AUIQ Technical Engineering Science

Manuscript 1063

Explainable Tree-Based Ensemble Models for Diabetes Prediction Using SHAP

Maan Y Anad Alsaleem

Omar Shakir Hasan

Yahya Albugg

AUIQ TECHNICAL ENGINEERING SCIENCE 2026;3:35–45

ORIGINAL STUDY

Explainable Tree-Based Ensemble Models for Diabetes Prediction Using SHAP

Maan Y Anad Alsaleem [a,*], **Omar Shakir Hasan** [a], **Yahya Albugg** [b]

[a] Directorate of Education in Nineveh
[b] Northern Technical University(NTU)

ABSTRACT

Due to the generally unqualified nature of prediction data and the difficulty of interpreting predictions, predicting diabetes remains a significant hurdle in the adoption of machine learning within the medical domain. In this study, several tree-based machine learning techniques (LightGBM, XGBoost, CatBoost, and Gradient Boosting) were applied to predict diabetes using the 2015 BRFSS dataset, while two ensemble methods (soft voting and stacking) were employed to improve predictive accuracy. The performance analysis of the individual models and ensemble approaches indicates that CatBoost achieved the highest accuracy among the single classifiers (0.871), with an F1-score of 0.871 and a ROC–AUC of 0.921. Both ensemble methods further improved predictive performance compared with individual models. The soft voting ensemble obtained an overall accuracy of 0.878 and a ROC–AUC of 0.928, whereas the stacking ensemble achieved the highest overall performance, with an accuracy of 0.883, an F1-score of 0.883, and a ROC–AUC of 0.934. Moreover, SHAP-based analysis identified general health, body mass index, and age as the most influential factors affecting the prediction outcomes. We conclude that ensemble learning improves predictive performance while still providing a level of interpretability when assessing risk for developing diabetes.

Keywords: Diabetes prediction, Ensemble learning, Tree-based models, SHAP

1. Introduction

Over the last several years, machine learning (ML) techniques have increasingly been used in the healthcare industry to enable data-based decisions, as well as the creation of predictive models that enhance clinical treatment [1, 2]. As a result, one area of increasing interest is the ability to predict chronic diseases such as diabetes; diabetic disease has become an international public health issue with a rise in the number of people diagnosed and the negative effect this will have on healthcare systems across the globe [3]. Accurate and timely prediction of diabetes will allow for timely interventions, improved patient outcomes, and a better opportunity for healthcare resources to be allocated efficiently [4]. A variety of ML methods exist to predict diabetes using structured health data sets, including decision trees, boosting algorithms, and ensemble learning methods [5, 6]. The ensemble learning models, such as LightGBM, XGBoost, Cat boost and Gradient boosting have gained a lot of attention from the research community due to their ability to deliver high levels of predictive accuracy in the analysis of imbalanced datasets while simultaneously being able to deal with missing data and also analysing high dimensional dataset [7–9]. Although these models have strong predictive capacity, a significant limitation associated with ensemble learning models is their inability to provide interpretability, which impacts their ability to be accepted into clinical practise [10]. To foster AI-enabled Decision Support Systems (DSS), Healthcare professionals require an explicit understanding of how the AI-DSS arrives at recommendations through an explicit

Received 16 December 2025; revised 20 January 2026; accepted 6 February 2026.
Available online 19 May 2026

* Corresponding author.
E-mail addresses: maanyounis1983@gmail.com (M. Y. A. Alsaleem), omarshakir06@gmail.com (O. S. Hasan), dr.yahya.albugg@ntu.edu.iq (Y. Albugg).

https://doi.org/10.70645/3078-3437.1063

visualisation of the key features that influence the predicted outcome [11, 12]. Sufficient comprehension of which features are most relevant to making a diagnosis of Diabetes allows for the development of better clinical practices to assist physicians and patients alike during their shared decision-making process [13]. The challenge of providing interpretable explanations behind an AI system's reasoning capabilities has been solved by the use of Explainable AI (XAI) methods. A popular technique that falls under the umbrella of XAI is SHapley Additive exPlanations (SHAP). SHAP is based upon a theory of cooperative game theory that by using calculated contributions of various input features assists in determining how much of the change in predicted outcomes could be attributed to each of the input features [14, 15]. The use of SHAP has enabled the interpretation of both the global and local predictions of individual machine learning (ML) models, in particular tree-based classifiers. This paper is focused on the investigation of the SHAP approach as an avenue for improving the interpretability of tree-based ensemble ML models trained on the publicly available 2015 BRFSS dataset containing balanced binary labels for Diabetes diagnoses. This paper contributes to the interpretation of voting results and ensembles using SHAP software. SHAP values are integrated across base models, and the interpretability of the base models and ensembles is compared.

