabetes Pedigree Function ≤ 0.00 (−0.0073) contributed negatively. Glucose ≤ 80.00 had the greatest negative contribution (−0.6861) for Instance 1. Insulin ≤ 0.00 (−0.0374) and systolic blood pressure ≤ 110 mmHg (−0.0257) also contributed negatively. The DPF was a small positive contributor (0.0125), whereas BMI and age were minor contributors.

The highest positive contribution was with Glucose (0.1332) for Instance 2. BMI ≤ 22.65 also contributed positively (0.0080). Conversely, Insulin ≤ 0.00 (−0.0396), Diabetes Pedigree Function ≤ 0.00 (−0.0209), Age

≤ 35 years (−0.0139), and systolic blood pressure between 110 and 120 mmHg (−0.0058) contributed negatively.

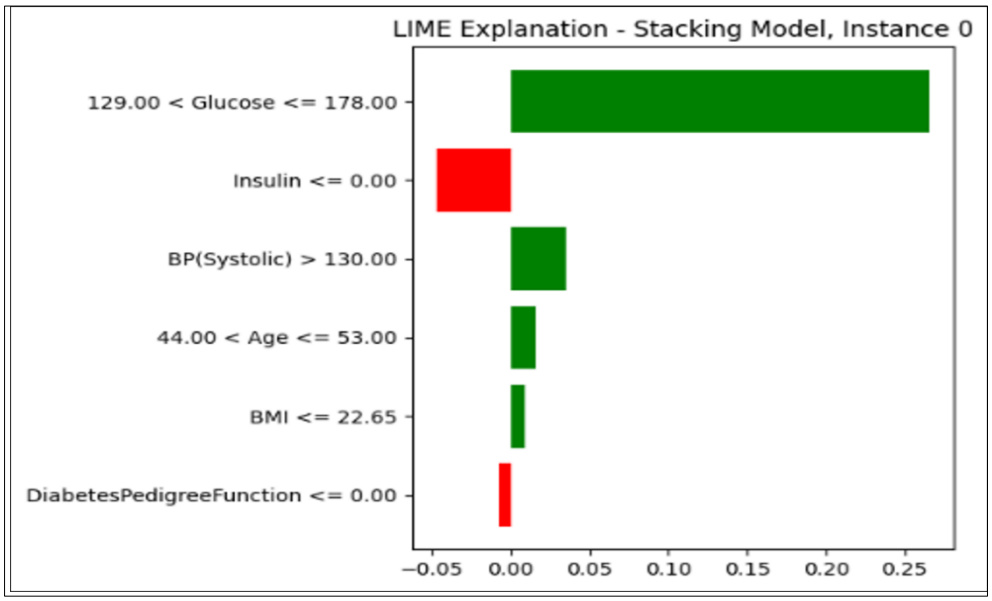

Figure 8(a): LIME Explanation, Instance (0)

