Enhancing Clinician Trust in Automated Diagnostic Systems: A Framework Combining Random Forest, XGBoost, and SHAP for Interpretable and Auditable Diabetes Risk Prediction

Authors

Rasa Kiki, Wiliams Gassimu, Carlson Geroge, Gideon Ifeanyi, Billy Elly

Date; June 26, 2026

Abstract

Diabetes mellitus affects over 537 million adults globally, with early detection critical for reducing long-term complications and healthcare costs. Despite advances in machine learning for disease prediction, the "black-box" nature of many high-performing models limits clinical adoption due to insufficient transparency and clinician trust. This study addresses this gap by developing a hybrid predictive framework that integrates Random Forest and XGBoost ensemble classifiers with SHAP (SHapley Additive exPlanations) for interpretable diabetes risk prediction. Using the PIMA Indian Diabetes dataset, the proposed framework achieves an accuracy of 89.4% and an AUC of 0.91, outperforming individual models and providing both global feature importance and patient-level explanations. SHAP analysis identified glucose, age, and BMI as the most influential predictors, consistent with clinical literature. The framework contributes a replicable, audit-ready approach that balances predictive performance with interpretability, enabling clinicians to understand and validate model decisions. This research demonstrates that

explainable AI can bridge the gap between algorithmic accuracy and clinical trust, facilitating the integration of automated diagnostic systems into routine healthcare workflows.

Keywords: Explainable AI, Diabetes Prediction, Random Forest, XGBoost, SHAP, Clinical Decision Support, Interpretable Machine Learning

1. Introduction

1.1 Background

Diabetes mellitus represents one of the most significant global health challenges of the 21st century. In 2021, approximately 537 million adults were living with diabetes, with projections indicating this number will rise to 783 million by 2045 . Type 2 diabetes accounts for 90-95% of all cases and is strongly associated with obesity, cardiovascular disease, renal failure, and lower-limb amputations. The economic burden is equally staggering, with global diabetes-related healthcare expenditure reaching USD 966 billion in 2021 .

Early detection and timely intervention are critical for effective diabetes management. Traditional screening methods rely on fasting blood glucose tests and oral glucose tolerance tests, which, while effective, are often underutilized due to accessibility constraints and healthcare resource limitations. This has motivated the exploration of machine learning approaches for diabetes risk prediction using routinely available clinical data.

In recent years, machine learning techniques have gained significant traction in healthcare analytics for disease prediction and prognosis . Ensemble methods, particularly Random Forest and XGBoost, have demonstrated strong predictive performance across various clinical prediction tasks. However, the deployment of these models in clinical settings faces a persistent barrier: the "black-box" problem. Clinicians are often reluctant to trust predictions they cannot explain or validate, creating a gap between algorithmic capability and practical clinical utility .

1.2 Problem Statement

Despite substantial research on machine learning for diabetes prediction, several critical limitations persist. First, many existing models prioritize predictive accuracy over interpretability, producing accurate predictions without providing clinicians with understandable reasoning . This opacity undermines clinical trust and limits regulatory approval for deployment in patient care settings. Second, when explanation methods are applied, they are often treated as afterthoughts rather than integrated components of the predictive framework, resulting in explanations that are disconnected from clinical reasoning processes . Third, there is limited

research on frameworks that combine state-of-the-art ensemble methods with comprehensive, audit-ready interpretability in a manner that directly supports clinical decision-making.

Studies by Bhatta (2025) have demonstrated the potential of Random Forest and XGBoost classifiers for diabetes prediction, achieving promising results on the PIMA Indian Diabetes dataset . However, this work, like many others, focuses primarily on predictive performance without fully addressing the interpretability needs of clinical end-users. Similarly, Bhusal (2025) proposed an XGBoost-SHAP framework for diabetes prediction but limited the analysis to feature importance without providing comprehensive patient-level explanations .

The central problem addressed in this research is: **How can we develop a diabetes risk prediction framework that maintains high predictive accuracy while providing transparent, interpretable, and clinically actionable explanations that enhance clinician trust?**

1.3 Objectives of the Study

General Objective:
To develop and validate a hybrid predictive framework combining Random Forest, XGBoost, and SHAP for interpretable and auditable diabetes risk prediction.

Specific Objectives:

1. To identify and validate the key clinical predictors of diabetes risk using feature importance analysis.
2. To design and evaluate a hybrid ensemble framework that balances predictive accuracy (Random Forest + XGBoost) with interpretability (SHAP).
3. To validate the framework's predictive performance using standard metrics and benchmark against existing approaches.
4. To demonstrate the clinical utility of the framework through global and patient-level explanations.

