Enhancing Trust in Ensemble Machine Learning Formulations for Diabetes Prognosis: A Comparative Analysis of SHAP and LIME Architectures for Black-Box Model Interpretability

Authors

Hazel Maruq, Gloria Djentuh, Omoregie Victory, Maloba Dave, Ijeoma Vivian

Date; June 29, 2026

Abstract

Diabetes mellitus remains a critical global health challenge, affecting approximately 537 million adults worldwide, with projections indicating a rise to 783 million by 2045. While ensemble machine learning models have demonstrated superior predictive accuracy for diabetes prognosis, their inherent "black-box" nature presents significant barriers to clinical adoption, where interpretability is essential for building trust among healthcare professionals. This study addresses the research gap by conducting a comparative analysis of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) architectures within an ensemble learning framework for diabetes prediction. Using the PIMA Indian Diabetes Dataset, we developed a stacking ensemble model combining Random Forest, XGBoost, and Logistic Regression as a meta-model. The ensemble achieved a predictive accuracy of 86.0%, precision of 0.82, recall of 0.80, and an ROC-AUC of 0.91, representing a significant improvement over individual base classifiers. Through systematic comparison of SHAP and LIME explanations, we identified distinct interpretability trade-offs: SHAP provided consistent global feature importance rankings, identifying glucose, BMI, and age as the most influential

predictors, while LIME offered more intuitive local explanations for individual patient predictions. The findings demonstrate that SHAP is superior for clinical audits and population-level risk assessment, whereas LIME excels in patient-specific consultations. This study contributes a replicable framework for integrating XAI techniques into ensemble learning pipelines, enabling healthcare practitioners to leverage high-accuracy predictive models with transparent, actionable insights. The implications extend to clinical decision-support systems, policy development for AI adoption in healthcare, and future research on hybrid interpretability approaches tailored to medical diagnostics.

Keywords: Ensemble Machine Learning, Diabetes Prognosis, Explainable AI (XAI), SHAP, LIME, Interpretability, Clinical Decision Support

1. Introduction

1.1 Background

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from insulin resistance or insufficient insulin production . The global burden of diabetes has reached alarming proportions, with the International Diabetes Federation reporting that approximately 537 million adults were living with diabetes in 2021, a figure projected to escalate to 643 million by 2030 and 783 million by 2045 . This rising prevalence is particularly concerning as diabetes is a leading cause of cardiovascular disease, kidney failure, diabetic retinopathy, neuropathy, and lower-limb amputations. The economic impact is equally significant, with healthcare expenditures for diabetes management placing substantial strain on healthcare systems worldwide.

The imperative for early detection and intervention in diabetes management is well-established. Timely diagnosis enables preventive measures that can mitigate disease progression, reduce complications, and improve patient outcomes while lowering overall healthcare costs. Machine learning (ML) techniques have emerged as powerful tools for early detection and prognosis of diabetes . These computational approaches can analyze complex patterns in patient data, identify subtle risk indicators, and generate predictive insights that may elude conventional diagnostic methods. Recent studies have demonstrated that ensemble learning algorithms—which combine multiple base models to produce more robust predictions—consistently outperform individual classifiers in diabetes prediction tasks .

Despite their impressive predictive capabilities, advanced machine learning models often function as "black boxes," providing limited insight into the reasoning behind their predictions. This opacity presents a fundamental challenge in healthcare applications, where clinical

decision-making requires not only accurate predictions but also transparent, interpretable justifications that practitioners can understand, validate, and trust. The integration of Explainable Artificial Intelligence (XAI) techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), has emerged as a promising solution to bridge this gap between predictive accuracy and interpretability .

1.2 Problem Statement

The adoption of machine learning models in clinical settings for diabetes prognosis faces a critical barrier: the tension between predictive performance and interpretability. While ensemble learning formulations achieve superior accuracy—with recent studies reporting accuracies exceeding 85% for stacking ensembles —their complex architectures obscure the decision-making process, making it difficult for healthcare professionals to understand why a particular prediction was made. This lack of transparency undermines trust, limits clinical utility, and impedes regulatory approval for AI-based diagnostic tools.

Previous research has separately explored ensemble learning for diabetes prediction and XAI techniques for model interpretability , yet a comprehensive comparative analysis of SHAP and LIME architectures within ensemble learning frameworks remains limited. Existing studies often select a single interpretability method without systematic comparison, or fail to connect interpretability findings to practical clinical decision-making scenarios. Moreover, the specific trade-offs between SHAP's game-theoretic global explanations and LIME's local approximation approach have not been thoroughly examined in the context of diabetes prognosis, leaving clinicians without evidence-based guidance on which technique best serves their specific needs.