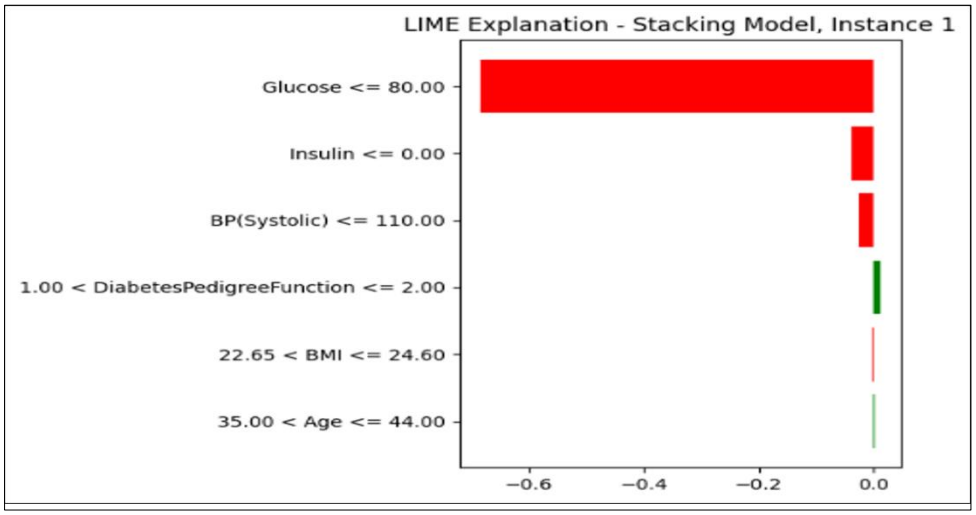

Figure 8(b): LIME Explanation, Instance (1)

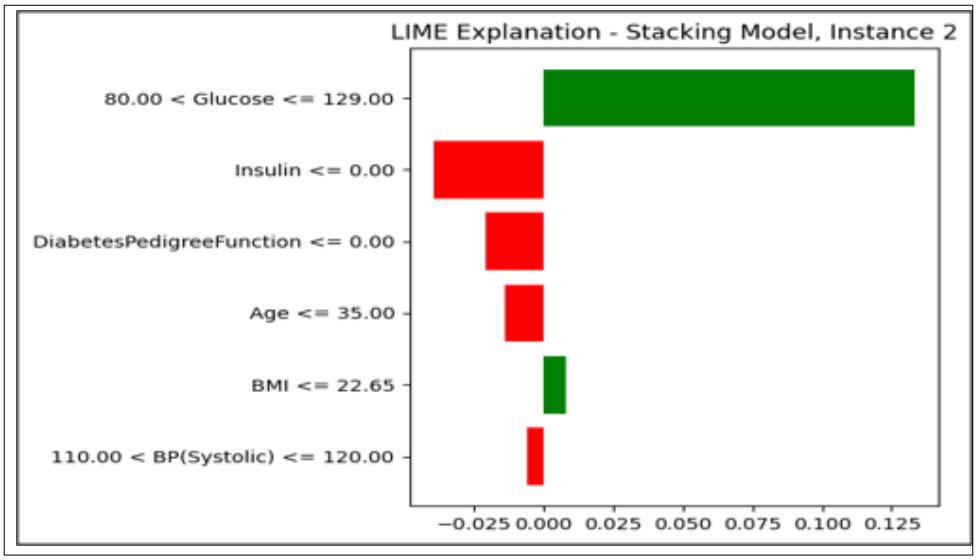

Figure 8(c): LIME Explanation, Instance (2)

Overall, the LIME results show that the proposed stacking model does not give equal importance to all the predictors, rather it makes instance-specific decisions. The direction and magnitude of the local contribution varied depending on the patient's characteristics but glucose always generated the largest local contribution in all three instances. This patient-level variability adds to the global SHAP analysis, and illustrates how global and local explainability methods can be used together to interpret T2DM prediction.

4.6 Comparison with Existing Literature

The performance of the proposed framework was evaluated with the diabetes prediction approaches reported in the literature as shown in Table 4. Studies using parameter optimized classifiers, voting based ensembles and stacking based classifiers are included in the comparison. The accuracy of the proposed framework is 98.50% while the selected comparative studies had accuracy values ranging from 79.08% to 93%. The enhanced performance can be attributed to the combination of RFECV-based feature selection, eight heterogeneous classifiers, stratified 10-fold cv, in-fold SMOTE, OOF prediction generation, greedy base-learner selection, and XGBoost-based stacking. Rather than directly combining all candidate classifiers, the proposed method first identifies the learners that provide complementary predictive information and then uses their OOF predictions for stacking. This alleviates unnecessary model redundancy and preserves useful diversity of base learners.

