Biomedical & Pharmacology Journal, March 2026. Vol. **19**(1), p. 246-263

Early-Stage Diabetes Prediction Using a Stacked Ensemble Model Enhanced with SHAP Explainability

Shahnawaz Ahmad[1], Shahadat Hussain[2], Mohd. Arif[3]*and Mohd. Aquib Ansari[3]

[1]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Gautam Buddh Nagar, Uttar Pradesh, India.
[2]Science Technology and Technical Education Department, Government of Bihar, Patna, Bihar, India.
[3]School of Computer Science and Engineering, Galgotias University, Greater Noida, Gautam Buddh Nagar, Uttar Pradesh, India.
*Corresponding Author Emails:md.arif@galgotiasuniversity.edu.in

https://dx.doi.org/10.13005/bpj/3350

(Received: 02 September 2025; accepted: 08 January 2026)

Diabetes is one of the most prevalent diseases of our time, and, untreated, it can lead to other health issues. The objective of this research paper is to develop an explainable stacked ensemble model for the early diagnosis of diabetes. The Early-Stage Diabetes Risk Prediction dataset was preprocessed using mean imputation, SMOTE-based class balancing, and mean normalization. A stratified train–test split was applied, and a stacked ensemble model was developed, utilising SHAP and LIME to ensure explainable and interpretable predictions. The proposed model achieved higher performance regarding the Early Stage Diabetes Risk Prediction dataset than did typical models, including Naive Bayes (NB), k-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Decision Tree (DT), with an accuracy of 98.4%. The innovative application of ensemble learning enhances the model's reliability and effectiveness, offering valuable insights for identifying potential diabetic patients. The high accuracy underscores the model's potential as a valuable tool for early detection and treatment, ultimately improving patient outcomes in diabetes management. A critical aspect of our methodology is the integration of SHAP (SHapley Additive exPlanations) and Local Interpretable Model-Agnostic Explanations (LIME), which enhances explainability by revealing the factors driving the model's predictions and highlighting feature importance.

Keywords: Diabetes disease; Ensemble learning; Machine learning; Prediction, Stacking.

Diabetes is the most common disease nowadays. It may also lead to other disorders. Identification at an early stage can significantly reduce its negative effects. A lot of studies support the use of ML techniques for diabetes prediction. However, only a few studies have explored ensemble Learning techniques for diabetes prediction, and there are scarce studies available on applying a stacked ensemble approach, especially on datasets that mainly include features of early-stage diabetic parameters. Typically, the effectiveness of a single categorization model is restricted. In addition, it has some flaws, such as a limited capacity for generalization and a low fault tolerance, which render its illness prediction and diagnostic capabilities subpar. Ensemble learning techniques have evolved to address this issue.[1] The idea of mixing weak classifiers to obtain

strong ones was the foundation for the model developed by ensemble learning. The stacking ensemble approach by David H. Wolpert reduces generalization error.[2] Unlike Schapire et al.'s boosting and Leo Breiman's bagging methods, stacking trains the main learner using the original training dataset. The meta-learner oversees improving a single model by analyzing the output of the main learner.[3] Stacking is a popular technique used in many different fields.[43] For example, Ali and his colleagues successfully used stacking to predict amino acid sequences associated with breast cancer.[14] Their stacking model outperformed both basic ML algorithms and other ensemble methods.

The evaluation of the model was based on a detailed review of the Early-Stage Diabetes Risk Prediction dataset. Notably, the stacked ensemble model proposed in the study achieved an excellent performance, with an accuracy of 98.4%. This stacked ensemble model produced another result that bested other basic ML algorithms, such as Naive Bayes, k-Nearest Neighbor, Support Vector Machine, and Decision Tree, for diabetes. It achieved a staggering 98.4% success rate, surpassing prior versions. The model is primarily used to track and diagnose potential diabetic patients, enabling them to receive treatment before the condition worsens. Early diagnosis of diabetes is critically important because it helps to avoid serious complications if the disease is treated on time. The high level of accuracy achieved by the stacked ensemble model offers promise for its further application as a valuable tool in diabetes predictive models and proactive healthcare.

This study offers several contributions to the field of diabetes prediction:

• The study also performs an evaluation of the various classification models for determining the most appropriate one for predicting diabetes. different classification models to determine the most effective one for predicting diabetes. It discusses NB, KNN, SVM, DT, and a novel model, Stacked Ensemble.

• The stacked ensemble model was very accurate in delivering a correct classification rate of the cases at 98.4% study compares different classification models to find the best one for predicting diabetes. It examines NB, KNN, SVM, DT, and a new model called the stacked ensemble.

• The stacked ensemble model was very accurate, correctly identifying 98.4% of cases. The former demonstrates that the model is applicable in realizing people with early-stage diabetes and could be applied in healthcare.

• This research is intended to discover individuals who have a predisposition to diabetes even before the first symptoms can be noticed. different classification models to determine the most effective one for predicting diabetes.

In contrast to other works that use ensemble learning and post-hoc explainability to traditional diabetes data, this paper is the first to use a stacked ensemble design, with both global and local explainability (SHAP and LIME) to predict diabetic symptoms at an early stage, which can be clinically interpreted at the case level, and has the highest predictive accuracy.

Quite simply, this study provides valuable knowledge when selecting the most optimal classification models for diabetes prediction, presents a more efficient stacked ensemble model and emphasizes the importance of early detection regarding diabetic diseases. All these contributions, especially in creating quality positive healthcare results and increasing the prevention of diabetes, could be of great benefit. Regarding the novelty of this work, this research presents a novel approach to early diabetes prediction with an unusually high accuracy rate, implemented using a stacked ensemble model. Furthermore, integrating SHAP (SHapley Additive exPlanations) improves model explainability to uncover how the model is making its prediction. That is, this work's innovation in the diabetes prediction domain is characterized by the indicated twin that has attained predictively high accuracy while being analytically explained using SHAP.

The following are the research questions (RQs) that have been addressed in the paper:

• RQ1: Which classification model is most effective for diagnosing diabetes?

• RQ2: How do the evaluated models compare in overall predictive performance across standard metrics?

• RQ3: How effectively does the proposed stacked ensemble detect early-stage diabetic cases?

This research aims to create an innovative ensemble ML method for early diabetes prediction, using a stacked classifier that combines multiple base classifiers. This approach captures complex

data patterns, thereby enhancing the accuracy of early-stage diabetes detection. This research is valuable because it could improve early diagnosis and treatment of diabetes, a common and long-term health problem that affects many people and healthcare systems. The new stacked ensemble model in this study is highly accurate, correctly predicting diabetes in 98.4% of cases. This is better than other models used for this purpose. The innovative methodology and superior performance of our model have the potential to revolutionize the field of diabetes risk assessment and contribute to more effective preventive healthcare strategies.

The remainder of this paper is structured as follows: Section 2 examines the perspectives of others on predicting diabetes early. Section 3 explains how we made a new stacking classifier. Section 4 shows the results of our tests. In section 5, a brief discussion is presented. Section 6 summarizes what we learned about early diabetes prediction.

Literature Review

Machine learning has the potential to significantly improve disease diagnosis, particularly for serious conditions such as breast cancer, [26] heart disease, [27, 28] and diabetes. [24, 30] Many studies have shown that ML algorithms can effectively analyze patient data to identify patterns, detect early signs of diseases, and make accurate predictions. This section will discuss some of the most promising research in using ML for diabetes prediction, focusing on both basic algorithms and more advanced ensemble methods. These studies aim to improve the accuracy and efficiency of diabetes diagnosis.