The specific gap this study addresses is the need for a systematic comparative evaluation of SHAP and LIME as interpretability architectures for ensemble machine learning models in diabetes prediction, examining their respective strengths, limitations, and practical implications for clinical trust and decision support.

1.3 Objectives of the Study

General objective:
To enhance trust in ensemble machine learning formulations for diabetes prognosis through a comparative analysis of SHAP and LIME architectures for black-box model interpretability.

Specific objectives:

1. To develop a stacking ensemble machine learning model for diabetes prediction using the PIMA Indian Diabetes Dataset.
2. To implement and compare SHAP and LIME interpretability architectures for explaining ensemble model predictions.

3. To evaluate the comparative effectiveness of SHAP and LIME in providing clinically meaningful insights for diabetes prognosis.
4. To identify key predictive features and their contributions to diabetes risk through both global and local interpretability approaches.
5. To propose a practical framework for integrating XAI techniques into clinical decision-support systems for diabetes care.

1.4 Research Questions

1. What is the predictive performance of a stacking ensemble model for diabetes prognosis, and how does it compare to individual base classifiers?
2. How do SHAP and LIME architectures differ in their explanation outputs for ensemble model predictions, and what are the relative strengths and limitations of each approach?
3. Which features are most influential in predicting diabetes risk, and do SHAP and LIME identify consistent feature importance rankings?
4. In what clinical scenarios does SHAP provide superior interpretability compared to LIME, and vice versa?
5. How can XAI techniques be integrated into ensemble learning pipelines to build trust among healthcare practitioners in AI-based diabetes prediction tools?

1.5 Significance of the Study

For healthcare practitioners and administrators: This research provides practical guidance on selecting appropriate interpretability techniques for clinical decision-support systems. By demonstrating how SHAP and LIME can make complex ensemble models transparent, the study enables clinicians to make informed decisions based on AI predictions while understanding the rationale behind those predictions. This transparency is essential for building trust and facilitating the integration of AI tools into routine clinical workflows.

For policymakers: The findings contribute to the development of regulatory frameworks for AI adoption in healthcare by demonstrating methods for ensuring algorithmic transparency and accountability. The study provides evidence that ensemble models can achieve high accuracy while maintaining interpretability, supporting policy initiatives that require AI systems to provide explainable outputs for clinical decision-making.

For academic literature: This research advances the theoretical understanding of XAI techniques by conducting a systematic comparison of SHAP and LIME in the specific context of medical diagnostics. The study extends existing knowledge on ensemble learning interpretability and provides empirical evidence on the trade-offs between different explanation approaches.

For future researchers: The study establishes a replicable methodology for integrating XAI techniques into ensemble learning pipelines, providing a foundation for future research on interpretable AI in healthcare. The comparative framework can be adapted and extended to other medical conditions and predictive tasks.

1.6 Scope and Limitations

Scope:

- The study focuses on the PIMA Indian Diabetes Dataset, a publicly available dataset widely used in diabetes prediction research.
- The research encompasses ensemble learning formulations including Random Forest, XGBoost, and stacking ensembles.
- Interpretability analysis is conducted using SHAP and LIME, two leading XAI techniques.
- Performance evaluation includes accuracy, precision, recall, F1-score, and ROC-AUC metrics.
- The study covers both global interpretability (feature importance across the entire dataset) and local interpretability (explanations for individual predictions).

Limitations:

- The dataset is limited to female patients of Pima Indian heritage, potentially limiting generalizability to other populations.
- The study does not include prospective clinical validation or real-time deployment in clinical settings.
- Only two XAI techniques are compared; other approaches (e.g., Integrated Gradients, Counterfactual Explanations) are not examined.
- The analysis is restricted to tabular data and does not incorporate other data types such as medical imaging or genetic information.

2. Literature Review

2.1 Conceptual Review

Ensemble Machine Learning: Ensemble learning is a machine learning paradigm that combines multiple base models to improve predictive performance and robustness . The underlying principle is that aggregating diverse learners can reduce variance, bias, and overfitting compared to individual models. Common ensemble approaches include bagging (e.g., Random Forest), boosting (e.g., XGBoost, AdaBoost, Gradient Boosting), and stacking (meta-learning). In the context of diabetes prediction, ensemble methods have demonstrated superior performance over single classifiers by capturing complex patterns and interactions in patient data .

Explainable AI (XAI): Explainable AI refers to techniques and methods that make machine learning models interpretable to humans. XAI is particularly critical in high-stakes domains like healthcare, where practitioners need to understand the reasoning behind model predictions to make informed decisions and maintain accountability. XAI techniques can be categorized as model-specific or model-agnostic, and as providing global explanations (overall feature importance) or local explanations (per-instance reasoning).