2. Related works

In healthcare, particularly in the prediction and treatment of diabetes, the use of explainable Artificial Intelligence techniques (EAIs) has become more common over the past couple of years, with one of the more popular methods being called the Shapley Additive Explanations (SHAP) technique. A framework for developing interpretable machine learning models that implements the Synthetic Minority Oversampling Technique (SMOTE) for balancing classes within datasets, coupled with the SHAP approach to explaining how predictive algorithms function, has been developed and applied to diabetes prediction datasets, yielding good results regarding its clinical usefulness as well as predictive accuracy [16]. Research comparing multiple models of EAI (Explainable AI) using the Shapley values method clearly demonstrated both good interpretability and good predictive accuracy for Type-2 diabetes prediction regression models [17]. Another independent study applied the SHAP methodology to select features that had the highest predictive potential for diabetes risk. The study found that the use of a boosting algorithm such as LightGBM or XGBoost in conjunction with the SHAP technique provided superior predictive results over the other methods evaluated while also providing the highest interpretability of the key contributing features (such as Age and BMI) contributing to predicted risk [18]. Additionally, the study of Shapley values in conjunction with traditional machine learning prediction algorithms is also ongoing; for example, a cohort study combined the SHAP approach with multiple regression and reported new insights into potential biomarkers of disease progression that could be targeted for intervention [19]. The development of an ensemble framework incorporating multiple machine learning (ML) models and explainable artificial intelligence (XAI) tools, provides an avenue to improve the accuracy and interpretability of ML prediction models through the identification of key clinical predictors (e.g., physical activity, general health) [20]. A follow-up study presented an ensemble-based prediction model that utilised transparent techniques to provide an accurate ML prediction model and the use of SHAP for post-hoc model interpretation of feature impact [21]. Using SHAP and LIME together for risk prediction systems constructed on large population health datasets like the BRFSS was demonstrated to allow clinicians and patients to visualise the contribution of various factors to each person's risk [22]; an example can be found in diabetes prediction research that combined SHAP with XGBoost models using feature selection methods such as recursive feature elimination (RFE) to determine clinically important factors to improve model transparency [23]. Recent studies using cohort data have examined using explainable AI techniques for XGBoost and CatBoost methods in predicting type-2 diabetes with high predictive validity and interpretability [24].

Despite the progress in applying explainable machine learning to individual models, limited attention has been given to interpreting ensemble-based models such as voting and stacking classifiers. This study addresses this gap by providing framework for interpreting these meta-models using SHAP, offering insights into their combined decision-making process.

3. Methodology

Fig. 1 presents an overview flowchart of the proposed methodological framework. It outlines the stages of data acquisition (BRFSS 2015 diabetes), preprocessing, independent experimentation with four tree-based models, explainability via SHAP, and ensemble construction (soft voting and stacking) followed by evaluation.

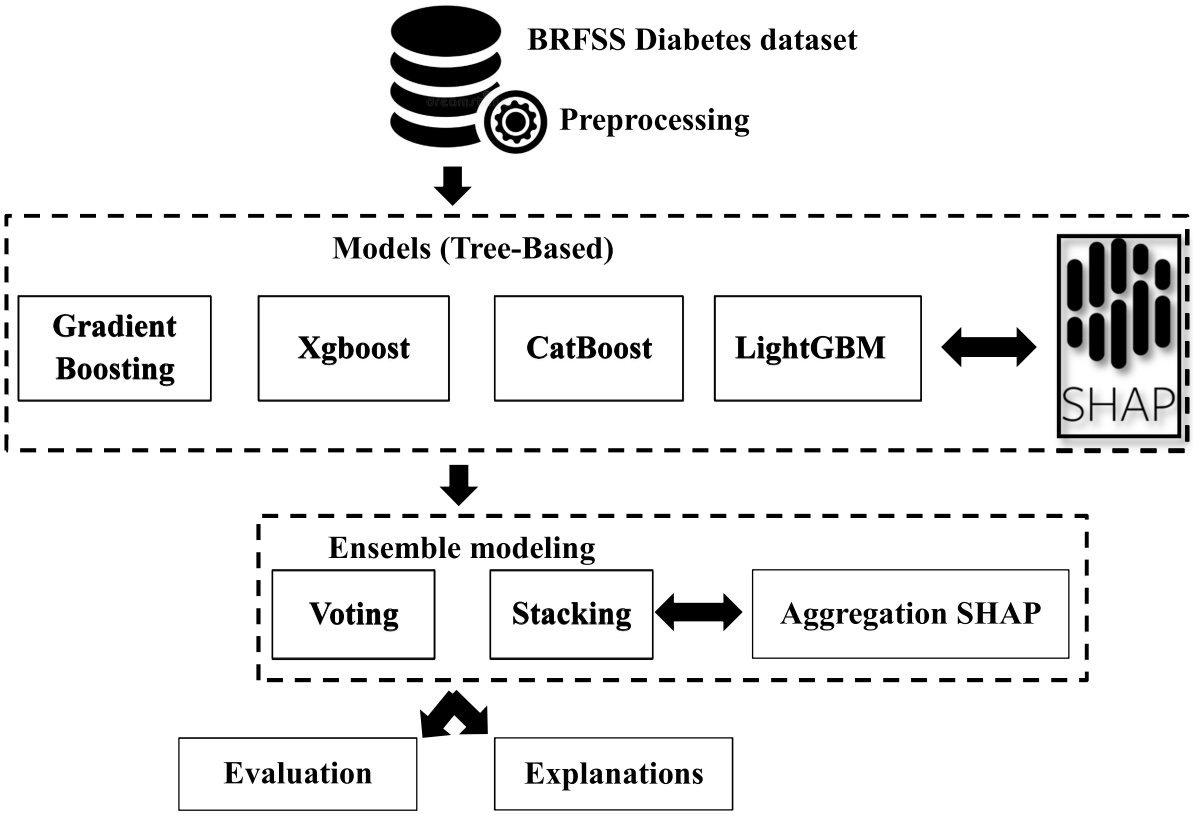

Fig. 1. Overview of the methodological workflow.