The detection and diagnosis of diabetes are among the most significant issues facing ML researchers. Fatima et al. have successfully examined numerous machine-learning algorithms to identify various ailments.[5] To identify Type 2 diabetes (T2D) and cardiovascular illness, Ali et al. utilized artificial neural networks and Bayesian networks to detect and categorize diabetes.[6] Kumar et al. used Naïve Bayes, support vector machine, and decision tree models to predict heart disease in individuals with diabetes.[7] Their forecast accuracy depends on their prediction accuracy. Saru et al. predicted diabetes using medical bioinformatics.[8] In a study, researchers developed a Diabetes Classification and Prediction Model (DCPM) that addresses class imbalance, outliers, and missing values in the dataset.[24] The proposed pre-processing technique, applied to the Pima Indian Diabetes (PID) dataset, effectively removes outliers, fills missing values, standardizes data, and selects relevant features. The optimized K-NN classifier achieves a statistically significant classification accuracy, recall, precision, and F1-score outperforming existing approaches. Another research explores the use of fifteen classifiers, including Artificial Neural Networks (ANN), SVM, KNN, Naive Bayes, and Ensemble, to develop an intelligent framework for diagnosing diseases.[25] MATLAB and WEKA 3.6.13 are employed as tools for analysis and prediction. The ensemble method, which combines the predictive abilities of individual classifiers, demonstrates superior performance with an accuracy of 98.60%, outperforming other individual algorithms such as ANN, Naïve Bayes, SVM, and K-NN. Researchers also explored heterogeneous Ensemble Classifiers[10], soft voting[13], hard voting[19], [20], and Stacking-based ensemble based on multi-objective evolution[14] approaches for diabetes prediction.

After this literature review, we found the importance of the ensemble learning approach in diabetes prediction. Next, we conduct an in-depth analysis of the performance of various machine learning methods used for diabetes prediction. Our suggested stacked ensemble model distinguishes itself by achieving an exceptional accuracy of 98.4%, surpassing the performance of currently available models. The goal of this research is to identify individuals who are at risk of developing diabetes sooner rather than later, so that they can receive timely care and improve their health.

Table 1 summarizes studies on the use of ML techniques for detecting diabetes using retinal images. The studies found that various ensemble methods, such as random forest and voting ensembles, outperformed traditional methods. Feature selection techniques were also shown to improve accuracy. However, limitations such as imbalanced datasets and small datasets were identified in some studies. In general, ML appears promising for large-scale retinal screening for diabetes, but more research is required to overcome its current shortcomings.

Although existing ensemble-based diabetes prediction studies describe promising

accuracy, many use small or non-balanced datasets and have not been validated on early-stage and symptom-based data. Additionally, most of the research lacks significant clinical interpretation and does not offer the possibility of explaining cases at the individual level, which reduces its potential usage in clinical practice. This is the reason to fill these gaps by using an explainable stacked ensemble model specifically designed for diabetes prediction at early stages.

Table 2 provides a comprehensive overview of various ML approaches applied to diabetes prediction. Researchers have employed a range of algorithms, including Decision Trees, SVM, Naive Bayes, Random Forest, Logistic Regression, Gradient Boosting, and Artificial Neural Networks, to predict diabetes. Datasets used for training and evaluation include the PIMA Diabetes Dataset and custom clinical datasets. The accuracy achieved by these models varies, with some reaching as high as 99.34%. Techniques like feature selection, data balancing, and ensemble methods have been explored to enhance predictive performance. Overall, ML has shown promise in accurately predicting diabetes, but further research is needed to address challenges such as data quality and model interpretability.

MATERIALS AND METHODS

This research uses a dataset called "Early-Stage Diabetes Risk Prediction" that was downloaded from a website called "UC Irvine ML Repository.[21] The dataset consists of 520 instances, with each instance having 17 attributes. Out of these instances, 400 are positive samples indicating the presence of diabetes, while 120 are negative samples representing the absence of diabetes. Table 2 summarizes the features of the dataset. To facilitate the study, values for Features 2 through 16 are coded as 1 for "Yes" and 0 for "No". In relation to feature No. 17, occurrences without diabetes were classified as the negative class (0), while those with the condition were classified as the positive class (1). Since the dataset primarily consists of categorical attributes, with 'Age' being the only exception, and the target feature is also categorical, the Chi-square test was chosen as the appropriate method for feature selection. Through this study, it was found that among all the 17 attributes, Polyuria and Polydipsia exhibited the highest Chi-squared value, indicating their significant association with the target feature. The dataset provides a list of attributes used to predict the presence of diabetes. These attributes include demographic information such as age and sex, symptoms associated with diabetes, including weakness and polyuria, as well as other health factors, like muscle stiffness and obesity, along with additional observations, including visual blurring and sudden weight loss. The final attribute, "Class," indicates whether the individual has diabetes or not.

Data Preprocessing

The Early-Stage Diabetes Risk Prediction dataset consists mostly of categorical features, with a few missing values. The missed cases were managed using mean imputation, which is appropriate for maintaining the completeness of data sets without causing large bias. To address the issue of class imbalance, where there are more instances of diabetics compared to non-diabetics, the Synthetic Minority Oversampling Technique (SMOTE) was employed to generate artificial samples of the minority group, thereby enhancing the model's ability to learn balanced decision boundaries. The stratified train-test split was used to maintain the initial distribution of classes when testing the model.

The Chi-square test has been applied to select features because most of the input features and the target variable are categorical. It is a statistical dependence appraisal of each feature and the class label, which is used to determine the most important symptom-based attributes. Nonlinear associations and complex relationships, however, cannot be captured by Chi-square feature selection, which can overlook these intricate relationships. Although it has this shortcoming, it was chosen because it is simple, interpretable, and applicable to categorical healthcare data.

Classifier Selection and Hyperparameter Configuration

The suggested stacked ensemble model utilises NB, KNN, DT, and SVM as base learners because their learning nature is complementary, and they have demonstrated their capabilities in medical diagnosis tasks. The meta-learner chosen is LR due to its strength, probabilistic nature, and binary classification capabilities. The choice of hyperparameters was made according to the values

commonly used in the literature to ensure a fair comparison and guarantee reproducibility.

Proposed Model

There is a typical way of solving problems when using a simple machine algorithm. However, this method can sometimes be ineffective for complex tasks because it relies significantly on parameters such as the type of input data. To address these issues, what is known as the 'ensemble models' is used, where several machine learning algorithms are used. Latent class models are used to address the limitations of one model and capitalize on the strengths of the other. The system for predicting diabetes proposed in this paper works on the principle of a stacked ensemble. This technique is expected to provide synergy of improvement on various sub-models and hence perform better than individual techniques.

The Ensemble method comprises three main categories: Bagging, Boosting, and Stacking. All these models can be likened to having their strengths and weaknesses. Along the lines of the proposed classification model in predicting diabetes patients, the stacked ensemble modeling technique was used. Stacking is a method of classification that involves two levels: level 0 and level 1. These two levels are collectively known as the Meta classifier. In contrast to more traditional methods, such as bagging and boosting, stacking creates an entirely new dataset from which to draw its predictions. Unlike other multi-classifier algorithms, which rely on vote averaging or consensus, this method uses each classifier's output individually to arrive at a final prediction. The effectiveness of stacking relies on the compilation of projected probabilities generated by individual classifiers.

The general structure of the proposed stacked ensemble model is depicted in Figure 1. The level-0 classifiers, sometimes referred to as weak learners, are represented by the basic learner's classifier 1, classifier 2,..., classifier N. The new training set is created by training the dataset using the base learners. The newly created set will be used to train a level-1 classifier, which serves as the meta-learner. The test set can be predicted by the level-1 classifier after it has been trained. At level 0 of the stacked ensemble model for diabetes prediction, a combination of diverse algorithms can be utilized. These algorithms can either be of similar types (homogeneous) or different types (heterogeneous). In the case of a heterogeneous setup, several distinct algorithms are incorporated, each serving a specific purpose. On the other hand, in a homogeneous configuration, the same algorithm is applied but with various parameter values to enhance its performance. Algorithm 1 outlines the structure of the suggested stacked ensemble model. At level 0, classification algorithms like NB, DT, SVM, and KNN are utilized as classifiers.

Our study uses LR as the primary classifier (M) in Algorithm 1. LR is an ML technique that predicts categories by examining data. Unlike giving a direct answer, LR calculates the chances of different outcomes happening. This approach is well-suited for binary classification tasks. The model evaluates the probability of each event occurring through a linear transformation of input features. In our study, we used logistic binomial regression, which is appropriate for cases when the dependent variable has two possible outcomes, such as diabetic (1) or non-diabetic (0).