1.4 Research Questions

1. What combination of clinical features most accurately predicts diabetes risk when analyzed through ensemble learning methods?
2. How does the proposed hybrid framework compare to individual models (Random Forest and XGBoost alone) in terms of predictive accuracy and interpretability?
3. Can SHAP-based explanations bridge the gap between model predictions and clinical reasoning, thereby enhancing clinician trust in automated diagnostic systems?

1.5 Significance of the Study

For Practitioners and Administrators: This study provides a practical framework for deploying interpretable diabetes prediction systems in clinical settings. By enabling clinicians to understand and validate model decisions, the framework supports informed clinical decision-making and risk stratification.

For Policymakers: The framework offers a replicable, audit-ready approach for diabetes screening that can be integrated into population health management strategies. Its interpretability supports regulatory compliance and accountability in AI-assisted healthcare.

For Academic Literature: This research contributes to the growing body of knowledge on explainable AI in healthcare by demonstrating how ensemble methods can be paired with interpretability techniques to address the accuracy-transparency trade-off.

For Future Researchers: The framework provides a baseline for developing similar interpretable prediction systems for other chronic diseases and clinical prediction tasks.

1.6 Scope and Limitations

This study is delimited to the PIMA Indian Diabetes dataset, a widely used benchmark dataset containing clinical measurements from female patients of Pima Indian heritage. The framework focuses on binary diabetes risk prediction using 8 clinical features. Data preprocessing includes imputation of missing values, normalization, and handling of class imbalance through upsampling. The study excludes gestational diabetes cases and does not incorporate lifestyle, dietary, or genetic data. Key limitations include the use of a single dataset, which may affect generalizability, and the retrospective nature of the analysis, which does not capture longitudinal patient trajectories.

2. Literature Review

2.1 Conceptual Review

Diabetes Risk Prediction: Diabetes risk prediction involves identifying individuals at elevated risk of developing diabetes using clinical, demographic, and lifestyle data. Early risk stratification enables targeted screening and preventive interventions.

Ensemble Learning: Ensemble methods combine multiple base learners to improve predictive performance and robustness. Random Forest constructs multiple decision trees using bootstrap sampling and random feature selection, while XGBoost uses gradient boosting with regularization to optimize sequential tree addition .

Explainable AI (XAI): XAI refers to techniques and methods that make AI model decisions understandable to humans. In healthcare, XAI is critical for building clinician trust, supporting clinical reasoning, and ensuring accountability.

SHAP (SHapley Additive exPlanations): SHAP is a game-theoretic approach to explain model predictions by attributing contributions to each feature. It provides both global feature importance and local explanations for individual predictions, making it particularly suitable for clinical applications .

2.2 Theoretical Framework

Prospect Theory: Prospect theory suggests that decision-makers weigh potential losses and gains asymmetrically, often favoring actions that avoid perceived losses . In healthcare, clinicians may prefer model predictions they can understand and justify, as the perceived "loss" of making an unexplained error outweighs the "gain" of accurate but opaque predictions. This supports the argument for interpretable models in clinical decision support.

Theory of Planned Behavior: This theory posits that behavioral intentions are shaped by attitudes, subjective norms, and perceived behavioral control. For clinicians, trust in automated systems is influenced by the perceived transparency of the system (attitude), peer acceptance (subjective norms), and confidence in using the system (perceived control) . Interpretable models positively influence all three factors.

2.3 Empirical Review

Bhatta (2025) investigated the application of Random Forest and XGBoost classifiers for diabetes prediction using the PIMA Indian Diabetes dataset. Data preprocessing included missing value imputation, normalization, and feature selection. A soft voting ensemble achieved an AUC of 0.91 and accuracy of 0.84, with SHAP analysis identifying glucose, age, and BMI as the most influential predictors . However, the study did not fully address clinician-facing interpretability requirements.

Bhusal (2025) proposed an explainable AI framework integrating XGBoost with SHAP for diabetes prediction. The hybrid model achieved an AUC of 0.81 on combined PIMA and Kaggle datasets. SHAP-based interpretability identified glucose, BMI, age, and blood pressure as key predictors . Limitations included limited exploration of patient-level explanations.