3.1. Dataset

We used the BRFSS 2015 health indicators Balanced (50/50 split) Dataset for developing our Diabetes Classification Model. The dataset consists of 70,692 Records of Adult Respondents and 21 Columns of Predictor Variables (including columns for demographic information, health-related information and lifestyle/risk factor information) (See Table 1). Each entry in this file represents an Adult Respondent and contains a set of attributes covering their Health status, Demographics, and Lifestyle/risk factors. The Target Variable diabetes_binary takes a value of (0) non-diabetic or (1) diabetic/prediabetic and is indicated by the dataset having an equal number of cases in each category (i.e., it has a balanced dataset). The target variable was converted into integer format (0 for non-diabetic and 1 for diabetic cases). The dataset was inspected for missing values and inconsistencies, and no missing entries or abnormal values were detected; therefore, no imputation or outlier treatment was required. The dataset was split into training and testing sets using an 80/20 stratified strategy to preserve class distribution. A fixed random seed (random_state = 42) was used to ensure reproducibility.

3.2. ML models (Tree-based)

Four tree models were used in the basic forecasting phase. Each model is trained independently on the training set, then evaluated.

LightGBM–A Boosting Learning System based off how we add to our prediction's leaf effect (the classification as yes or no) based on the weight of the Leaf Node. It uses Histogram data as well as the 1st Order statistics and 2nd order Statistics as the function of calculation to speed up the optimization process of building the tree.

XGBoost–Is a Boosting Learning System based on how we learn to determine the left/right balance of each tree by adding a tree to an already predicted tree (or adding additional predictions into a previously determined tree). Additionally, it uses row-based or column-based data and learning-rate reductions during the training of the model to add consistency.

CatBoost–A Boosting Learning System based on Symmetric Trees. It includes built-in handling of categorical data using Target Statistics and ordered tree construction to minimize loss from overfitting. Its training remains consistent when fixed seed values are used and minimizes the need for additional data processing prior to using mixed-feature sets.

Table 1. Predictor features and brief descriptions.

Feature	Type	Brief description
HighBP	Binary (0/1)	Ever told high blood pressure.
HighChol	Binary (0/1)	Ever told high cholesterol.
CholCheck	Binary (0/1)	Cholesterol checked in past 5 years.
BMI	Numeric	Body Mass Index.
Smoker	Binary (0/1)	Ever smoked ≥100 cigarettes.
Stroke	Binary (0/1)	Ever told had a stroke.
HeartDiseaseorAttack	Binary (0/1)	Coronary heart disease or myocardial infarction.
PhysActivity	Binary (0/1)	Any physical activity/exercise in past 30 days
Fruits	Binary (0/1)	Consumes fruit ≥1 time/day.
Veggies	Binary (0/1)	Consumes vegetables ≥1 time/day.
HvyAlcoholConsump	Binary (0/1)	Heavy alcohol consumption indicator.
AnyHealthcare	Binary (0/1)	Has any kind of health care coverage.
NoDocbcCost	Binary (0/1)	Could not see a doctor due to cost (past 12 months).
GenHlth	Ordinal (1–5)	General health (1=Excellent . . . 5=Poor).
MentHlth	Numeric (0–30)	Days of poor mental health (past 30 days).
PhysHlth	Numeric (0–30)	Days of poor physical health (past 30 days).
DiffWalk	Binary (0/1)	Serious difficulty walking or climbing stairs.
Sex	Binary (0/1)	Sex of respondent.
Age	Ordinal (coded bands)	Age category (ascending ordered codes).
Education	Ordinal (coded levels)	Highest education level (ascending ordered codes).
Income	Ordinal (coded bands)	Annual household income category (ascending ordered codes).

Gradient Boosting-A boosting Learning System that builds on the Prediction Tree through repeated training by fitting the Shallow Tree's' predictions on the negative gradient or "reaction" effect of the chosen Loss Function (The more Stable the Loss Function the Easiest the Prediction). Gradient Boosting offers consistent (steady) predictions based on the Hyper-Parameter constants. There is also no specific processing for categorical features so this could be a good model for multi-class/multi-label use cases.

For each model, post-hoc explainability is conducted with SHAP using TreeExplainer to obtain (i) global importance (summary/beeswarm and bar plots) and (ii) local rationale (force plots for representative instances).

Table 2 illustrates the basic classification models along with their settings to ensure reproducibility. All models were chosen to be either tree-based or boosted, allowing the SHAP TreeExplainer framework to be directly applied across all experiments. The hyperparameters were manually selected to control model complexity and ensure stable training behavior. A fixed number of trees (n_estimators = 200) was used to provide sufficient model capacity while avoiding excessive growth of the ensemble. Shallow tree depths were adopted to limit overfitting and improve generalization. Learning rates in the range of 0.03–0.1 were employed to allow gradual model updates and reduce sensitivity to noise. These settings were chosen to maintain consistency across models and to support reproducible comparison.

3.3. Ensemble model

The Ensemble combines four different types of tree-based Learner: LightGBM, XGBoost, CatBoost, and Gradient Boosting with the methods of soft voting and stacking. Soft voting combines the individual models' (i.e., LightGBM, XGBoost, CatBoost, and Gradient Boosting) prediction by averaging their predicted positive class expectations using either the equal-weight average of the predicted class probabilities or weighted averages of the predicted class probabilities. These averages provide a stable and straightforward means of arriving at a combined prediction. Stacking, in contrast, uses the outputs of base learners' probability values for training a logistic regression model to learn how to readjust the complement error patterns of the base learners using the outputs from the base learners (i.e., out of fold

Table 2. Base classifiers and baseline configuration.