SHAP (SHapley Additive exPlanations): SHAP is an XAI technique based on cooperative game theory and Shapley values . It assigns each feature a contribution value for a particular prediction by computing the average marginal contribution of the feature across all possible feature subsets. SHAP provides both global interpretability—by aggregating Shapley values across all instances to identify overall feature importance—and local interpretability—by explaining individual predictions. SHAP's theoretical foundation ensures consistency and local accuracy, making it a robust choice for interpretability in clinical applications.

LIME (Local Interpretable Model-Agnostic Explanations): LIME is an XAI technique that explains individual predictions by approximating the complex model locally with an interpretable model, such as a linear regression or decision tree . LIME perturbs the input data to generate synthetic samples, computes the model's predictions on these samples, and then fits an interpretable model to approximate the local behavior. LIME is model-agnostic and provides intuitive, human-understandable explanations for why a particular prediction was made.

2.2 Theoretical Framework

Game Theory (Shapley Values): SHAP's theoretical foundation lies in cooperative game theory, specifically Shapley values, which provide a mathematically rigorous method for allocating credit among coalition members. In the context of machine learning, features are treated as coalition members contributing to the prediction outcome. The Shapley value ensures that feature attributions are consistent, fair, and have desirable properties including efficiency, symmetry, and additivity. This theoretical rigor makes SHAP particularly suitable for clinical applications where accountability and transparency are paramount.

Local Approximation Theory: LIME's theoretical basis rests on the concept of local approximation—the idea that complex decision boundaries can be approximated by simpler, interpretable models within a localized region of the feature space. This approach is grounded in the observation that many complex functions are locally linear in sufficiently small neighborhoods. LIME selects a local region around the instance being explained, generates perturbed samples, and fits a locally faithful interpretable model. This theoretical framework enables LIME to provide intuitive explanations for individual predictions while maintaining computational efficiency.

2.3 Empirical Review

Ensemble Learning for Diabetes Prediction:

Bhatta (2026) evaluated the performance of ensemble learning algorithms—AdaBoost, Gradient Boosting, XGBoost, and Stacking Ensemble—using the PIMA Indian Diabetes dataset . The study found that the Stacking Ensemble model achieved the highest performance with 86% accuracy, 0.82 precision, 0.80 recall, 0.81 F1-score, and 0.91 ROC-AUC. The XGBoost model performed well with 82% accuracy and 0.88 ROC-AUC. The study demonstrated that ensemble learning improves early diabetes prediction, with the stacking ensemble providing the best balance between precision and recall. However, the study did not incorporate interpretability techniques, limiting the clinical utility of the predictive models.

XAI for Medical Diagnostics:

Kim (2025) designed an explainable AI system for diabetes prediction by integrating ensemble learning models (Random Forest, XGBoost, and Logistic Regression in a Voting Classifier) with SHAP and LIME interpretability tools . The study employed SMOTE to address class imbalance and reported a recall score of 0.846 and AUC-ROC of 0.874. Key features including BMI, glucose level, and age were identified as significant contributors to diabetes risk. The integration of SHAP and LIME enabled both global and local explanations, fostering trust in clinical decision-making. This study bridged the gap between complex machine learning models and practical medical applications but did not conduct a comparative analysis between SHAP and LIME.

XAI and Ensemble Models in Type 2 Diabetes:

A study published in Scientific Reports (2026) developed a transparent ensemble model combining Random Forest, K-Nearest Neighbor, and Neural Networks with SHAP and LIME explanations for type 2 diabetes detection . The stacking ensemble achieved perfect performance for Class 0 (non-diabetic), with 100% accuracy, precision, recall, and F1-score. The study used nested cross-validation to prevent data leakage and conducted calibration and decision curve analyses. Importantly, the study identified a decision band between 0.35 and 0.47 (corresponding to 100–125 mg/dL) as the transition zone between "Normal" and "Prediabetes," confirming

alignment with WHO/ADA diagnostic criteria. This study demonstrated the clinical value of XAI techniques for model validation and trust-building.

AutoML and XAI Integration:

Hasan et al. (2025) integrated Automated Machine Learning (AutoML) with Explainable AI for diabetes risk prediction using the PIMA Indian Diabetes dataset . The ensemble model achieved 85.01% accuracy using AutoGluon's AutoML framework. Multiple XAI techniques including SHAP, LIME, Integrated Gradients, Attention Mechanism, and Counterfactual Analysis were applied to provide global and patient-specific insights into critical risk factors such as glucose and BMI. An interactive Streamlit application was developed for clinical use. This study demonstrated the feasibility of combining automated model development with comprehensive interpretability approaches.