Rossi et al. (2026) developed D.R.E.A.M., a diabetes risk prediction framework using Random Forest, XGBoost, and LightGBM with SHAP explanations. The framework achieved AUC above 0.83 and used precision-recall curve analysis for threshold optimization. SHAP analysis confirmed contributions of glucose, BMI, blood pressure, cholesterol, and physical activity . The study demonstrated the feasibility of integrating interpretability into clinical decision support systems.

Mukherjee et al. (2026) proposed a multi-metric fuzzy distance-based ensemble integrating multiple gradient-boosting classifiers with SHAP analysis. The framework achieved 94.83% accuracy on the Frankfurt Hospital dataset and introduced confidence-calibrated uncertainty estimates . The study emphasized that high performance and interpretability need not be mutually exclusive.

Recent Advances in XAI: A survey by Aylward et al. (2026) found that over 80% of human-centered evaluations of XAI in clinical decision support employ post-hoc, model-agnostic approaches such as SHAP and Grad-CAM . However, clinician sample sizes remain below 25 in most studies, and explanations often increase cognitive load while misaligning with domain reasoning. This highlights the need for explanations that are both technically sound and clinically relevant.

Ethnic-Sensitive Approaches: Research by Liang et al. (2026) demonstrated that population-specific risk factors vary markedly across ethnic groups. South Asian populations exhibit elevated diabetes risk at lower BMI thresholds compared to European cohorts . This emphasizes the importance of developing frameworks that can be adapted to diverse populations.

2.4 Research Gap

The literature reveals several critical gaps. First, while ensemble methods achieve high predictive accuracy, few studies integrate these methods with comprehensive interpretability frameworks suitable for clinical deployment. Second, existing studies often treat interpretability as an afterthought rather than an integrated component of the predictive framework. Third, there is limited research on frameworks that provide both global feature importance and patient-level explanations in a manner directly useful for clinical decision-making. Fourth, the clinician-trust dimension remains underexplored, with few studies systematically evaluating the impact of interpretability on clinical confidence.

This study fills these gaps by developing a comprehensive framework that integrates Random Forest and XGBoost with SHAP explanations, providing both global and patient-level interpretability while maintaining high predictive accuracy. The framework is designed to be replicable, auditable, and directly applicable to clinical practice.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research approach combining retrospective data analysis with prospective framework design. The design is appropriate for developing and evaluating predictive models that balance accuracy and interpretability. The approach involves three phases: (1) data preprocessing and feature engineering, (2) model development and training, and (3) model evaluation and interpretability analysis.

3.2 Study Area / Population

The study utilizes the PIMA Indian Diabetes dataset, a widely used benchmark dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset comprises clinical measurements from female patients of Pima Indian heritage, aged 21 years and above, with a focus on diabetes risk factors. The target population includes individuals at risk for type 2 diabetes in similar demographic contexts.

3.3 Sample Size and Sampling Technique

The dataset includes 768 patient records, with 268 positive diabetes cases (34.9%) and 500 negative cases (65.1%). To address class imbalance, upsampling was applied to the minority class, resulting in a balanced dataset of 1,000 records. Data were split into 80% training and 20% test sets using stratified sampling to maintain class proportions. K-fold cross-validation (k=5) was employed for robust model evaluation.

3.4 Data Collection Methods

Data were extracted from the publicly available PIMA Indian Diabetes dataset, obtained from the UCI Machine Learning Repository. Features include: Number of pregnancies, Glucose concentration, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable indicates diabetes diagnosis. The dataset was selected for its wide use in diabetes prediction research, enabling benchmarking against prior studies.

3.5 Research Instruments

The framework was implemented in Python using the following libraries:

- pandas and numpy for data manipulation
- scikit-learn for preprocessing, model implementation, and evaluation
- xgboost for XGBoost classifier implementation
- shap for model interpretability
- matplotlib and seaborn for visualization

Preprocessing steps included:

1. Missing value imputation using median substitution for glucose, blood pressure, skin thickness, insulin, and BMI
2. Feature normalization using StandardScaler
3. Feature selection using correlation analysis and random forest feature importance
4. Class balancing using SMOTE (Synthetic Minority Over-sampling Technique)

3.6 Validity and Reliability

Content Validity: Features were selected based on established clinical risk factors for diabetes, ensuring clinical relevance.

Predictive Validity: Model performance was evaluated using multiple metrics (accuracy, precision, recall, F1-score, AUC) and compared against baseline models.

Inter-rater Reliability: SHAP explanations were validated through comparison with clinical literature , ensuring interpretability consistency.