Model	Library	Selected hyperparameter
LightGBM	lightgbm	n_estimators = 200, learning rate in [0.03–0.1]
XGBoost	xgboost	n_estimators = 200, max_depth = 6, eval_metric = logloss
CatBoost	catboost	Depth = 4, learning_rate = 0.03–0.1
Gradient Boosting	sklearn	n_estimators = 200, max_depth = 4

predictions). Using the same splits of data to evaluate the performance of both the ensemble learners and the individual learners and by aggregating the SHAP importance scores from each base learner for soft voting and using the inputs from the base learners when fitting the meta learner to aggregate SHAP values for all inputs from the base learners for stacking.

3.4. Evaluation measures

Performance is measured on the held-out test set. We report Accuracy and F1-score as primary metrics; ROC–AUC and the confusion matrix are optionally included. Let $TP,\ TN,\ FP,\ and\ FN$ denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall proportion of correctly classified instances among all samples:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

Precision quantifies the proportion of correctly predicted positive cases among all predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall reflects the model's ability to correctly identify positive cases:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score represents the harmonic mean of Precision and Recall, providing a balanced measure of classification performance:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Receiver Operating Characteristic – Area Under Curve (ROC–AUC)

ROC–AUC evaluates the discriminative ability of the model across all possible classification thresholds. It is defined as the area under the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

3.5. Explainable ML using SHAP

The SHAP method is based on cooperative game theory, as well as using this method to quantify the contribution of all of the model's inputs to the model's prediction, so that a model's prediction can be decomposed into a baseline prediction and the effect of each feature. By decomposing a model's prediction, the SHAP method allows for consistent and comparable predictions of models from different datasets.

Formally, for a trained model $f(x)$ and an input instance $x = (x_1, x_2, \ldots, x_M)$ with M features, the SHAP formulation expresses the prediction as:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

Where $\phi_0 = \mathbb{E}[f(\mathbf{x})$ represents the expected model output over the dataset (baseline), and ϕ_i denotes the SHAP value associated with feature iii, reflecting its marginal contribution to the prediction. Each SHAP value is computed as the weighted average of the feature's contribution across all possible feature subsets:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f(S \cup \{i\}) - f(S) \right]$$

where F is the full set of features and S represents a subset excluding feature i. This formulation ensures that feature contributions are fairly distributed based on all possible coalitions.

The calculation of TreeExplainer SHAP values for all of the following tree-based classifiers: Gradient Boosting, LightGBM, CatBoost, and XGBoost, which is matched with an set of SHAP values from TreeExplainer given the requirements for decision-tree ensembles. The use of TreeExplainer is limited to tree-based classifiers and produces either an exact SHAP value or an approximate one that maintains consistency with the Shapley values that are used in building TreeExplainer. The SHAP analysis is only conducted on the held-out test set in order to avoid any potential information leakage, which ensures that any findings will truly represent the generalization behavior and not some artifacts from the training process.

Local interpretability is concerned with providing explanations for individual predictions. Specifically, for each test instance SHAP values show the contribution of each feature in terms of increasing/decreasing likelihood of diabetes (compared to baseline output). Instance level explanations are provided visually through the use of SHAP force plots wherein positive SHAP values will push the prediction further into the diabetic class, and negative SHAP values will pull it back away from that class. The ability to provide an explanation for an individual instance allows for a more patient-centred interpretation to be made, and supports clinical reasoning by relating model outputs

Table 3. Overall performance comparison of all models.

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
LightGBM	0.862	0.859	0.866	0.862	0.912
XGBoost	0.868	0.871	0.864	0.867	0.918
CatBoost	0.871	0.874	0.869	0.871	0.921
Gradient Boosting	0.851	0.847	0.856	0.851	0.904
Voting Ensemble	0.878	0.876	0.880	0.878	0.928
Stacking Ensemble	0.883	0.885	0.881	0.883	0.934

to individual health and lifestyle information that are known.

Global interpretability is an overview of all the feature importance and direction of how the features are impacting the likelihood of diabetes risk for all the test population combined. This is done by looking at how the SHAP values are distributed and their magnitude across all instances together and using two complementary methods of visualising this information.

SHAP summary (beeswarm) plot shows both how important each feature was along with how it was impacting the model output (in a positive vs negative direction) for each instance.

SHAP bar plot orders the features according to the mean absolute value of their SHAP values represents the top predictors over the entire dataset.

SHAP analysis has also been conducted for the ensemble models. For soft voting ensembles, we obtain global feature importance by averaging all absolute SHAP values associated with a particular feature across the different base learners to obtain a single ranking that reflects the overall decision-making behaviour of the ensemble model. In the case of stacking ensemble models, we are conducting a SHAP analysis on the meta-learner, where we treat the predicted probabilities from the base models as input features.

4. Results

4.1. Numerical performance evaluation

This section reports the experimental results obtained from evaluating the individual tree-based models and the ensemble approaches. Table 3 presents the overall performance of all evaluated models on the test set using Accuracy, Precision, Recall, F1-score, and ROC–AUC.

The results of Table 3 show that all individual tree based models consistent performance, indicating the effectiveness of tree based learning when applied to structured health data. CatBoost achieved the best accuracy and F1 score, and the highest ROC-AUC value, reflecting its superior discriminative ability. Gradient Boosting was slightly lower in performance than CatBoost.

Both ensemble models outperformed all individual classifiers and validated the value of model aggregation. The voting ensemble gave a significant boost to the performance of the best individual classifier and showed that averaging probabilistic predictions decreases the variance of predictions. The stacking ensemble outperformed all other models on average for each metric, and all metrics demonstrated incremental and consistent improvement in Accuracy, F1-score, and ROC-AUC of stacked models.