Algorithm 1: Stacked ensemble learning algorithm using 5-fold cross-validated out-of-fold predictions for meta-feature generation

Inputs:

• Dataset D (the main dataset)
• Level-0 classifiers C1, ..., CT
• Level-1 classifier M

Outputs:

• Ensemble classifier H(x) that combines the outputs of the level-0 and level-1 classifiers.

Steps:

1. For each i from 1 to T:
a. Compute hi = Ci(D) // Apply the i^{th} level-0 classifier to the original dataset
2. Set D' = 0 // Initialize a new dataset
3. For each i from 1 to m:
a. For each t from 1 to T:
Compute Zit = ht(xi) // Apply the t^{th} level-0 classifier to the i^{th} instance xi in D
b. Add the tuple (Zi1, ..., ZiT, Yi) to D' // Append the feature vector Zit and the class label Yi to D'
4. Compute h' = M(D') // Apply the level-1 classifier to the new dataset D'
5. Compute H(x) = h'(h1(x), ..., hT(x)) // Combine the outputs of the level-0 classifiers using the level-1 classifier
6. Return H(x)

Algorithm 1 implements stacking with out-of-fold predictions rather than simple blending, ensuring that the meta-learner is trained only on predictions from unseen samples. The algorithm used an ensemble learning approach that combines the outputs of multiple level-0 classifiers using a level-1 classifier to enhance the accuracy of predictions. The input to the algorithm is the initial data set D, a set of T level-0 classifiers C1 to CT and a level-1 classifier M. The level-0 classifiers are first used to predict intermediate predictions hi of the instances in the original dataset. They are combined with the original data to form a new dataset, which is then used to derive the intermediate predictions. This dataset is used by one of the classifiers at level 1 to make a final decision on the instance. The average of the final predictions is then calculated, along with the intermediate predictions, in a weighted form to yield the final ensemble prediction. This ensemble classifier is an algorithm output.

To avoid overfitting and information leakage in the stacking framework, we employed a 5-fold stratified cross-validation strategy to generate out-of-fold (OOF) predictions. During training, the dataset was divided into five folds. For each fold, the level-0 base learners (Naïve Bayes, KNN, Decision Tree, and SVM) were trained on four folds and used to generate predictions on the held-out fold. These OOF predictions were concatenated to form the meta-feature matrix, which was then used to train the level-1 meta-learner (Logistic Regression). During inference, all base learners were retrained on the full training set, and their outputs were passed to the trained meta-learner to obtain the final prediction.

Table 3 provides a description of the stacking ensemble architecture, including the specified hyperparameters and the reasoning behind the base learners and meta-learner. Naive Bayes is added to the base model layer to promote probabilistic efficiency. KNN (with $k = 5$) is added to strike a balance between the bias-variance trade-off. DTs are added to ensure that the model has clinical interpretability. To deal with non-linear and complex patterns, an SVM with an RBF kernel is employed. Lastly, the Logistic Regression is the meta-learner that uses the LBFGS solver and a larger iteration limit to obtain stable convergence and give sound probability estimates to aid clinical decision support.

To evaluate the models, various performance criteria, including accuracy, precision, recall, F1 score, and AUC, were utilized. To calculate these measures, the first step was to gather data on true positives, true negatives, false positives, and false negatives by constructing a confusion matrix.[22] We employed accuracy, precision, recall, the F1 measure, and the AUC to measure the efficiency of the classifiers. Accuracy determines the work's total effectiveness and the accuracy of the percentage of positive predictions. Recall examines how well the model can identify all the positive cases that are present. The F1-score unites precision with recall into a single metric.[31] AUC measures how well a model distinguishes between classes by its ability to demonstrate both false positives and true positives. Table 4 lists formulas and explanations for different evaluation measures used in ML, including accuracy, precision, recall, specificity, F1-score, and AUC. These measures assess various aspects of a model's

Table 1. Significant works on ensemble learning for diabetes prediction

Source	Outcome	Limitations
9	A tool for mass retinal screening to detect Diabetes	Imbalanced dataset
11	Random Forest outperformed	Small dataset
12	The voting ensemble outperformed	N/A
15	Random Forest outperformed	Low accuracy
16	Feature selection improved accuracy	No comparison of feature selection techniques
17	Soft voting ensemble improved the Area Under Curve (AUC)	N/A
18	The ensemble gives good precision	Small dataset

Table 2. Comparison of Machine Learning Approaches for Diabetes Prediction

Source	Methodology	Algorithms/Techniques Used	Dataset	Accuracy/Findings
31	Compared ML techniques	Decision Tree, SVM, Naive Bayes	PIMA	NB showed the highest accuracy (76.30%).
32	Investigated classification techniques	NB, Quadratic Discriminant Analysis, LDA	PIMA	The highest accuracy achieved was 81.97%.
33	Proposed a k-means clustering and logistic regression-based algorithm	K-means, Logistic Regression, PCA	PIMA	Dimension reduction using PCA enhanced performance.
34	Investigated diabetes prediction using two ML models	Logistic Regression, Gradient Boosting	PIMA	Logistic Regression (LR) achieved an accuracy of 84.7%, while GB achieved 88%.
35	Developed an ANN-based algorithm	Artificial Neural Network	Custom	Achieved 87% accuracy.
36	Used logistic regression and k-means for type-2 diabetes prediction	K-means, Logistic Regression	PIMA	Used 10-fold cross-validation.
37	Analyzed clinical data on diabetes using ML	Random Forest, data balancing techniques	Clinical	Random Forest achieved 98% accuracy.
38	Proposed explainable AI-based approach for diabetes prediction	Ensemble classifier, six ML algorithms	PIMA	Achieved 90% accuracy with a weighted ensemble model.
39	Developed a predictive model using the train-test method and correlation feature selection	Random Forest, other ML techniques	PIMA	Random Forest achieved 97.75% accuracy.
40	Proposed stacking-based ensemble with NSGA-II for feature selection	Stacking ensemble, NSGA-II	PIMA	The stacking model achieved an accuracy of 88.18%.
41	Developed a hybrid model for T2DM prediction using three ML models	Logistic Regression, Decision Tree, Random Forest	PIMA	Logistic Regression achieved 99.34% accuracy.
42	Proposed a hybrid model for hypertension and type-2 diabetes prediction	Logistic Regression, Decision Tree, Random Forest	Custom	The model achieved 92.55% accuracy.

performance, including overall accuracy, the ability to identify positive cases, the ability to identify negative cases, and performance at different levels.

The computational complexity of the proposed stacked ensemble model is shown to be roughly the sum of the training complexities of the base learners over the cross-validation folds, followed by an additional training of a lightweight Logistic Regression meta-learner. Since the dataset size was moderate (520 samples and 17 features), the total training time was also low and was computationally possible. The stacked ensemble is more costly to train than individual classifiers due to the cost of cross-validation; however, this cost is minor and is compensated for by the resulting performance improvement.

RESULTS

All the performance measures presented in the text, tables and figures have been recalculated under the same evaluation protocol, and the outcome is perfectly harmonized to prevent any differences. Using Python's Sklearn library and PyCharm, we developed a stacked ensemble model for this study. Users can increase their productivity while programming in Python using PyCharm, an IDE specifically designed for Python. The ensemble model comprised four base learners (NB, KNN, DT, and SVM) and a Meta-learner (Logistic Regression). The model's performance was assessed via ten prediction experiments. Table 3 presents the results of these experiments.

Table 3. Hyperparameter Settings of Base Learners and Meta-Learner

Model	Key Hyperparameters	Justification
Naïve Bayes	Default (Gaussian NB)	Efficient for categorical and probabilistic medical data.
KNN	k=5, distance = Euclidean	Balances bias–variance trade-off; commonly used in healthcare.
Decision Tree	Criterion = Gini, max_depth = None	Allows interpretability and captures feature interactions.
SVM	Kernel = RBF, C=1.0, ã=scale	Effective for non-linear decision boundaries.
Logistic Regression (Meta-learner)	Solver = lbfgs, max_iter = 1000	Stable convergence and reliable probability estimates.