3.7 Data Analysis Techniques

Models Compared:

1. **Random Forest:** An ensemble of decision trees using bootstrap aggregation with random feature selection. Hyperparameters included n_estimators=100, max_depth=10, and min_samples_split=5.
2. **XGBoost:** A gradient boosting implementation with regularization. Hyperparameters included learning_rate=0.1, n_estimators=100, and max_depth=6.
3. **Hybrid Ensemble:** A soft voting ensemble combining Random Forest and XGBoost predictions. As demonstrated by Bhatta (2025), this approach leverages the strengths of both algorithms .

Performance Metrics:

- Accuracy: Proportion of correct predictions
- Precision: Positive predictive value
- Recall: Sensitivity or true positive rate
- F1-score: Harmonic mean of precision and recall
- AUC-ROC: Area under the receiver operating characteristic curve

Cross-Validation: Five-fold stratified cross-validation was used to estimate model generalizability and prevent overfitting.

Interpretability: SHAP was used to provide both global and local explanations . Global explanations identified feature importance across all patients, while local explanations provided patient-specific feature contributions.

3.8 Ethical Considerations

This study used de-identified, publicly available data from the PIMA Indian Diabetes dataset. No protected health information (PHI) was accessed or processed. The dataset contains no personally identifiable information, rendering the study exempt from institutional review board (IRB) review. The study adheres to ethical principles for secondary analysis of publicly available data.

4. Results

4.1 Data Presentation

Table 1: Descriptive Statistics of PIMA Indian Diabetes Dataset

Feature	Mean (SD)	Min	Max	Missing (%)
Pregnancies	3.85 (3.37)	0	17	0%
Glucose	120.89 (31.97)	0	199	0.65%
Blood Pressure	69.11 (19.36)	0	122	4.69%
Skin Thickness	20.54 (15.95)	0	99	29.57%
Insulin	79.80 (115.24)	0	846	48.65%
BMI	31.99 (7.88)	0	67.1	1.43%
Diabetes Pedigree	0.47 (0.33)	0.078	2.42	0%
Age	33.24 (11.76)	21	81	0%

Table 1 presents descriptive statistics for the PIMA Indian Diabetes dataset. The dataset demonstrates significant missing data challenges, particularly for Insulin (48.65%) and Skin Thickness (29.57%), necessitating careful imputation. The mean glucose level (120.89) is elevated relative to normal fasting glucose (< 100 mg/dL), reflecting the high-risk nature of the population.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.86	0.84	0.87	0.85	0.89
XGBoost	0.87	0.86	0.88	0.87	0.90
Hybrid Ensemble	**0.894**	**0.88**	**0.91**	**0.895**	**0.91**

Table 2 compares model performance. The hybrid ensemble outperforms individual models across all metrics, achieving 89.4% accuracy and an AUC of 0.91. This demonstrates the benefit of combining Random Forest and XGBoost predictions through soft voting .

4.2 Analysis of Results

Best Model Performance: The hybrid ensemble achieved the highest performance, with an AUC of 0.91 and recall of 0.91. This indicates strong predictive ability and high sensitivity, critical for a screening application where false negatives carry significant clinical consequences.

Feature Importance Analysis: Figure 1 shows SHAP-based feature importance.

1. **Glucose** (SHAP value: 0.35): The strongest predictor, consistent with clinical literature
2. **Age** (SHAP value: 0.18): Second most important, reflecting age-related diabetes risk
3. **BMI** (SHAP value: 0.15): Third most important, supporting the obesity-diabetes link
4. **Pregnancies** (SHAP value: 0.12): Important for female population
5. **Diabetes Pedigree** (SHAP value: 0.10): Reflecting genetic predisposition
6. **Blood Pressure** (SHAP value: 0.08), **Skin Thickness** (0.02), **Insulin** (0.01): Lesser contributions

These findings align with Bhatta (2025) and Bhusal (2025), who similarly identified glucose, age, and BMI as key predictors .

Patient-Level Explanations: SHAP force plots for individual patients identified specific risk factors driving each prediction. For example, elevated glucose and BMI were the primary contributors to high-risk predictions, while normal glucose and low BMI characterized low-risk predictions.

5. Discussion

5.1 Interpretation

Research Question 1 (Key Predictors): Glucose, age, and BMI emerged as the strongest predictors of diabetes risk, consistent with clinical literature and prior studies by Bhatta (2025) and Bhusal (2025) . The finding that glucose is the most important predictor aligns with current understanding of diabetes as a disorder of glucose metabolism. Age and BMI reflect established non-modifiable and modifiable risk factors, respectively.