4.2. Explainability analysis using SHAP

In this subsection, SHAP is employed to analyze the explainability of the individual tree-based models. Both global and local interpretations are provided to highlight feature influence patterns and instance-level decision behavior. The global feature contribution patterns are illustrated in Figs. 2 and 3, while local instance-level explanations are presented in Fig. 4. These visualizations collectively demonstrate how the individual model captures feature relevance at both global and local levels.

Fig. 5 presents the explainability analysis of the stacking ensemble model. SHAP is used to examine both the overall feature influence and the instance-level behavior of the ensemble decision mechanism. The global importance ranking and feature contribution patterns of the stacking ensemble are illustrated in Figs. 5 and 6, whereas representative local explanations are shown in Fig. 7. These results provide insight into the internal behavior of the stacking model and highlight differences compared with individual classifiers.

For the voting ensemble, explainability is derived using an aggregated SHAP approach that summarizes feature importance across the participating base models. Fig. 8 presents the global importance ranking of features based on the mean absolute SHAP values aggregated across all base learners.

5. Discussion

The results obtained numerically indicate that all the individual models and the ensemble models

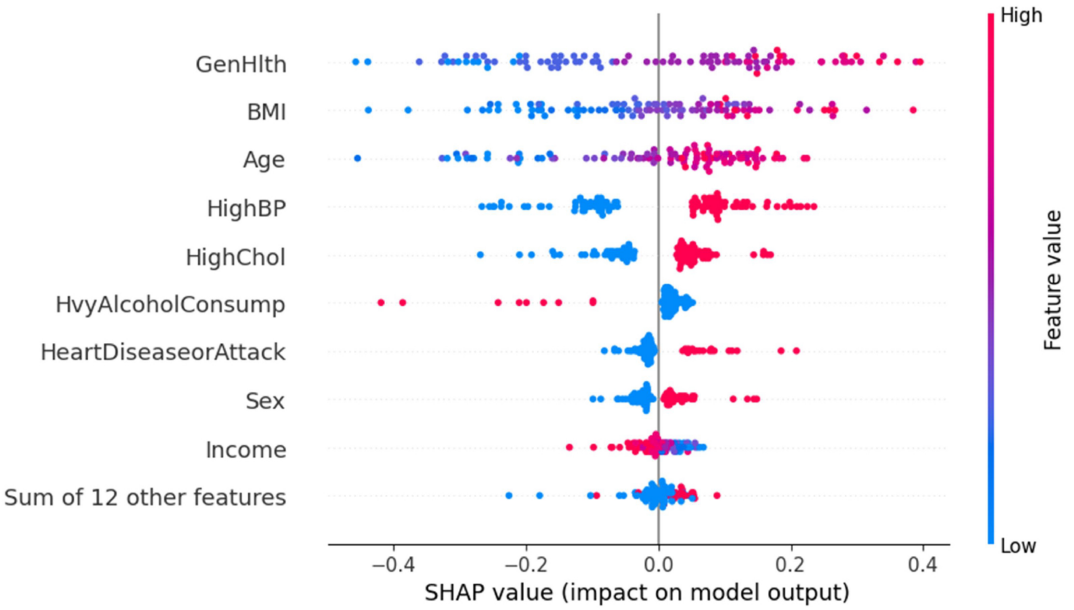

Fig. 2. SHAP Beeswarm plot for the best-performing individual model, illustrating the global distribution and direction of feature contributions.

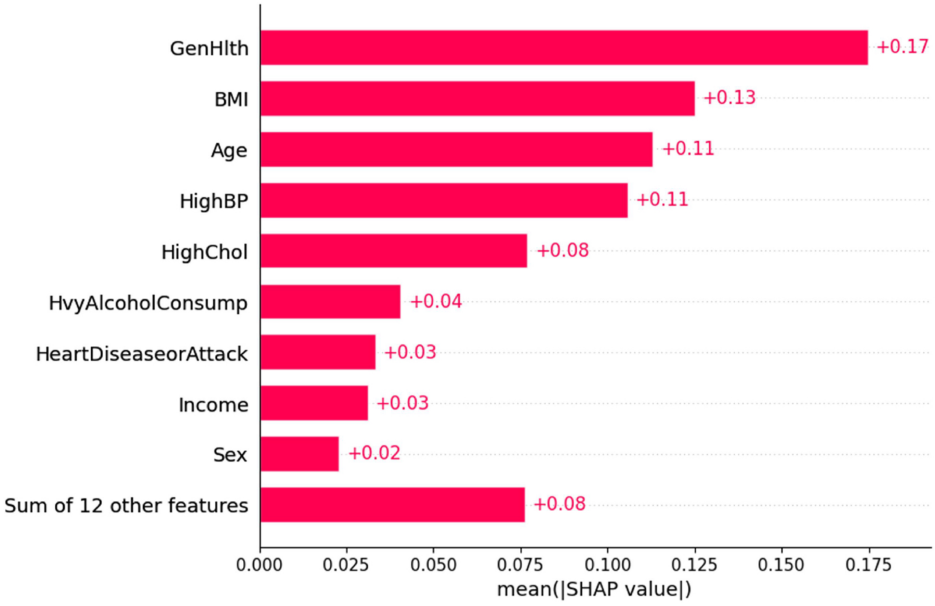

Fig. 3. SHAP Bar plot for the same model, showing the mean absolute SHAP values and ranking the most influential features.