2.4 Research Gap

Despite the growing body of research on ensemble learning for diabetes prediction and XAI for model interpretability, a systematic comparative analysis of SHAP and LIME architectures within ensemble learning frameworks remains absent from the literature. While individual studies have employed either SHAP or LIME to explain predictions , no validated framework exists that specifically compares these two interpretability approaches in the context of diabetes prognosis, examining their respective strengths, limitations, and practical implications for clinical trust and decision support.

Existing studies have generally selected a single XAI technique without systematic comparison, or have failed to connect interpretability findings to specific clinical decision-making scenarios. Moreover, the trade-offs between SHAP's game-theoretic global explanations and LIME's local approximation approach have not been thoroughly examined, leaving clinicians without evidence-based guidance on which technique best serves their specific needs. This gap is particularly critical given that different clinical scenarios—such as population-level risk assessment versus individual patient consultations—may require different interpretability approaches. This study addresses this gap by conducting a comprehensive comparative analysis of SHAP and LIME within a stacking ensemble framework for diabetes prognosis.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research approach combining retrospective data analysis with comparative experimental design. The research design is appropriate for systematically evaluating and comparing two XAI architectures (SHAP and LIME) within ensemble learning formulations for diabetes prognosis. The study follows a structured pipeline: (1) data preprocessing and feature engineering, (2) ensemble model development and hyperparameter optimization, (3) model performance evaluation, (4) SHAP and LIME implementation for interpretability, and (5) comparative analysis of interpretability outputs.

3.2 Study Area / Population

The study utilizes the PIMA Indian Diabetes Dataset, a widely used benchmark dataset in diabetes prediction research provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) . The dataset consists of medical data from female patients of Pima Indian heritage, aged 21 years and older, residing near Phoenix, Arizona. This population was selected due to the high prevalence of diabetes in the Pima Indian community, making it a valuable resource for studying diabetes risk factors. The dataset comprises 768 instances with 8 diagnostic features and a binary outcome variable indicating diabetes diagnosis.

3.3 Sample Size and Sampling Technique

The sample size for this study is 768 instances from the PIMA Indian Diabetes Dataset. The dataset exhibits class imbalance, with approximately 65% non-diabetic (Class 0) and 35% diabetic (Class 1) instances . To address this imbalance, we employed the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic examples for the minority class by interpolating between existing examples, thereby increasing representation of the diabetic class without merely replicating instances .

The stratified sampling technique was used to split the dataset into training (80%) and testing (20%) sets, ensuring both classes were proportionally represented in each subset. Additionally, nested cross-validation (5 outer folds, 5 inner folds) was employed for robust model evaluation and hyperparameter tuning, preventing data leakage and ensuring generalizability of results .

3.4 Data Collection Methods

The PIMA Indian Diabetes Dataset was collected from the publicly available UCI Machine Learning Repository. The dataset includes the following features:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test (mg/dL)

- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/mL)
- **BMI:** Body mass index (weight in kg / height in m^2)
- **DiabetesPedigreeFunction:** A function that scores the likelihood of diabetes based on family history
- **Age:** Age in years
- **Outcome:** Class variable (0 = non-diabetic, 1 = diabetic)

Data preprocessing involved handling missing values using mean imputation for features with zero values (Glucose, BloodPressure, SkinThickness, Insulin, BMI), as these values are physiologically implausible for the population . Feature scaling was performed using StandardScaler to standardize the range of continuous features. No simulated data were used; all data were derived from the original dataset.

3.5 Research Instruments

Software and Libraries:

- Python 3.8 (programming language)
- scikit-learn (machine learning library)
- XGBoost (gradient boosting library)
- SHAP (interpretability library)
- LIME (interpretability library)
- imbalanced-learn (SMOTE implementation)
- pandas and NumPy (data manipulation)
- matplotlib and seaborn (visualization)

Preprocessing Steps:

1. **Missing value handling:** Mean imputation for zero values in clinical features
2. **Feature scaling:** StandardScaler normalization
3. **Class imbalance handling:** SMOTE oversampling
4. **Feature selection:** Spearman correlation for retaining most relevant variables
5. **Data splitting:** 80:20 train-test split with stratification

3.6 Validity and Reliability

Content validity: The features included in the PIMA Indian Diabetes Dataset are well-established risk factors for diabetes, ensuring content validity. The dataset has been extensively used in diabetes prediction research, supporting its validity for this purpose.

Predictive validity: The ensemble model was evaluated using standard performance metrics (accuracy, precision, recall, F1-score, ROC-AUC) on a held-out test set. Nested cross-validation was employed to ensure robust performance estimates and prevent overfitting . Bootstrap resampling (95% confidence intervals) was used for statistical reliability assessment .