Table 4: Comparison Results with Existing Literature

Reference	Model	Methodology	Accuracy%	Dataset
Proposed	Level 0: SVM, RF, DT, XGBOOST, NB, Extra Trees, KNN, LR, Level 1: Stacking (XGBoost)	Cross Validated Greedy Forward Selection	98.50%	Type 2 Diabetes Prediction Bangladesh
[52]	LR, RF, SVM, and KNN with grid search	Parameter optimization using grid search	83%	PIDD
[53]	Level 0 (RF, DT, gradient boost, Gaussian NB, SVM KNN), Level 1 (LR)	Stacking ensemble	93%	PIDD
[54]	CATBoost, XGBoost, RF, LR, SVM	AIC-based forward selection (CATBoost outperform)	82.10%	NHANES 1999–2020
[55]	LR, SVM, k-NN, gradient boost, NB, and RF with majority voting	Voting ensemble	82%	PIDD
[56]	SMOTE + KNN, RF, CatBoost, ExtraTrees + stacking	Stacking ensemble	85.94%	MIMIC-IV

In addition to its predictive power, the proposed framework makes a methodological contribution. It provides a unified workflow that includes RFECV-based feature optimisation, heterogeneous classifier evaluation, in-fold SMOTE, OOF prediction analysis, greedy learner selection, XGBoost-based stacking, and SHAP-LIME explainability. In addition, the performance indicators of the model were computed using stratified 10-fold CV and the mean and standard deviation were obtained over the folds. Mean values are shown in the main tables for a concise and consistent comparison of predictive performance, whilst standard deviations are provided to quantify variability and stability of models over folds. This joint reporting strategy provides better evaluation of both prediction efficacy and cross-validation stability.

4.7 Discussion

The proposed CV-GSE demonstrated good performance for T2DM prediction with 98.50% accuracy, 98.25% precision, 99.70% recall, 98.62% F1-score, and 96.85% ROC-AUC. These results show the benefit of using the predictions of a set of different classifiers instead of only one classifier.

The framework combines RFECV feature selection, stratified 10-fold CV, in-fold SMOTE, OOF-based greedy learner selection and XGBoost stacking. The selection strategy allows to keep the complementary learners, avoiding unnecessary redundancy of the models, resulting in a compact and efficient ensemble. The explainability analysis showed that the most important feature was Glucose, followed by BMI and systolic blood pressure. Moreover, LIME showed that the feature contribution was unique to each case, thus confirming the patient-specific prediction of the model. These results show the complementary value of global SHAP and local LIME explanations. All performance measures were computed using 10-fold CV and

the mean and standard deviation were computed across the folds. Mean values are reported in the main tables for a concise comparison of model performance. Standard deviations were calculated to assess variability and stability across folds. The experimental results indicate that the proposed framework is an effective and interpretable approach for T2DM prediction. However, independent dataset evaluation, nested CV, calibration and prospective clinical data are needed before clinical deployment can be considered.

5. Conclusion

The proposed study was an explainable CV-GSE for predicting T2DM, which combined RFECV-based feature selection, heterogeneous ML classifiers, stratified 10-fold CV, in-fold SMOTE, OOF-based greedy learner selection, and XGBoost stacking. Six informative predictors Age, BMI, SBP, Diabetes Pedigree Function, Insulin, and Glucose were identified by RFECV with the highest cross-validated ROC-AUC of 0.9558. The ensemble model outperformed individual baseline models with the accuracy of 98.50%, precision of 98.25%, recall of 99.70%, F1 score of 98.62% and ROC-AUC of 96.85%. The OOF-based greedy selection strategy helped the framework to find complementary predictive information without unnecessary combinations of models. In addition, SHAP and LIME gave both global and local explanations, with Glucose being the most influential predictor, followed by BMI and systolic blood pressure. The results overall indicate that the proposed feature selection, cross-validated ensemble learning, OOF-based learner selection, and XAI is a promising approach for the prediction of T2DM that is interpretable. However, the results are based on one set of data only and cross-validation experiments and should not be interpreted as clinical readiness. In the future, a multimodal hybrid deep learning framework is developed to further enhance the prediction performance. Feature selection and hyperparameter tuning will be discussed with the use of optimization techniques. Furthermore, clinically approved heterogeneous datasets from multiple hospitals will be gathered to assess the robustness, generalizability, and applicability of the framework in real-world settings.