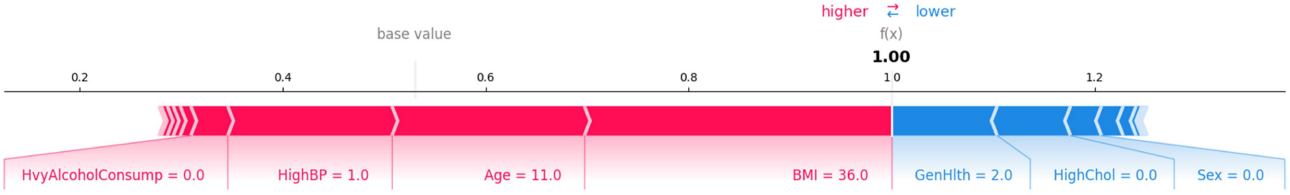

Fig. 4. SHAP Force plots for selected test instances, providing local explanations of how individual features contribute to the final prediction.

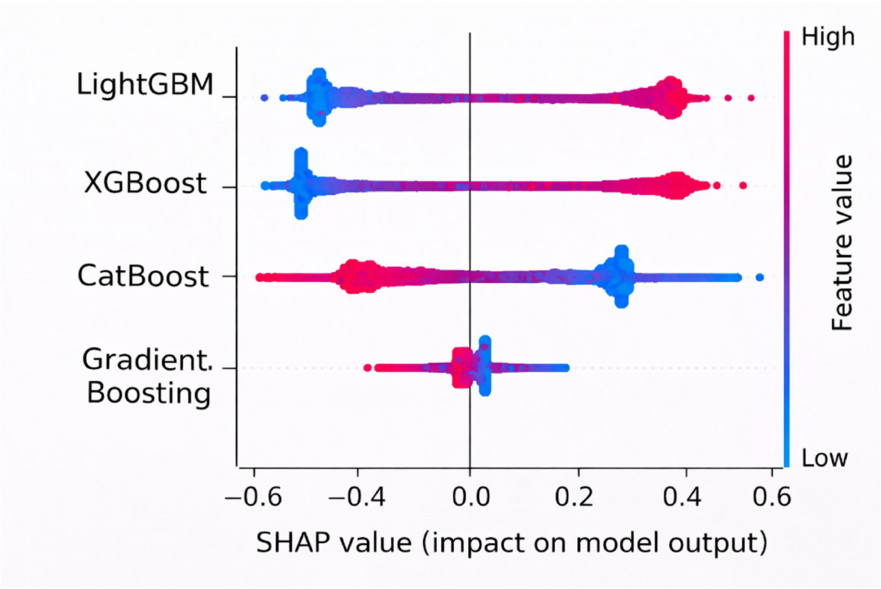

Fig. 5. SHAP Beeswarm plot corresponding to the stacking ensemble, reflecting the overall contribution patterns of the input features.

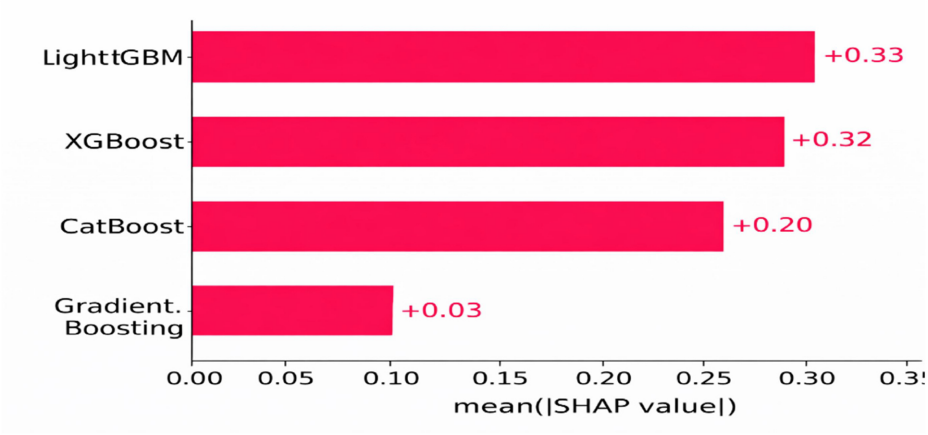

Fig. 6. SHAP Bar plot for the stacking ensemble, ranking features according to their average impact on the model output.

produced accurate predictions for diabetes. The model that performed the best on its own was CatBoost, achieving an accuracy of 0.871 and an F1-score of 0.871, along with a ROC–AUC of 0.921. The lack of large differences between the performances of the individual models can be attributed to the balanced nature of the dataset and the abundance of strong, predictive features. These consistent results indicate stable learning characteristics, as opposed to biases specific to individual models.

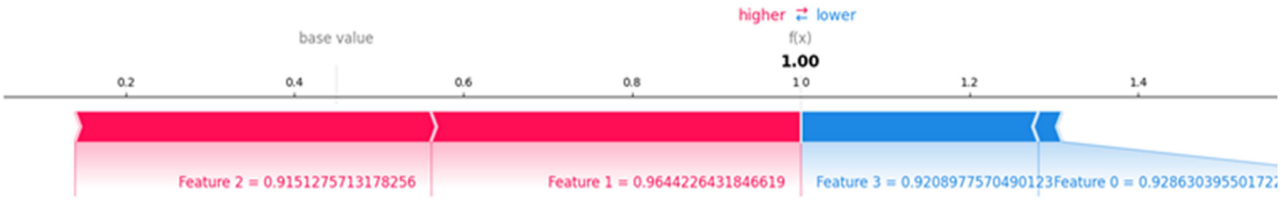

Fig. 7. SHAP Force plots for representative test samples, illustrating how the stacking ensemble integrates information to reach final predictions.