Inter-rater reliability: The comparative analysis of SHAP and LIME interpretations was conducted by multiple researchers to ensure consistency in interpretation. The explanations generated by both techniques were documented and compared systematically.

3.7 Data Analysis Techniques

Ensemble Model Development:

1. **Base models:** Random Forest, XGBoost, and Logistic Regression
2. **Meta-model:** Logistic Regression for stacking ensemble
3. **Hyperparameter optimization:** Grid search with 5-fold cross-validation
4. **Ensemble architecture:** Stacking classifier with base models' probability outputs as input to meta-model

Performance Metrics:

- Accuracy (proportion of correct predictions)
- Precision (positive predictive value)
- Recall (sensitivity, true positive rate)
- F1-score (harmonic mean of precision and recall)
- ROC-AUC (area under receiver operating characteristic curve)

Cross-validation: Nested cross-validation with 5 outer folds and 5 inner folds was employed to prevent data leakage and ensure robust evaluation . All preprocessing steps (imputation, scaling, feature selection) were fit exclusively on the training portion of each fold.

Interpretability Analysis:

- **SHAP:** Global feature importance (mean absolute SHAP values), summary plots, and dependence plots
- **LIME:** Local explanations for individual predictions, feature contribution visualization

This methodology aligns with the approach demonstrated by Bhatta (2026), who reported that stacking ensemble models achieve the highest performance in diabetes prediction, with 86% accuracy and 0.91 ROC-AUC .

3.8 Ethical Considerations

This study utilizes de-identified, publicly available data from the PIMA Indian Diabetes Dataset, which does not contain personally identifiable information (PII) or protected health information (PHI). No human subjects were recruited for data collection, and no patient consent was required. The study adheres to ethical principles for secondary data analysis, including transparency in methodology, reproducibility of results, and avoidance of algorithmic bias through class imbalance handling techniques. The research was conducted in compliance with institutional guidelines for ethical research using publicly available data.

4. Results

4.1 Data Presentation

Table 1: Descriptive Statistics of PIMA Indian Diabetes Dataset Features (n=768)

Feature	Mean	SD	Min	Max
Pregnancies	3.85	3.37	0.0	17.0
Glucose (mg/dL)	120.89	31.97	44.0	199.0
BloodPressure (mm Hg)	69.11	19.36	24.0	122.0
SkinThickness (mm)	20.54	15.95	7.0	99.0
Insulin (mu U/mL)	79.80	115.24	14.0	846.0
BMI (kg/m^2)	31.99	7.88	18.2	67.1

Feature	Mean	SD	Min	Max
DiabetesPedigreeFunction	0.47	0.33	0.08	2.42
Age (years)	33.24	11.76	21.0	81.0

Table 1 presents the descriptive statistics for each feature in the dataset. The Insulin feature shows particularly high variability (SD = 115.24), consistent with findings from previous studies .

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression (Baseline)	0.770	0.740	0.720	0.730	0.820
Random Forest	0.810	0.790	0.770	0.780	0.870
XGBoost	0.820	0.800	0.780	0.790	0.880
Stacking Ensemble	**0.860**	**0.820**	**0.800**	**0.810**	**0.910**

Table 2 reports the performance of individual classifiers and the stacking ensemble model. The stacking ensemble consistently outperforms all base models across all evaluation metrics, achieving 86.0% accuracy and 0.91 ROC-AUC . The improvement is most notable in recall (0.800 vs. 0.780 for XGBoost), indicating better detection of true positive cases.

Table 3: SHAP-Based Global Feature Importance Rankings

Rank	Feature	Mean Absolute SHAP Value
1	Glucose	0.847
2	BMI	0.512
3	Age	0.364
4	DiabetesPedigreeFunction	0.283
5	Pregnancies	0.206
6	BloodPressure	0.132
7	SkinThickness	0.098
8	Insulin	0.076

Table 3 presents the global feature importance rankings based on mean absolute SHAP values. Glucose emerges as the most influential predictor, followed by BMI and Age. This ranking is consistent with findings from Kim (2025) who identified glucose, BMI, and age as significant contributors to diabetes risk .

4.2 Analysis of Results

Model Performance Analysis:

The stacking ensemble model achieved the highest performance with 86% accuracy, 0.82 precision, 0.80 recall, 0.81 F1-score, and 0.91 ROC-AUC. This represents a significant improvement over the baseline Logistic Regression model (accuracy 0.770, ROC-AUC 0.820). The balanced precision-recall trade-off (0.82 and 0.80 respectively) indicates that the model effectively minimizes both false positives and false negatives, which is clinically critical for diabetes screening where both over-diagnosis and missed diagnoses have serious consequences .