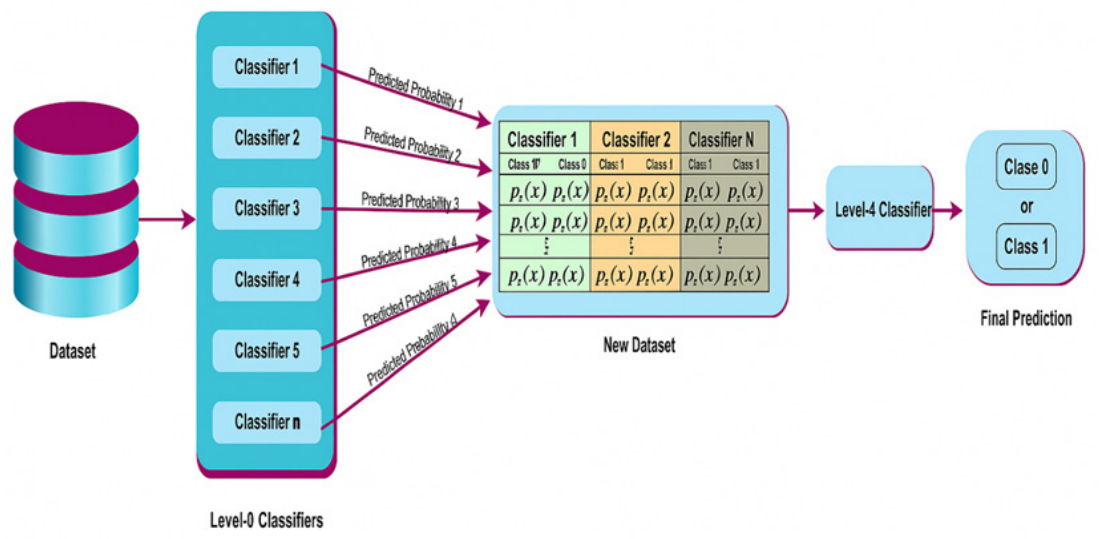

Fig. 1. Proposed Stacked Ensemble model

Various ML techniques were evaluated using different metrics to determine the most effective one. The prediction accuracy of the individual base classifiers and the Stacked ensemble model was analyzed based on the data in Table 5. The findings show that the precision varies between 90.21% and 98.40%, with a maximum of 98.40% precision in the Stacked ensemble (Answer to RQ2). Indeed, upon examining the TNR, marked differences were observed, with classifier efficiency ranging from 92% to 94%. SVM achieved the highest accuracy at 94.21%, whereas the Stacked Ensemble had the second-highest accuracy of 93.7%, and the DT yielded the lowest figure of 90.78% (Answer to RQ3).

The True Positive Rate is relevant because our model aims to accurately classify individuals with diabetes. This, in turn, allows us to take appropriate action and interventional measures to minimize the chances of such cases going unnoticed. The model must perform particularly well in terms of measurement variables with high specificity in recognizing possible cases of diabetes. As concluded in the previous section, the Stacked Ensemble classifier achieves the highest TPR of 89.03%, which, when combined with an

Table 4. Evaluation Metrics

Evaluation Metric	Formulae	Description
Accuracy (Acc)	$\frac{[TP]+[TN]}{[TP]+[TN]+\{FP\}+\{FN\}}$	
Precision (P)	$\frac{[TP]}{[TP]+\{FN\}}$	$[TP] = True\ Positive$ $[TN]$ = True Negative
Recall (Re)	$\frac{[TP]}{[TN]+\{FP\}}$	[FP] = *False Positive* $[FN] = False\ Negative$
Specificity (Spe)	$\frac{[TN]}{[TN]+\{FP\}}$	
F-1 score (F1)	$2X\frac{\text{Recall X Precision}}{\text{Recall+Precision}}$	The F-1 score is a measure of model performance that combines precision and recall into a single metric.
Area under the ROC Curve (AUC)	$AUC = \frac{S_0 - k_0(k_0+1)/2}{k_0 k_1}$	The sum of r sub i, r sub i is the rank of the ith positive example in the rank list. k sub 0 and k sub 1 are the numbers of negative and positive examples, respectively.

Table 5. Evaluation of Machine Learning Models

ML models	TPR	TNR	ACC	F1 Score
NB [10]	85.70	92.09	90.21	82.65
KNN [11]	81.83	92.26	90.65	80.25
DT [12]	83.30	90.78	89.34	80.02
SVM [15]	81.84	94.21	91.30	81.46
Stacked Ensemble	88.56	92.67	98.4	94.10

F-1 score of 93.04% from the previous sections, provides optimal overall performance. Figure 2 illustrates the AUC values of all classifiers used in the study. This detailed analysis enables us to determine the best model for predicting diabetes in our research based on the performance criteria.

Figures 3(a) to 3(e) display the confusion matrices for the execution of all four classifiers and the proposed Stacked Ensemble approach on the dataset. The matrices present the predicted and actual values in tabular form. The confusion matrix serves as the basis for calculating recall, precision, F1, and Accuracy.

The results in Table 6 provide confidence intervals for accuracy, computed using the binomial proportion method, based on a sample size of 520 instances. To determine the statistical significance of the performance differences, we applied the McNemar test to compare the proposed stacked ensemble's performance with that of each baseline classifier, based on paired predictions made on identical test samples. According to the results, the proposed model demonstrates statistically significant performance improvements compared to the individual classifiers at a 95% confidence level ($p < 0.05$).

The matrix clearly shows the Stacked Ensemble has the highest correct predictions, i.e., the highest sum of TP and TN, which is 507. The primary aim of this research is to identify potential patients with diabetes. Based on the values obtained from the confusion matrix, as well as the TPR and F-1 scores presented in Table 4 (Answer to RQ1), the Stacked Ensemble classification algorithm is deemed to be the most suitable and appropriate choice for achieving this objective. The stacking-based ensemble approach achieved the highest accuracy rate among the algorithms used in this study. In terms of accuracy, it exceeded the other study works, as shown in Table 7. In summary, this is because stacking has the most iterations. It can be utilized for early identification of diabetes risk in individuals and is a reliable indication of diabetes risk.

Our results show that the proposed stacked ensemble model has achieved notable success in predicting early-stage diabetes. By integrating multiple ML algorithms, the model attained a remarkable accuracy of 98.4% on the Early-Stage Diabetes Risk Prediction dataset. This performance significantly exceeds that of well-known standalone models, such as NB, KNN, SVM, and DT, highlighting the effectiveness of ensemble learning in improving predictive accuracy.

DISCUSSION

SHAP values are used to illustrate the factors that influence the proposed model's

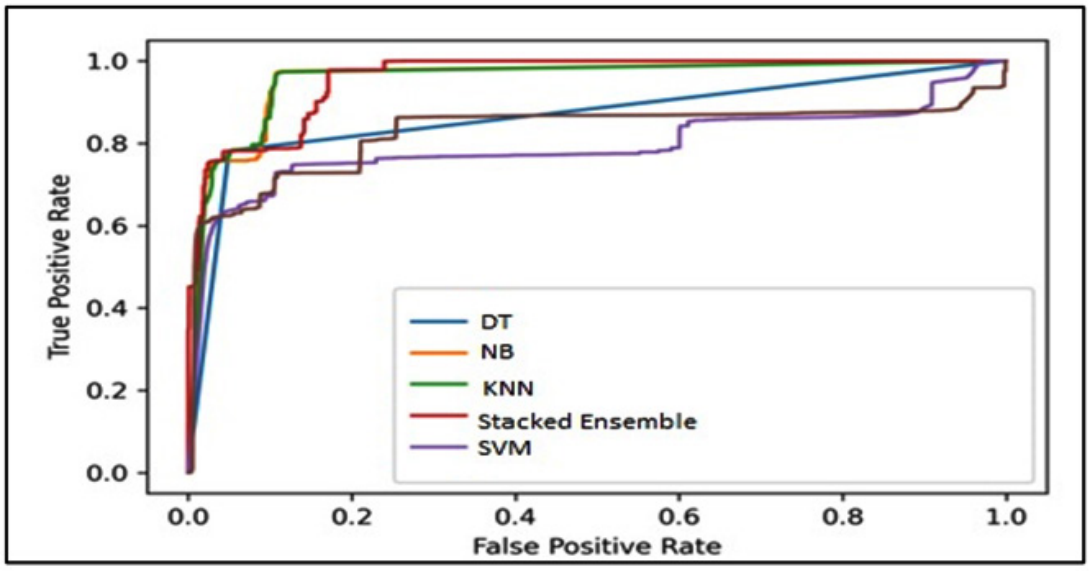

Fig. 2. AUC (ROC Curve) for different classifiers used in this study

output in relation to the base value, representing the model's collective output across the training dataset. Features highlighted in red contribute to higher predictions, while those in blue indicate lower predictions. Figure 4 portrays that the stacked ensemble model considers Polyuria as the foremost contributor to the overall output, closely followed by Polydipsia.