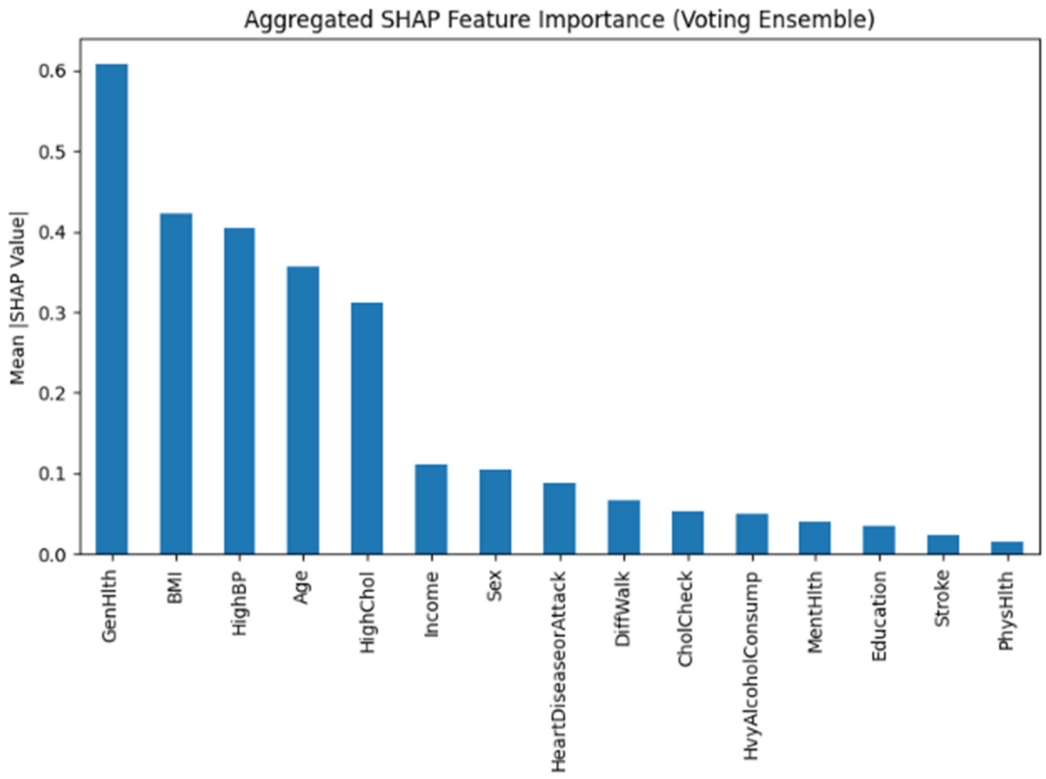

Fig. 8. Aggregated SHAP Bar plot for the voting ensemble.

All of the ensemble models showed an overall improvement from their individual counterparts. The accuracy of the voting ensemble increased by 0.871 to 0.878, while its ROC–AUC improved from 0.921 to 0.928. This relatively small improvement indicates that the use of probability-based aggregation decreases the variability in the predictions of the individual classifiers without adding complexity to the models.

The stacking ensemble exhibited the highest level of accuracy, with an accuracy of 0.883 and an F1-score of 0.883, as well as an ROC–AUC of 0.934. These increases in accuracy and ROC–AUC correspond to an improvement over the best individual model of 1.2% and 1.3%, respectively. Such increases in performance are consistent with the principles outlined in ensemble learning theory, whereby it takes advantage of the differences in error distribution of the individual learners, while also limiting the effects of overfitting.

From the perspective of interpretability, the SHAP-based explanations validate the numerical findings of this study. All models utilize strong clinically relevant predictors, including general health, body mass index, and age, and their predictive performance aligns with feature importance, supporting the validity of the methodology employed in this study.

While social determinants such as education level and household income—have measurable impacts on the SHAP values, they are small relative to the effect size of the dominant physiological factors (i.e., general health status, BMI, and age). As such, their impact on the decision-making process of each model is significantly smaller than the impact of the physiological predictors.

The comparison of local SHAP values across individual instances shows that the ensemble (stacking) models produce similar predictions for borderline cases. There are improvements in the F1-score, indicating an acceptable balance between precision and recall, for this model as well. An aggregated analysis of SHAP values for the voting ensemble indicates that while its structure preserves the feature importance ranking of the models that comprise it, the level of variation in SHAP contributions is less than that seen across the constituent models.

General Health is a major contributor to the development of Type 2 Diabetes Mellitus. Recent evidence from clinical studies and epidemiologic research demonstrate that patients' self-rated health is a multidimensional risk factor for Type 2 Diabetes; including aspects of social, behavioral and biological health factors. Jansana et al. [25] demonstrated that patients rating their health as poor are at an increased

risk for the development of Type 2 Diabetes when demographic and other lifestyle-related factors are accounted for as well. Similarly, Brückner et al. [26] demonstrated that diabetes patients typically rate themselves lower for general health status and quality of life; thus, it appears that self-rated health is a measure of the cumulative impact of metabolic and functional impairments. The consistency of these findings lends credence to the clinical likelihood that SHAP-based explanations can be used to understand the burden of these types of health conditions on the general population. The potential benefit of using SHAP-based explanations lies in the fact that clinicians and public health stakeholders can utilize Explainable Ensemble Models to improve risk stratification, identify the highest clinical and social risks, and give precedence to those with the highest risks for early screening and prevention. Incorporating SHAP-based explanations also provides the opportunity for clinicians and public health practitioners to make evidence-based decisions regarding patient care, as they are able to review the output of the Explainable Ensemble Models in conjunction with established clinical practices. However, several limitations must be acknowledged before real-world deployment, including the reliance on a single dataset, the absence of external validation, and the use of self-reported health indicators that may introduce reporting bias. Future work should therefore focus on validating the model on independent cohorts, assessing robustness across different populations, and evaluating clinical utility through prospective or real-world studies prior to operational integration into decision support systems.