DECLARATIONS

Statement of Approval. This study has been approved by Department of Computer Application, Integral University, Lucknow, India.

SDG Alignment: This study is relevant to SDG 3, 4, 9 and 10.

Conflict of Interest There is no conflict of interest among the authors as well as consent has been provided for the participation in this study.

Funding There is no funding applicable to this study.

MCN No: IU/ R&D / 2026-MCN0004998

Data Availability: Type-2 Diabetes Dataset Bangladesh (Original data) (Mendeley Data).

References

[1] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review. Diabetology & Metabolic Syndrome, 14, 196. https://doi.org/10.1186/s13098-022-00969-9.

[2] Dinanthi, D. A., Ramadanti, E., Aditya, C. S. K., & Chandranegara, D. R. (2024). Diabetes detection using Extreme Gradient Boosting (XGBoost) with hyperparameter tuning. Indonesian Journal of Electronics Electromedical Engineering and Medical Informatics, 6(2). https://doi.org/10.35882/qr3hw926.

[3] American Diabetes Association. (2024). Standards of Medical Care in Diabetes—2024. Diabetes Care, 47(Suppl. 1), S1–S350. https://doi.org/10.2337/dc24-S001.

[4] Yang, W., Wu, Y., Chen, Y., et al. (2024). Different levels of physical activity and risk of developing type 2 diabetes among adults with prediabetes: A population-based cohort study. Nutrition Journal, 23. https://doi.org/10.1186/s12937-024-01013-4.

[5] A. Fagg, T. J. Olds, M. J. H. et al., "How do we identify people at high risk of Type 2 diabetes and help prevent the condition from developing?" Diabetic Medicine, vol. 36, no. 3, pp. 316–325, 2019, doi: 10.1111/dme.13867.

[6] Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis, 113–125. Springer. https://doi.org/10.1007/978-981-13-8798-2_12.

[7] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based

on data mining. Informatics in Medicine Unlocked, 10, 100–107. https://doi.org/10.1016/j.imu.2017.12.006.

[8] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004.

[9] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using machine learning classification methods. Procedia Computer Science, 167, 706–716. https://doi.org/10.1016/j.procs.2020.03.336.

[10] Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2021). A framework for feature selection through boosting. Expert Systems with Applications, 187, 115895. https://doi.org/10.1016/j.eswa.2021.115895.

[11] Alkhalefah, S., AlTuraiki, I., & Altwaijry, N. (2025). Advancing diabetic foot ulcer care: AI and generative AI approaches for classification, prediction, segmentation, and detection. Healthcare, 13(6), 648. https://doi.org/10.3390/healthcare13060648.

[12] Abu-Jbara, A., Anwar, A., & Mohammed, H. (2022). Diabetes prediction model using data mining techniques. Data Science and Engineering, 7(1), 123–131. https://doi.org/10.1016/j.dse.2022.03.005

[13] Rastogi, R., & Bansal, M. (2022). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605. https://doi.org/10.1016/j.measen.2022.100605.

[14] A. Aich, M. Murshed, S. Hewage, and A. Mayeaux, "A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction," Scientific Reports, vol. 16, Art. no. 12132, 2026.

[15] Awad, M., & Fraihat, S. (2023). Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature selection method for machine learning-based intrusion detection systems. Journal of Sensor and Actuator Networks, 12(5), 67.

[16] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389–422. https://doi.org/10.1023/A:1012487302797.

[17] Iftikhar, K., Javaid, N., Ahmed, I., & Alrajeh, N. (2025). A novel explainable deep learning framework for accurate diabetes mellitus prediction. Applied Sciences, 15(16), 9162. https://doi.org/10.3390/app15169162.

[18] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[19] Pima Indians Diabetes Database. (2016). Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[20] Early Stage Diabetes Risk Prediction. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5VG8H.