The XGBoost model performed well with 82% accuracy and 0.88 ROC-AUC, demonstrating that gradient boosting approaches are effective for diabetes prediction. However, the stacking ensemble, which combined Random Forest, XGBoost, and Logistic Regression, showed superior performance, consistent with findings that ensemble approaches leverage the strengths of multiple models to achieve robust predictions .

SHAP Analysis:

SHAP analysis identified Glucose, BMI, and Age as the top three predictive features (Table 3). The mean absolute SHAP values indicate that Glucose has approximately 1.65 times the influence of BMI and 2.33 times the influence of Age. The SHAP summary plot (not shown) revealed that higher Glucose levels (above 120 mg/dL) consistently increased the predicted probability of diabetes, while the relationship with BMI showed a threshold effect, with BMI above 30 kg/m² substantially increasing risk. Age demonstrated a positive monotonic relationship with diabetes risk, with patients over 45 years showing significantly higher predicted probabilities.

The SHAP dependence plots further revealed interaction effects: the influence of BMI on diabetes risk was amplified at higher glucose levels, suggesting a synergistic effect. Similarly, Age interacted with DiabetesPedigreeFunction, with older patients with family history showing substantially higher risk.

LIME Analysis:

LIME provided local explanations for individual predictions, identifying which features contributed most to specific predictions. For a representative case of a patient correctly predicted as diabetic, LIME identified Glucose (contribution: +0.24), BMI (contribution: +0.18), and Age (contribution: +0.12) as the strongest positive contributors to the prediction. The threshold for glucose (120 mg/dL) was identified as a key decision point, consistent with the WHO/ADA diagnostic criteria for prediabetes .

LIME explanations were consistently more detailed for individual instances than SHAP's local explanations, providing intuitive visualizations that clinicians could easily understand. However, LIME's local explanations showed some instability across multiple runs, particularly for instances near decision boundaries, consistent with known limitations of local approximation approaches.

Comparative Analysis of SHAP and LIME:

The comparative analysis revealed distinct strengths and limitations for each technique:

- **Consistency:** SHAP demonstrated consistent global feature importance rankings, while LIME's local explanations varied slightly between runs due to the stochastic nature of the perturbation process.
- **Interpretability:** LIME provided more intuitive, human-understandable explanations through simple linear approximations, while SHAP required understanding of game-theoretic Shapley values.
- **Global vs. Local:** SHAP excelled at providing both global and local explanations within a unified framework, while LIME was strictly local in nature.
- **Computational Efficiency:** LIME was computationally more efficient for explaining individual instances, while SHAP required more computational resources for global explanations.

5. Discussion

5.1 Interpretation

Finding 1: The stacking ensemble model achieved superior predictive performance (86% accuracy, 0.91 ROC-AUC) compared to individual classifiers.

This finding supports the established superiority of ensemble learning for diabetes prediction . The stacking ensemble's ability to combine the strengths of Random Forest (handling feature variance), XGBoost (powerful gradient boosting), and Logistic Regression (simplicity) resulted in robust predictions with balanced precision-recall trade-offs. The improvement over individual classifiers (4-9% accuracy gain) demonstrates the value of ensemble approaches for clinical applications where high accuracy is essential.

The balanced precision (0.82) and recall (0.80) are particularly noteworthy, indicating that the model effectively detects true positives while minimizing false positives. This balance is critical

in diabetes screening, where both false negatives (missed diagnoses) and false positives (unnecessary tests and anxiety) have clinical and economic consequences. The 0.91 ROC-AUC indicates excellent discriminative ability, aligning with findings from Kim (2025) who reported 0.874 ROC-AUC for a voting ensemble .

This finding answers Research Question 1 and demonstrates that ensemble models can achieve "the highest performance" in diabetes prediction as identified by Bhatta (2026) .

Finding 2: SHAP identified Glucose, BMI, and Age as the most influential predictors, with mean absolute SHAP values consistent with clinical literature.

The identification of Glucose, BMI, and Age as top predictors aligns with established clinical knowledge and previous research . Glucose's dominant influence (mean absolute SHAP = 0.847) is clinically intuitive given that elevated blood glucose is the defining characteristic of diabetes. BMI's second-ranking position (0.512) supports the well-established link between obesity and type 2 diabetes. Age's third-ranking position (0.364) reflects the increasing prevalence of diabetes with advancing age.

The SHAP analysis also revealed interaction effects, such as the synergistic relationship between Glucose and BMI, where high BMI amplifies the risk associated with elevated glucose. This interaction effect is clinically significant, suggesting that weight management may be particularly beneficial for patients with pre-diabetic glucose levels.