6. Conclusion

A comparison of various Tree-Based Machine Learning Models & Ensemble Learning Techniques to Predict/Diagnose Diabetes was done with the 2015 Balanced dataset of the BRFSS Survey. Individual Tree-Based ML Models yielded varying levels of performance, with the CatBoost ML Model yielding the highest predictive ability. Additionally, combining the base ML models into an ensemble format further improves prediction performance, with the highest performing ensemble model for prediction being the Stacking Ensemble, followed by the Vote Ensemble. The effectiveness of aggregation through ensembling results supports as much as possible accuracy in predicting/diagnosing Diabetes. The SHAP-Based Explanations for the ML Models revealed similar trends with the General Health Status, BMI & Age being the most influential factors in predicting/diagnosing diabetes. In conclusion, using this approach to enhance the predictive robustness while maintaining the ability to explain interpretations of results has application in Healthcare Decision Support Systems, and should prove practical to utilize in future research.

Conflict of interest

The authors affirm that there is no conflict of interest regarding this work

Ethics information

This research complies with the ethical information for conducting scientific studies.

Funding

This research did not receive any funding.

References

1. Y. Zheng, S. Ley, and F. B. Hu, “Global aetiology and epidemiology of type 2 diabetes mellitus and its complications,” *Nature Reviews Endocrinology*, vol. 14, no. 2, pp. 88–98, 2018.
2. International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium, 2021. [Online]. Available: https://diabetesatlas.org
3. J. Zheng, Y. Zhang, and J. Li, “A comprehensive review of machine learning-based diabetes prediction,” *Artificial Intelligence in Medicine*, vol. 128, p. 102315, 2022.
4. S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance analysis of data mining classification techniques to predict diabetes,” *Procedia Computer Science*, vol. 167, pp. 191–200, 2020.
5. M. Hasan, M. Z. Hasan, and A. Rahman, “Diabetes prediction using ensemble machine learning models,” *IEEE Access*, vol. 9, pp. 76590–76604, 2021.
6. H. Alshammari, M. Alshammari, and S. Alqahtani, “Machine learning techniques for diabetes prediction: A comparative study,” *Sensors*, vol. 21, no. 15, p. 5072, 2021.
7. A. Ashraf *et al.*, “Explainable AI for diabetes prediction using ensemble learning,” *Computers in Biology and Medicine*, vol. 145, p. 105495, 2022.
8. S. Kabir, *TreeSHAP and Explainable AI for Tree-Ensemble Models: A Comprehensive Review*, independent research, 2025.
9. F. Xu, Z.-J. Zhou, J. Ni, and W. Gao, “Interpretation with baseline Shapley value for feature groups on tree models,” *Frontiers of Computer Science*, vol. 19, no. 5, p. 195316, 2025.
10. M. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.
11. R. Bodria *et al.*, “Benchmarking and survey of explanation methods for black box models,” *Data Mining and Knowledge Discovery*, vol. 35, pp. 1–49, 2021.
12. A. K. Shukla *et al.*, “Explainable artificial intelligence (XAI) in healthcare: A systematic review,” *Applied Soft Computing*, vol. 107, p. 107558, 2021.
13. Y. Zhang, X. Wang, and H. Chen, “Interpretable ensemble learning for medical risk prediction,” *Knowledge-Based Systems*, vol. 235, p. 107675, 2022.

14. S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
15. J. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
16. P. Netayawijit, W. Chansanam, and K. Sorn-In, “Interpretable machine learning framework for diabetes prediction: Integrating SMOTE balancing with SHAP explainability for clinical decision support,” *Healthcare*, vol. 13, no. 20, p. 2588, 2025.
17. K. Pang, “A comparative study of explainable machine learning models for type 2 diabetes prediction,” *Informatics in Medicine Unlocked*, vol. 44, 2025.
18. F. Rahman, “Diabetes prediction using feature selection algorithms with boosting classifiers and SHAP,” *Diagnostics*, vol. 15, no. 20, p. 2622, 2025.
19. M. S. Islam, “Explainable machine learning for efficient diabetes classification,” *Engineering Reports*, 2025.
20. L. Rafie *et al.*, “Advanced predictive modeling of type 2 diabetes using XGBoost and explainable AI,” *Research Square*, Dec. 2024.
21. P. B. Khokhar, V. Pentangelo, F. Palomba, and C. Gravino, “Towards transparent and accurate diabetes prediction using machine learning and explainable artificial intelligence,” *arXiv*, Jan. 2025.
22. U. Allani, “Interactive diabetes risk prediction using explainable machine learning: A Dash-based approach with SHAP, LIME, and comorbidity insights,” *arXiv*, May 2025.
23. İ. Kırbaş and A. Çifci, “Leveraging SHAP for interpretable diabetes prediction: A study of machine learning models on the Pima Indians diabetes dataset,” *Balkan Journal of Electrical and Computer Engineering*, vol. 13, no. 2, 2025.
24. Q. Sun, X. Cheng, H. Ren *et al.*, “Machine learning-based assessment of diabetes risk,” *Applied Intelligence*, vol. 55, p. 106, 2025.
25. M. Jansana *et al.*, “Validated diabetes risk scores and their associations with poor self-rated health,” *International Journal of Environmental Research and Public Health*, vol. 6, no. 10, 2025.
26. R. M. Brückner *et al.*, “Exploring factors associated with self-rated health in individuals with diabetes,” *Journal of Diabetes*, 2024.