[21] Abdelhafez, H. A., & Amer, A. A. (2024). Machine learning techniques for diabetes prediction: A comparative analysis. Journal of Applied Data Sciences, 5(2), 792–807. https://doi.org/10.47738/jads.v5i2.219.

[22] Yadu, S., Chandra, R., & Sinha, V. K. (2024). Comparing different machine learning techniques in predicting diabetes on early stage. Engineering Proceedings, 62, 20. https://doi.org/10.3390/engproc2024062020.

[23] Akhtar, R., & Ahamad, M. K. (2026). A comparative evaluation of various machine learning techniques for prediction of Type 2 diabetes mellitus. ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), 12(60), 1215–1231. https://doi.org/10.5935/jetia.v12i60.4052.

[24] Ganie, S. M., Pramanik, P. K. D., Bashir Malik, M., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics, 14, 1252159. https://doi.org/10.3389/fgene.2023.1252159.

[25] Jena, S. K., Behera, D. K., Jena, A. K., & Rout, J. K. (2025). A novel ensemble machine learning framework for improved diabetes prediction and complication prevention. Procedia Computer Science, 258, 4008–4017. https://doi.org/10.1016/j.procs.2025.04.652.

[26] Alsadi, B., Musleh, S., Al-Absi, H. R. H., et al. (2024). An ensemble-based machine learning model for predicting type 2 diabetes and its effect on bone health. BMC Medical Informatics and Decision Making, 24, 144. https://doi.org/10.1186/s12911-024-02540-0.

[27] Jamal, M. K., & Faisal, M. (2024). Machine learning-driven implementation of workflow optimization in cloud computing for IoT applications. Internet Technology Letters, 8(3). https://doi.org/10.1002/itl2.571.

[28] Daza, A., Sa´nchez, C. F. P., Apaza-Perez, G., Pinto, J., & Ramos, K. Z. (2023). Stacking ensemble approach to diagnosing the disease of diabetes. Informatics in Medicine Unlocked, 44, 101427.