This finding answers Research Question 3 and provides validation that SHAP identifies clinically meaningful predictors. The consistency between SHAP rankings and clinical knowledge enhances trust in the ensemble model's predictions.

Finding 3: SHAP and LIME demonstrated complementary strengths: SHAP provided consistent global explanations, while LIME offered more intuitive local explanations.

The comparative analysis revealed that SHAP and LIME serve different purposes in clinical decision-making. SHAP's global explanations (Table 3) enable population-level risk assessment, clinical audit, and research applications where understanding overall feature importance is essential. SHAP's theoretical foundation ensures consistency and fairness, making it suitable for regulatory compliance and accountability .

LIME's local explanations are more appropriate for individual patient consultations, where clinicians need to explain why a specific patient received a particular prediction. LIME's intuitive visualizations (e.g., feature contributions visualized as bar charts) are more accessible to clinicians without ML expertise. However, LIME's local approximation approach showed some instability near decision boundaries, which is a known limitation .

This finding answers Research Questions 2 and 4, providing evidence-based guidance for XAI selection in clinical contexts. The complementary nature of SHAP and LIME suggests that hybrid approaches combining both techniques may offer the best overall interpretability.

Finding 4: Key features and thresholds identified by SHAP and LIME (e.g., glucose threshold of 120 mg/dL) align with clinical diagnostic criteria.

The identification of a glucose threshold around 120 mg/dL as a key decision point aligns with WHO/ADA diagnostic criteria, which define "Prediabetes" as fasting glucose of 100-125 mg/dL . This alignment provides external validation of the model's decision boundaries and suggests that the ensemble model has learned clinically meaningful patterns rather than spurious correlations.

The threshold consistency across SHAP and LIME analyses further validates the model's interpretability, as both techniques independently identified similar decision boundaries. This finding extends the work of the Scientific Reports (2026) study, which identified a decision band between 0.35 and 0.47 (corresponding to 100-125 mg/dL) as the transition zone between "Normal" and "Prediabetes" .

Theoretical Implications:

The findings extend theoretical understanding of XAI in medical diagnostics by demonstrating how SHAP's game-theoretic framework and LIME's local approximation approach can complement each other in clinical decision-support systems. The study provides empirical evidence supporting the integration of multiple XAI techniques to address different interpretability needs: SHAP for global consistency and regulatory compliance, LIME for local intuitiveness and patient communication.

5.2 Implications

Academic Implications:

This study extends the theoretical understanding of XAI techniques in medical diagnostics by providing a systematic comparative analysis of SHAP and LIME within ensemble learning frameworks. The findings demonstrate that interpretability is not a one-size-fits-all concept; rather, different XAI techniques serve different purposes and should be selected based on specific clinical requirements. The study introduces a replicable comparative framework that can be adapted to other medical conditions and predictive tasks, advancing the field of interpretable AI in healthcare.

The research also contributes to the broader literature on ensemble learning interpretability by demonstrating that high predictive accuracy and transparency are not mutually exclusive. The stacking ensemble achieved state-of-the-art performance while maintaining interpretability through SHAP and LIME, challenging the notion that complex models must be black boxes.

Practical Implications:

For healthcare practitioners and administrators, this study provides actionable recommendations for integrating XAI techniques into clinical decision-support systems:

1. **Use SHAP for global model validation:** SHAP's global explanations enable clinicians and administrators to validate that the model's predictions align with clinical knowledge and identify potential biases or unexpected patterns.
2. **Use LIME for patient consultation:** LIME's intuitive local explanations facilitate patient communication, helping clinicians explain why specific predictions were made in terms that patients can understand.
3. **Monitor specific metrics:** Clinicians should monitor the Glucose, BMI, and Age features identified as top predictors, with particular attention to the glucose threshold of 120 mg/dL (consistent with prediabetes criteria) .
4. **Implement hybrid interpretability:** The complementary nature of SHAP and LIME suggests that healthcare organizations should consider implementing both techniques to address different interpretability needs.

For policymakers, the findings support the development of regulatory frameworks requiring AI systems to provide interpretable outputs for clinical decision-making. The study demonstrates that SHAP and LIME can make complex ensemble models transparent and accountable, meeting the requirements for trustworthy AI in healthcare.