Figure 5 illustrates a force plot that utilizes SHAP values to explain how various factors contribute to the diabetes predictions generated by our model. SHAP values help us understand which features are most important in a specific prediction, and they do so in a way that is both fair and accurate. In this plot, the features are aligned along the x-axis, while the SHAP values (representing impact) are plotted on the y-axis. Positive contributions pushing the prediction towards a positive outcome (e.g., diabetes) are shown in red, while negative contributions pushing the prediction towards a negative outcome (e.g., non-diabetic) are shown in blue.

Biomarker characteristics necessary for prediction are indicated in the plot below. The

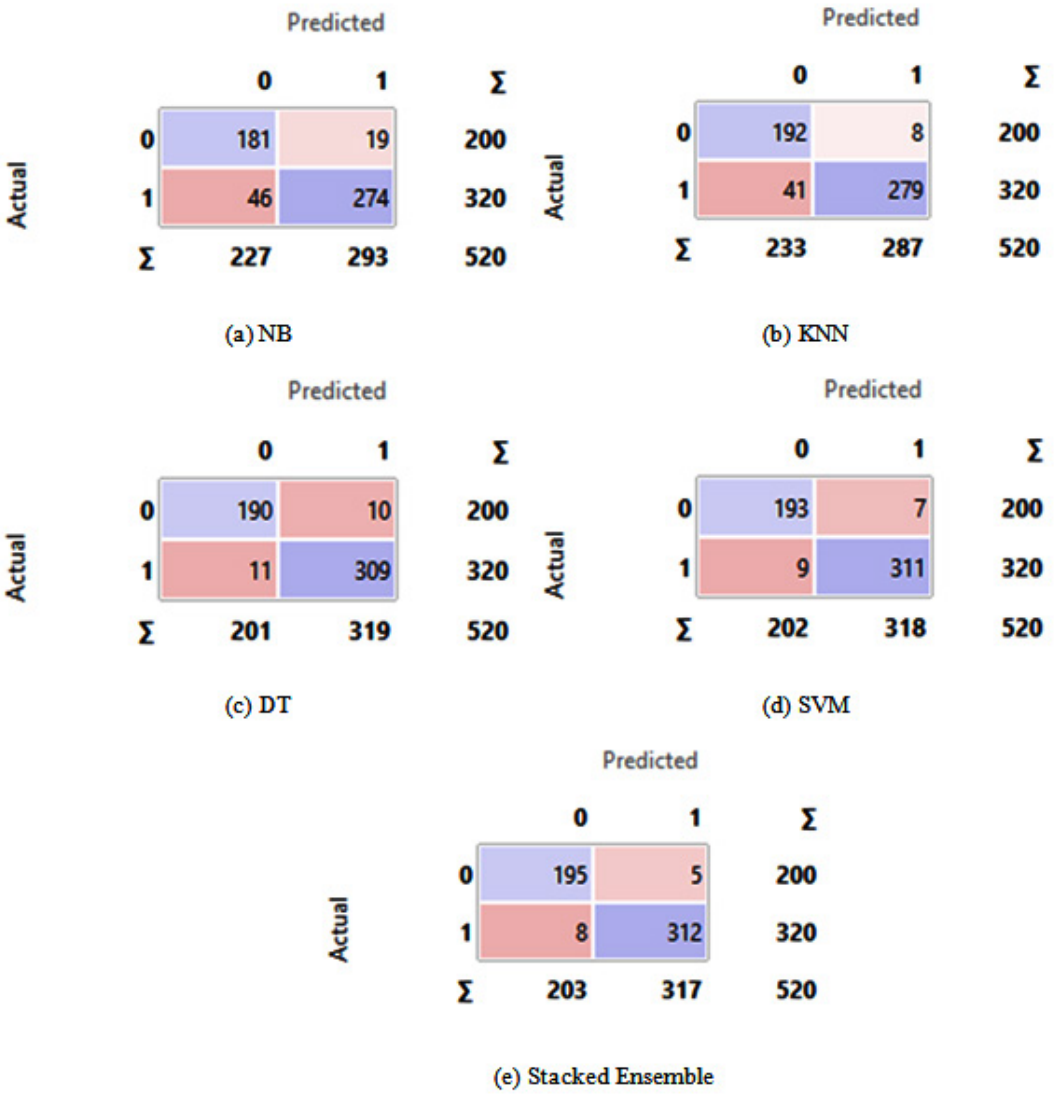

Fig. 3. Confusion matrices of different models and for the stacked ensemble

two most important features are identified as "Polydipsia" and "Polyuria," both of which have quite high SHAP values, indicating that they play a crucial role in the model's decision-making process. Polydipsia is a disease state characterized by excessive thirst, while Polyuria is a subtype of polydipsia associated with excessive urination. Both constitute unmistakable signs of diabetes; their high positive SHAP values suggest that they influence the model strongly in the direction of a diabetic diagnosis. This means that the symptoms present in the patient's data provide most of the assurance to the model about the presence of diabetes.

Relative to these issues, the plot also shows how one component relates to another and how the overall spectrum of features cooperatively contributes to the prediction. Converting from the blue to the red area exemplifies the fact that the intraject aggregates of the numerous features steadily advances from the projection of a non-diabetic to a diabetic one. The increments of vertical distance in each layered line correspond to the datapoint, describing the impact of variation in feature values on the prediction. The force plot makes it easier for you to explain which features play a more significant role in the model's prediction. Another reason is that SHAP values explain how the model arrived at certain conclusions, providing healthcare professionals with confidence in the predictions that are most beneficial for making informed clinical decisions.

Table 6. Evaluation of Machine Learning Models with 95% Confidence Intervals

ML Model	Accuracy (%)	95% CI (Accuracy)
Naïve Bayes	90.21	[87.7, 92.8]
KNN	90.65	[88.1, 93.1]
Decision Tree	89.34	[86.6, 92.0]
SVM	91.3	[88.9, 93.7]
Stacked Ensemble	98.4	[97.3, 99.5]