[29] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[31] Al-Dabbas, L., & Abu-Shareha, A. A. (2024). Early detection of female type-2 diabetes using machine learning and oversampling techniques. Journal of Applied Data Sciences, 5(3), 1237–1245. https://doi.org/10.47738/jads.v5i3.298.
[32] Khan, M. A., Ali, M., & Raza, S. (2020). Effective prediction of Type II Diabetes Mellitus using data mining classifiers and SMOTE. In Advances in Data Mining, 189–201. Springer. https://doi.org/10.1007/978-981-15-0222-4_17.
[33] Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing k-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00973-y.
[34] Demirciog˘lu, A. (2021). Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights into Imaging, 12, 172. https://doi.org/10.1186/s13244-021-01115-1.
[35] A. K. M. et al., "Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles," Applied Soft Computing, vol. 111, Art. no. 107689, 2021, doi: 10.1016/j.asoc.2021.107689.
[36] Ansari, Z.A. et al. (2026). Empowering Breast Cancer Diagnostics: SHAP-Enhanced Explainable AI. In: Nayak, S.C., Dehuri, S., Cho, SB., Favorskaya, M. (eds) Green Artificial Intelligence and Industrial Applications (G-AIIA). GreenAI 2025. Artificial Intelligence-Enhanced Software and Systems Engineering, vol 8. Springer, Cham. https://doi.org/10.1007/978-3-031-99882-9_22.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[38] J. Kaliappan et al., "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," Frontiers in Artificial Intelligence, vol. 7, 2024, Art. no. 1421751, doi: 10.3389/frai.2024.1421751.
[39] Agliata, A., Giordano, D., Bardozzo, F., et al. (2023). Machine learning as a support for the diagnosis of Type 2 diabetes. International Journal of Molecular Sciences, 24(7), 6775. https://doi.org/10.3390/ijms24076775.
[40] Roobini, M. S., Lakshmi, M., Rajalakshmi, R., Sujihelen, L., & Babu, K. (2023). Type 2 diabetes mellitus classification using predictive supervised learning model. Soft Computing. https://doi.org/10.1007/s00500-023-08726-4.
[41] Li, X., Ding, F., Zhang, L., et al. (2025). Interpretable machine learning method to predict the risk of pre-diabetes using national-wide cross-sectional data: Evidence from CHNS. BMC Public Health, 25. https://doi.org/10.1186/s12889-025-22419-7.
[42] J. Kaliappan, I. J. Saravana Kumar, S. Sundaravelan, T. Anesh, R. Rithik, Y. Singh, D. V. Vera-Garcia, Y. Himeur, W. Mansoor, S. Atalla, and K. Srinivasan, "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," Frontiers in Artificial Intelligence, vol. 7, Art. no. 1421751, 2024, doi: 10.3389/frai.2024.1421751.
[43] Z. Zhang, Y. Lu, M. Ye, W. Huang, L. Jin, G. Zhang, Y. Ge, A. Baghban, Q. Zhang, H. Wang, and W. Zhu et al., "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis," Journal of King Saud University—Computer and Information Sciences, vol. 36, no. 1, Art. no. 101873, 2024, doi: 10.1016/j.jksuci.2023.101873.
[44] Mohsen, F., et al. (2025). ECG features improve multimodal deep learning prediction of diabetes. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-12633-z.
[45] P. K. Mohanty, S. A. John Francis, R. K. Barik, D. S. Roy, and M. J. Saikia, "Leveraging Shapley Additive Explanations for Feature Selection in Ensemble Models for Diabetes Prediction," Bioengineering, vol. 11, no. 12, Art. no. 1215, 2024, doi: 10.3390/bioengineering11121215.
[46] Lee, H., et al. (2025). AI machine learning–based diabetes prediction in older adults. JMIR Formative Research, 9, e57874. https://doi.org/10.2196/57874.
[47] Abnoosian, K., et al. (2023). Prediction of diabetes disease using an ensemble machine learning approach. Journal of Biomedical Informatics, 130, 103984. https://doi.org/10.1016/j.jbi.2022.103984.
[48] K. O. et al., "A stacked ensemble machine learning approach for the prediction of diabetes," Journal of Diabetes & Metabolic Disorders, vol. 23, no. 1, pp. 603–617, 2024, doi: 10.1007/s40200-023-01321-2.
[49] V. Q. Tran, Y. Choi, and H. Byeon, "Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction," Journal of Men's Health, 2024, doi: 10.22514/jomh.2024.184.

[50] Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artificial Intelligence in Medicine, 164, 103132. https://doi.org/10.1016/j.artmed.2025.103132.
[51] I. A. Elgendy, M. Hosny, M. A. Albashrawi, and S. Alsenan, "Dual-stage explainable ensemble learning model for diabetes diagnosis," Expert Systems with Applications, vol. 274, Art. no. 126899, 2025, doi: 10.1016/j.eswa.2025.126899.
[52] R. Krishnamoorthi et al., "A novel diabetes healthcare disease prediction framework using machine learning techniques," Journal of Healthcare Engineering, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.
[53] S. R., S. M., M. K. Hasan, R. A. Saeed, S. A. Alsuhibany, and S. Abdel-Khalek, "An empirical model to predict the diabetic positive using stacked ensemble approach," Frontiers in Public Health, vol. 9, Jan. 2022, doi: 10.3389/fpubh.2021.792124.
[54] Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., & Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. International Journal of Environmental Research and Public Health, 19(22), 15027. https://doi.org/10.3390/ijerph192215027
[55] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques," Mobile Information Systems, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/6521532.
[56] I. A. Elgendy, M. Hosny, M. A. Albashrawi, and S. Alsenan, "Stacked Ensemble Learning for Type 2 Diabetes Prediction Using the MIMIC-IV Dataset," 2025 International Conference on Decision Aid Sciences and Applications (DASA), 2025, doi: 10.1109/DASA68193.2025.11498899.