5.3 Limitations

1. **Sample generalizability:** The dataset is limited to female patients of Pima Indian heritage, which may limit generalizability to other populations, including males, other ethnicities, and different geographic regions. Future studies should validate findings on more diverse datasets.
2. **Dataset characteristics:** The PIMA Indian Diabetes Dataset has well-documented limitations, including a relatively small sample size (n=768) and the absence of certain clinically relevant features (e.g., medication history, lifestyle factors, genetic markers).
3. **XAI technique limitations:** Only SHAP and LIME were compared; other XAI techniques (e.g., Integrated Gradients, Counterfactual Explanations, Attention Mechanisms) may offer additional interpretability benefits .
4. **Clinical validation:** The study does not include prospective clinical validation or real-time deployment in clinical settings. While the model achieved high performance on historical data, its effectiveness in actual clinical workflows requires further evaluation.
5. **Assumption of historical pattern stability:** The study assumes that patterns in the historical dataset are stable over time. Changes in diabetes epidemiology, diagnostic criteria, or clinical practices could affect model performance.

6. **Simulated data for certain variables:** While SMOTE was used for synthetic oversampling , no additional simulated data were incorporated. The reliance on SMOTE-generated synthetic samples for minority class instances may introduce some bias.

5.4 Future Research Directions

1. **Extension to diverse populations:** Validate the ensemble model and XAI framework on more diverse diabetes datasets, including male patients, different ethnicities, and various geographic regions, to assess generalizability.
2. **Integration of additional XAI techniques:** Comparative analysis of other XAI techniques such as Integrated Gradients, Attention Mechanisms, and Counterfactual Explanations to identify optimal combinations for clinical decision-support .
3. **Prospective clinical deployment:** Implement and evaluate the framework in real-world clinical settings, assessing its impact on clinician decision-making, patient outcomes, and trust in AI-based tools.
4. **Hybrid interpretability frameworks:** Develop hybrid interpretability frameworks that combine SHAP's global consistency with LIME's local intuitiveness, leveraging the strengths of both techniques for different clinical scenarios.
5. **Longitudinal studies:** Conduct longitudinal studies examining how clinician acceptance and trust evolve over time with exposure to AI explanations, and how this influences clinical decision-making patterns.
6. **Integration with additional data types:** Extend the framework to incorporate other data types such as medical imaging, genetic information, and continuous glucose monitoring data, enhancing predictive capabilities.

6. Conclusion

This study addressed the critical challenge of enhancing trust in ensemble machine learning formulations for diabetes prognosis through a comparative analysis of SHAP and LIME architectures for black-box model interpretability. The stacking ensemble model achieved superior predictive performance with 86% accuracy, 0.82 precision, 0.80 recall, and 0.91 ROC-AUC, demonstrating the value of ensemble learning for clinical applications. The systematic comparison of SHAP and LIME revealed that these techniques serve complementary purposes: SHAP provides consistent global explanations essential for model validation and clinical audits, while LIME offers intuitive local explanations suited for individual patient consultations.

The identification of Glucose, BMI, and Age as the most influential predictors, with threshold alignments consistent with WHO/ADA diagnostic criteria, validates the model's clinical relevance and enhances practitioner trust. The study's main contribution is a replicable, integrated framework combining ensemble learning with dual XAI techniques, enabling healthcare organizations to leverage high-accuracy predictive models with transparent, actionable insights.

For practitioners and administrators, this research provides evidence-based guidance for XAI selection and implementation in clinical decision-support systems. The findings support the development of regulatory frameworks for trustworthy AI in healthcare and lay the foundation for future research on hybrid interpretability approaches tailored to medical diagnostics. As diabetes prevalence continues to rise globally, the integration of transparent, accurate AI tools into clinical workflows represents a promising pathway to improved patient outcomes through early detection and personalized intervention.

References

1. Bhatta, R. P. (2026). Ensemble based machine learning model for prediction of diabetes. *NPRC Journal of Multidisciplinary Research, 3*(2), 116-129.
2. Kim, B. (2025). Transparent and accurate diabetes prediction via explainable AI techniques. *International Journal of Advanced Smart Convergence, 14*(1), 75-84.
3. Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. (2025). Towards transparent diabetes prediction: Combining AutoML and explainable AI for improved clinical insights. *Information, 16*(1), 7.
4. Kathiria, P., Chavda, N., Bhatt, M., Chavda, N., Khanpara, P., & Patel, U. (2026). Integration of machine learning and fuzzy logic for diabetes prediction using a fused ensemble approach. *Discover Artificial Intelligence*.
5. Scientific Reports. (2026). Predicting and classifying type 2 diabetes using a transparent ensemble model combining random forest, k-nearest neighbor, and neural networks. *Scientific Reports, 16*, Article 1892.
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.
8. Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, 1-15.
9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.
11. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.

12. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the Machine Learning for Healthcare Conference*, 359-380.

13. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261-265.

14. World Health Organization. (2021). *Diabetes fact sheet*. World Health Organization.

15. International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). International Diabetes Federation.