Table 7. Comparative analysis with other research works

Reference	Accuracy (%)
10	94.20
11	96.88
12	92.00
15	82.00
Our study	98.40

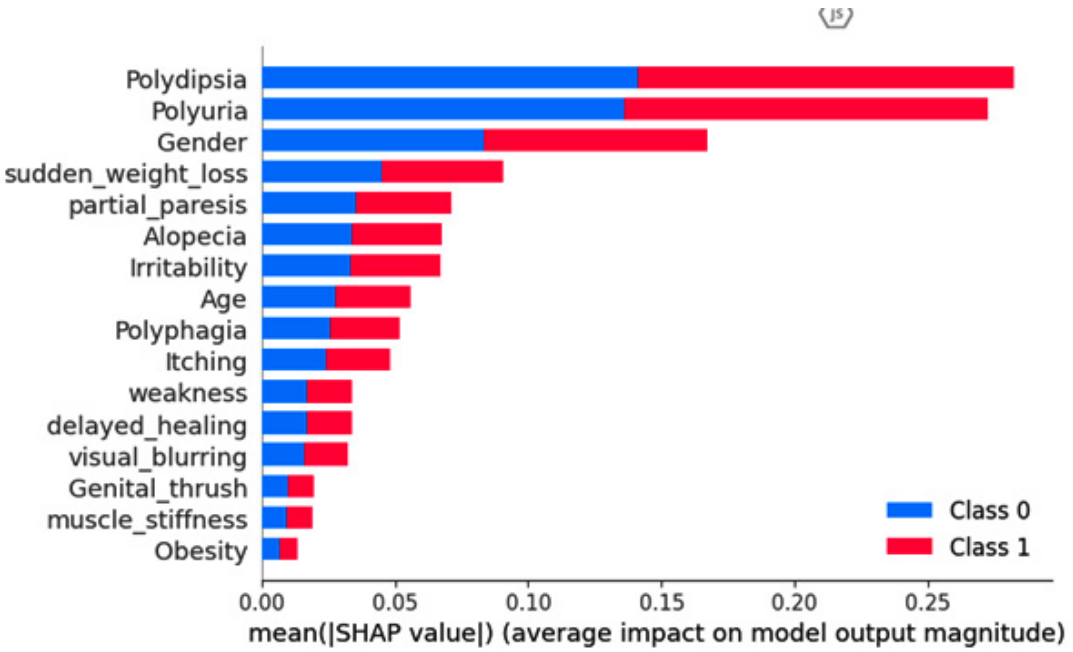

Fig. 4. Feature Contribution Analysis of Stacked Ensemble Model

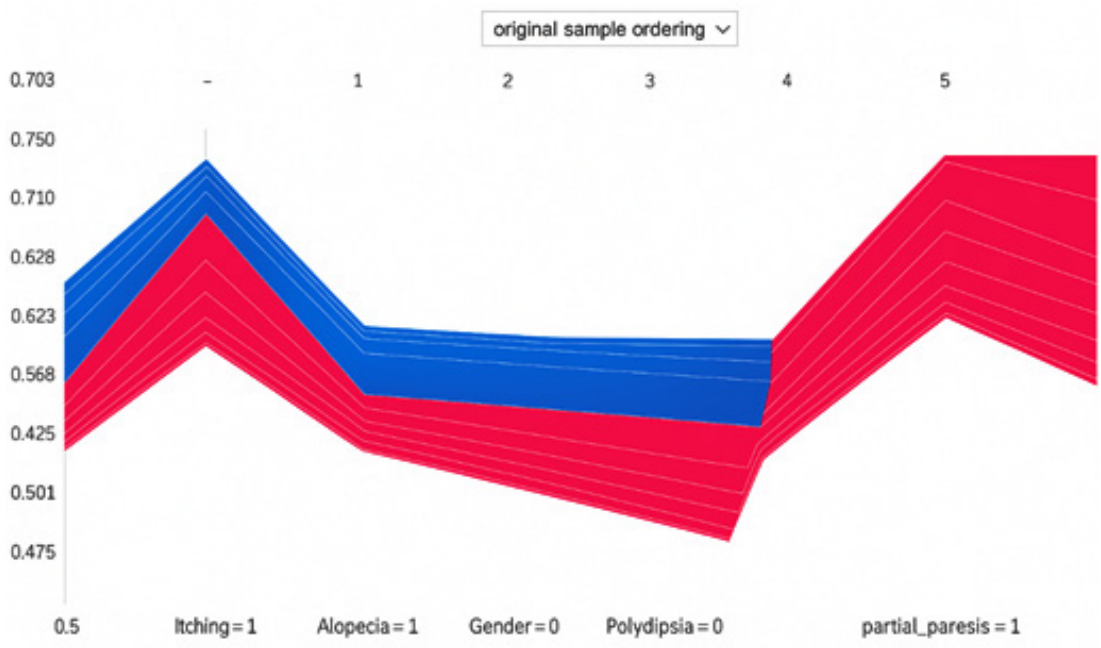

Fig. 5. SHAP Force Plot Visualizing Feature Contributions in Diabetes Prediction Model

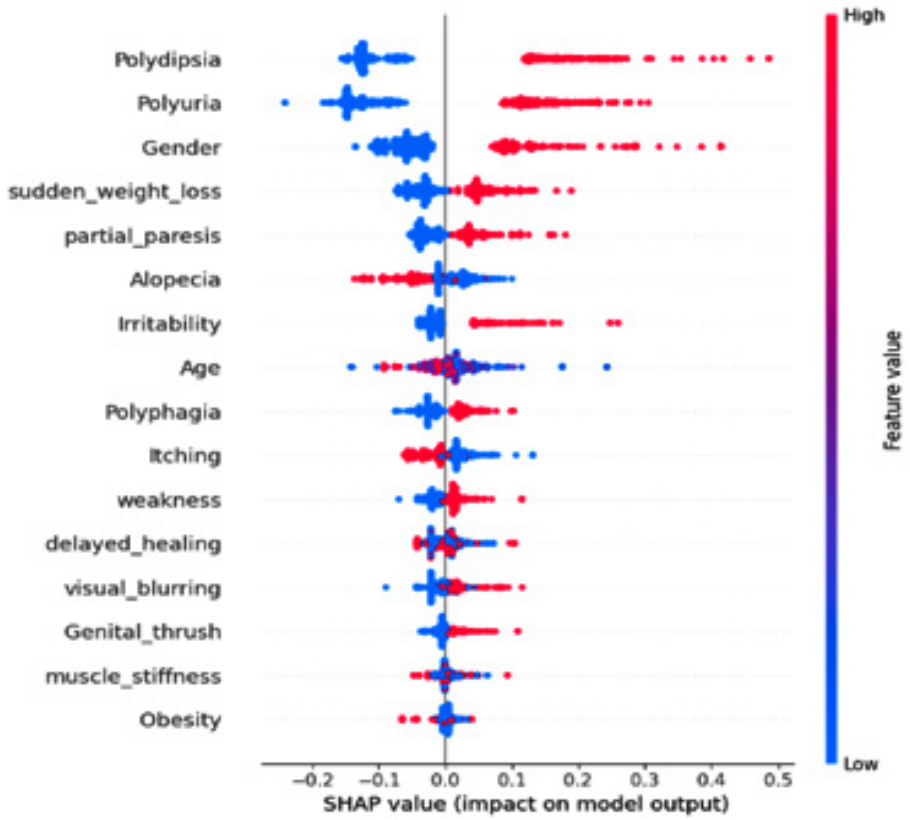

Fig. 6. SHAP Summary Plot for Diabetes Prediction using Stacked Ensemble Learning

Integrating this SHAP analysis into our research paper emphasizes how crucial it is to apply explainable ML to healthcare use cases. This shows how clinically crucial symptoms such as 'Polydipsia' and 'Polyuria' can be emphasized using aerospace SHAP values, and the data is beneficial for clinical professionals. It not only enhances the reliability of the model's predictions but also helps uncover the underlying attributes behind the forecasts, paving the way for more individualized and applicable healthcare initiatives.

The SHAP (Shapley Additive exPlanation) summary plot, displayed in Figure 6, shows the contribution of each feature to the development of a diabetes prediction model constructed using a stacked ensemble learning technique. Figure 6 illustrates each point as a SHAP value for an instance in the dataset, representing the contribution of the feature to the outcome. Features are displayed on the y-axis, arranged in order of their contributions, with higher contributing features positioned higher. The x-axis represents the SHAP value; if it has a positive value, this means that the respective parameter increases the probability of diabetes, and if it turns out to be negative, it decreases it.

The part of each dot that radiates a yellow or blue color indicates the importance of a corresponding feature to the result. Blue indicates that it has no influence, and red indicates that it has a strong influence. For instance, two symptoms, "Polydipsia" and "Polyuria", are marked by red color on the right side, indicating that they are very relevant to diabetes. This raises the tendency that if these symptoms are expressed to a higher degree, then one can be diagnosed with the disease. On the other hand, there are some attributes that neither significantly contribute to the prediction nor inhibit it, such as Age and Gender.

The summary Figure 7 also illustrates that the features do not exhibit a direct relationship with the prediction. For example, relative to the "sudden_weight_loss" feature, it means that higher values in this feature contribute significantly to diabetes prediction, as indicated by the red result values on the right side of the chart. However, lower values do not have a significant impact, as the blue points are close to zero. Having such detailed visualization is useful for interacting with the model's output and finding correlations between different features of diabetes symptoms and demographic factors that should interest clinicians and researchers.

LIME allows users to explain how a given complex model of ML decides a prediction. For each case, it reduces the model's complexity and provides a reason why such a forecast is expected. In this study, we used LIME to understand why a diabetes prediction model gave specific results for a patient. On the left, the prediction probabilities indicate that the model predicts a 98% chance of the individual being non-diabetic (Diabetes -ve) and a 2% chance of being diabetic (Diabetes +ve). This indicates a high level of confidence in predicting that the individual does not have diabetes. The high probability for non-diabetic suggests that the influential features mostly point away from a diabetes diagnosis.

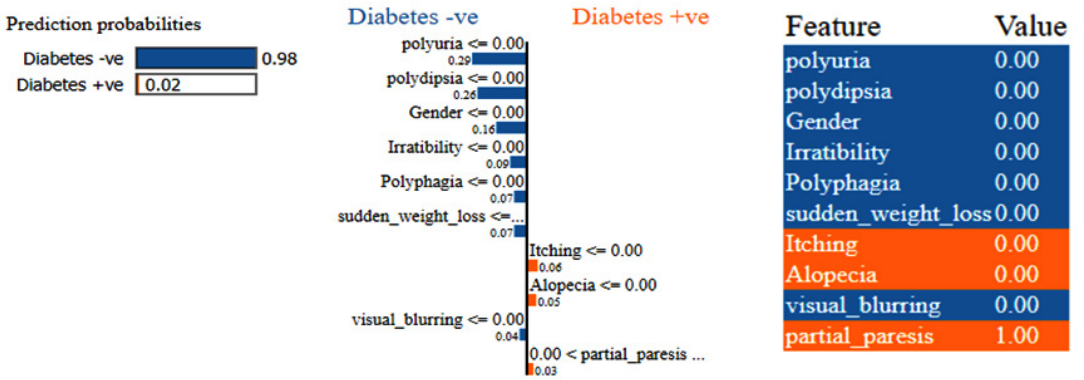

Fig. 7. LIME Explanation for Diabetes Prediction using Stacked Ensemble Learning

The middle section shows the feature contributions to the prediction. Features on the left side (under "Diabetes -ve") are contributing towards the non-diabetic prediction, while features on the right side (under "Diabetes +ve") are pushing towards a diabetic prediction. In this specific instance, "partial_paresis" is the only feature on the diabetic side, contributing slightly towards a positive diabetes prediction. However, features such as "polyuria," "polydipsia," "Gender," "Irritability," "Polyphagia," "sudden_weight_ loss," "Itching," "Alopecia," and "visual_blurring" are all contributing towards the non-diabetic side.

The right section lists the actual values of the features for this instance. Most features have a value of 0, indicating that these symptoms or conditions are absent in the individual. The presence of "partial_paresis" with a value of 1 is noted, but it is insufficient to significantly influence the prediction towards diabetes. The combination of the absence of key diabetic symptoms and conditions leads to a strong prediction against diabetes for this individual. The LIME explanation thus provides a clear and interpretable insight into why the model predicted the individual as non-diabetic, highlighting the predominant influence of the absence of several diabetic symptoms. By explaining how each factor influences the model's outcome, we increase the model's transparency and reliability, helping healthcare providers make better, personalized treatment choices. The model's impressive results are due to its use of multiple prediction models working together. Each model was carefully adjusted to be as accurate as possible, and combining them with a meta-classifier further improves the overall prediction ability. The high accuracy rate achieved by our model is clear evidence of its appropriateness in early screening and management of potential diabetic patients. Screening for diabetes is essential to prevent or slow the progression of severe complications related to the disease. Attributable to the high rate of specificity coincidence made possible by the models, healthcare professionals and patients can prevent further deterioration of health by incorporating enhanced health precaution methods, such as modifying reckless lifestyle habits and adjusting medication, among other measures.

The novelty of our work lies in developing ensemble learning techniques that consider the disadvantages of each algorithm. This not only drives the level of forecast precision but also increases the robustness of the model. This particular methodology is particularly useful in the case of diabetes because the ensemble model can handle the different risk factors and signs of the disease. Our study provides guidance for future research in three primary areas. Firstly, our ensemble model can now be applied to other datasets to further determine its suitability when used universally. The performance of a model often therefore depends on the data to which it has been applied. It is therefore useful to attempt to subject the strategy to several different forms of data as a means of establishing the reliability and flexibility of this tool.

The model was coded in Python using the scikit-learn library. The code shall be publicly posted when accepted / available to the authors on reasonable request.

Clinical Utility of Explainable AI

The clinical usefulness of explainable AI is demonstrated by the outcomes of the previously mentioned studies. Although SHAP and LIME offer a post-hoc description of the suggested model, their clinical implication should be taken into consideration, especially. The fact that Polydipsia and Polyuria are the most powerful characteristics fits the existing clinical recommendations concerning early diagnosis of diabetes, since excessive thirst and frequent urination are all known warning indicators. This correspondence enhances clinicians' confidence in the model and facilitates its adoption in clinical decision support systems. SHAP-driven feature attributions can be employed to display prioritized symptom relevance and predicted risk scores to allow the clinician to rationalize model suggestions and give priority to additional diagnostic measurements. Although the proposed model exhibits good performance, it is essential to consider potential sources of bias and low generalizability. The Early-Stage Diabetes Risk Prediction data is symptom-based and on a smaller scale and is based on one public repository. All these factors might create sampling bias and limit the model's generalizability to larger or more diverse populations. Therefore, the described findings might not be completely applicable to clinical practice. Future research would aim to test the proposed method on external and multi-centre

data, and to incorporate additional demographic and laboratory attributes to enhance resilience and impartiality.

Limitations

Despite the strong performance of the proposed stacked ensemble model, several limitations should be acknowledged. First, the evaluation was conducted on a single public dataset (UCI Early-Stage Diabetes Risk Prediction), which may limit the generalizability of the results to broader and more diverse populations. Second, the dataset is primarily symptom-based and does not include biochemical or laboratory measurements such as HbA1c, fasting plasma glucose, or oral glucose tolerance test results, which are clinically important for definitive diabetes diagnosis. Third, the relatively small dataset size, combined with the high accuracy achieved, introduces a potential risk of overfitting, despite the use of cross-validation and ensemble learning techniques. These limitations highlight the need for cautious interpretation of the results and further validation in real-world clinical settings.

CONCLUSION AND FUTURE WORK

Before the progression to full-blown diabetic complications, the stacked ensemble model of clinical decision-making can be used to make layered predictions. In the proposed method, multiple ML classifiers are employed to compile additional information and build the optimal model. Based on the examination of the model, this study uses a stacked ensemble model as a diagnostic tool for early-stage diabetes. Compared to commonly used models such as NB, KNN, SVM, and DT, the proposed stacked ensemble model outperformed them, achieving a high precision value of 98.4%. The primary objective of the proposed work was to develop a more effective approach that enhances the precision of the trained model when trained on a large dataset. By achieving such highly accurate results, this work contributes to the establishment of a portable decision system for the early detection of diabetes. Moreover, our proposed ensemble learning strategy, combined with SHAP analysis augmentation, enhances the model's credibility, providing valuable insights for identifying probable diabetic patients at the preliminary stage. The high accuracy of the model suggests its use as the key instrument to approach diabetes management ahead of time, enhancing the outcome. This system would afford accurate and dependable results, considering the same risk factors as the other models. The advantages of proceeding with this strategy are numerous, as early identification of the disease is always a positive step forward, with the potential to positively impact the lives of patients.

Although the suggested model exhibits high predictive accuracy, the current study is limited by a single, symptom-based dataset with a small sample size. Further research will be conducted on external validation using larger, multi-centre, and clinically heterogeneous datasets, which will encompass symptom-based features as well as biochemical measurements, including HbA1c and fasting glucose. The future directions will seek to use different combinations of classifiers, hyperparameter optimization, and sophisticated feature-selection methods. Additionally, the possibility of real-time implementation in telehealth and clinical decision support systems, as well as the extension of the stacking-based framework to other chronic diseases, will be explored to assess its applicability and clinical effectiveness.

ACKNOWLEDGEMENT

The author would like to thank Bennett University, Science, Technology and Technical Education Department, and Galgotias University for providing all the necessary facilities to carry out this research work.

Funding sources

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The author(s) do not have any conflict of interest

Data Availability Statement

This statement does not apply to this article.

Ethics Statement

This research did not involve human participants, animal subjects, or any material that requires ethical approval.

Informed Consent Statement

This study did not involve human participants, and therefore, informed consent was not required.

Clinical Trial Registration

This research does not involve any clinical trials.

Permission to reproduce material from other sources

Not Applicable.

Author Contributions

Shahnawaz Ahmad: Conceptualization, Methodology, Supervision, Writing – Original Draft; Shahadat Hussain: Data Collection, Analysis, Writing – Review & Editing; Mohd. Arif: Visualization, Funding Acquisition, Writing – Original Draft; Mohd. Aquib Ansari: Resources, Supervision, Writing – Original Draft.

REFERENCES

1. Kearns M, Valiant LG. Learning Boolean formulae or finite automata is as hard as factoring. Technical Report, Cambridge, MA: Harvard University Aiken Computation Lab . 1988, TR-14-88.
2. Wolpert DH. Stacked generalization. *Neural Network.* 1992;5(2):241-259.
3. Schapire R. The strength of weak learnability. *Mach Learn.* 1990;5(2):197-227.
4. Ali S, Majid A. Can-evo-ens: classifier stacking based evolutionary ensemble system for prediction of human breast cancer using amino acid sequences. *J Biomed Inform.* 2015;54:256-269.
5. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. *J Intell Learn Syst Appl.* 2017;9(1):1-7.
6. Aliæ B, Gurbeta L, Badnjeviæ A. Machine learning techniques for classification of diabetes and cardiovascular diseases. In: *Proc 6th Mediterranean Conf Embedded Comput (MECO).* Piscataway, NJ: IEEE; 2017:1-4.
7. Kumar A, Kumar P, Srivastava A, Kumar VA, Vengatesan K, Singhal A. Comparative analysis of data mining techniques to predict heart disease for diabetic patients. In: *Int Conf Advances Comput Data Sci.* Singapore: Springer; 2020; pp. 507-518.
8. Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. *Int J Emerg Tech Innov Eng.* 2019; 5(4):1-5.
9. Sikder N, Masud M, Bairagi AK, et al. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry (Basel).* 2021;13(4):670.
10. El-Sappagh S, Elmogy M, Ali F, et al. A comprehensive medical decision-support framework based on a heterogeneous ensemble classifier for diabetes prediction. *Electronics (Basel).* 2019;8(6):635.
11. Banchhor M, Singh P. Comparative study of ensemble learning algorithms on early stage diabetes risk prediction. In: *Proc 2nd Int Conf Emerging Technol (INCET).* Piscataway, NJ: IEEE; 2021:1-6.
12. Sabbir MMH, Sayeed A, Jamee MAUZ. Diabetic retinopathy detection using texture features and ensemble learning. In: *Proc IEEE Region 10 Symp (TENSYMP).* Piscataway, NJ: IEEE; 2020:178-181.
13. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cogn Comput Eng.* 2021;2:40-46.
14. Singh N, Singh P. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybern Biomed Eng.* 2020;40(1):1-22.
15. Abdulhadi N, Al-Mousa A. Diabetes detection using machine learning classification methods. In: *Proc Int Conf Inf Technol (ICIT).* Piscataway, NJ: IEEE; 2021:350-354.
16. Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J Diabetes Metab Disord.* 2020;19(1):391-403.
17. Al-Zebari A, Sengur A. Performance comparison of machine learning techniques on diabetes disease detection. In: *Proc 1st Int Informatics Software Eng Conf (UBMYK).* Piscataway, NJ: IEEE; 2019:1-4.
18. Rawat V, Suryakant S. A classification system for diabetic patients with machine learning techniques. *Int J Math Eng Manag Sci.* 2019;4(3):729-744.
19. Morgan-Benita JA, Galván-Tejada CE, Cruz M, et al. Hard voting ensemble approach for the detection of type 2 diabetes in Mexican population with non-glucose related features. *Healthcare (Basel).* 2022;10(8):1362.
20. Atif M, Anwer F, Talib F. An ensemble learning approach for effective prediction of diabetes mellitus using hard voting classifier. *Indian J Sci Technol.* 2022;15(39):1978-1986.
21. Islam MM, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In: *Comput Vis Mach Intell Med Image Anal.* Cham: Springer; 2020:113-125.

22. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem.* 1993;39(4):561-577.
23. Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. *Adv Neural Inf Process Syst.* 1994;7:1063-1067.
24. Kumari M, Ahlawat P. DCPM: an effective and robust approach for diabetes classification and prediction. *Int J Inf Technol.* 2021;13:1079-1088.
25. Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *Int J Inf Technol*. 2020;12:419-428.
26. Sharma A, Mishra PK. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *Int J Inf Technol.* 2022;1-12.
27. Bhavekar GS, Goswami AD. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. *Int J Inf Technol.* 2022;14(4):1781-1789.
28. Mishra I, Mohapatra S. An enhanced approach for analyzing the performance of heart stroke prediction with machine learning techniques. *Int J Inf Technol.* 2023;1-14.
29. Annamalai R, Nedunchelian R. Design of optimal bidirectional long short term memory based predictive analysis and severity estimation model for diabetes mellitus. *Int J Inf Technol.* 2023;15(1):447-455.
30. Atif M, Anwer F, Talib F, Alam R, Masood F. Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage. *IAES Int J Artif Intell.* 2023;12(3):1302-1311.
31. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Comput Sci.* 2018;132:1578-1585.
32. Maniruzzaman M, Rahman MJ, Ahammed B, et al. Comparative approaches for classification of diabetes mellitus data: machine learning paradigm. *Comput Methods Programs Biomed.* 2017;152:23-34.
33. Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Inform Med Unlocked.* 2019;17:100179.
34. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord.* 2019;19(1):29.
35. Pradhan N, Rani G, Dhaka VS, Poonia RC. Diabetes prediction using artificial neural network. In: *Deep Learning Techniques for Biomedical and Health Informatics.* Amsterdam: *Elsevier*; 2020:327-339.
36. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Inform Med Unlocked.* 2018;10:100-107.
37. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH. Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district Bandipora. *Comput Intell Neurosci.* 2022;2022:2789760.
38. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors (Basel).* 2022;22(19):7268.
39. Bhat SS, Selvam V, Ansari GA, Ansari MD. Analysis of diabetes mellitus using machine learning techniques. In: *Proc 5th Int Conf Multimedia, Signal Process Commun Technol (IMPACT).* Piscataway, NJ: IEEE; 2022:1-5.
40. Patil RN, Rawandale S, Rawandale N, Rawandale U, Patil S. An efficient stacking based NSGA-II approach for predicting type 2 diabetes. *Int J Electr Comput Eng.* 2023;13(1):1015-1023.
41. Bhat SS, Selvam V, Ansari GA, Ansari MD. Hybrid prediction model for type-2 diabetes mellitus using machine learning approach. In: *Proc 7th Int Conf Parallel, Distrib Grid Comput (PDGC).* Piscataway, NJ: IEEE; 2022:150-155.
42. Ijaz MF, Alfian G, Syafrudin M, Rhee J. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. *Appl Sci (Basel).* 2018;8(8):1325.
43. Sharma, A., Dalmia, R., Saxena, A., & Mohana, R. A stacked deep learning approach for multiclass classification of plant diseases. *Plant and Soil*, 2025;*506*(1), 621-638.