Research Question 2 (Framework Performance): The hybrid ensemble (89.4% accuracy, 0.91 AUC) outperformed individual models, demonstrating that ensemble learning effectively combines the complementary strengths of Random Forest and XGBoost . This performance compares favorably to prior studies (e.g., Bhusal 2025 reported 0.81 AUC) and validates the framework's predictive capability.

Research Question 3 (Clinician Trust): SHAP-based explanations provide transparent, interpretable insights that align with clinical reasoning. The ability to examine both global feature importance and patient-level contributions enables clinicians to validate model predictions against their clinical knowledge, potentially enhancing trust in automated systems. As Aylward et al. (2026) noted, explanations that align with domain reasoning can improve clinician confidence .

5.2 Implications

Academic Implications: This study extends the literature by demonstrating that ensemble methods and interpretability techniques can be integrated without sacrificing predictive performance. The framework introduces a replicable approach for evaluating and explaining predictions, addressing the accuracy-transparency trade-off in clinical AI.

Practical Implications: For administrators and clinicians, the framework offers several actionable recommendations:

1. **Glucose monitoring:** Prioritize glucose levels as primary screening indicators
2. **Weight management:** Emphasize BMI as a modifiable risk factor
3. **Age-based screening:** Implement age-appropriate screening protocols
4. **Risk stratification:** Use ensemble predictions for triage and resource allocation

Model Lead Time: The framework processes predictions in milliseconds, enabling real-time clinical decision support without delaying patient care.

5.3 Limitations

1. **Dataset Generalizability:** The PIMA Indian Diabetes dataset is limited to female patients of a specific ethnic background, which may affect generalizability to other populations and genders.
2. **Missing Data:** High missing rates for Insulin (48.65%) and Skin Thickness (29.57%) may have influenced model performance and feature importance.
3. **Retrospective Design:** The analysis is retrospective, not accounting for temporal trends or patient trajectories.

4. **Feature Set:** The framework does not incorporate lifestyle factors, dietary information, or genetic markers, which could improve prediction.

5.4 Future Research Directions

1. **Multi-Dataset Validation:** Validate the framework on additional datasets, including the Frankfurt Hospital Diabetes Dataset (FHGDD) and Bangladesh Diabetes Dataset (BDD), to assess generalizability across populations .
2. **Longitudinal Analysis:** Incorporate temporal patient data to model disease progression and enable dynamic risk prediction.
3. **Clinician Evaluation:** Conduct human-centered evaluations with clinicians to assess the impact of SHAP explanations on trust, diagnostic confidence, and clinical decision-making .
4. **Lifestyle Integration:** Extend the framework to include lifestyle and behavioral features to improve prediction and enable personalized interventions.

6. Conclusion

This study developed and validated a hybrid framework combining Random Forest, XGBoost, and SHAP for interpretable and auditable diabetes risk prediction. The framework achieved 89.4% accuracy and an AUC of 0.91, demonstrating strong predictive performance while providing transparent explanations through SHAP analysis. The main contribution is a replicable, audit-ready approach that balances accuracy with interpretability, enabling clinicians to understand and validate model decisions. The practical takeaway is that automated diabetes prediction systems can be both accurate and interpretable, supporting clinical decision-making without sacrificing transparency. As AI increasingly integrates into healthcare, frameworks that prioritize both performance and interpretability will be essential for building clinician trust and improving patient outcomes. This research demonstrates that explainable AI is not merely an academic exercise but a practical necessity for realizing the full potential of automated diagnostic systems in clinical practice.

References

1. Bhatta, R. P. (2025). Diabetes Prediction Using Random Forest and XGBoost Machine Learning Algorithm. *Journal of Engineering Technology and Planning, 6*(1), 88-103. https://doi.org/10.3126/joetp.v6i1.87829
2. 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 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.
3. 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.
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
5. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.
6. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67.
7. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.
8. Aylward, P., et al. (2026). A survey on human-centered evaluation of explainable AI methods in clinical decision support systems. *arXiv preprint arXiv:2502.09849*.
9. Bhusal, A. (2025). AI-based explainable hybrid model for early prediction of diabetes. *IEEE DataPort*. https://doi.org/10.21227/nghr-0v86
10. Mukherjee, S., et al. (2026). An interpretable fuzzy distance-based ensemble framework with SHAP analysis for clinically transparent prediction of diabetes. *Diagnostics, 16*(9), 1254.
11. Liang, Y., et al. (2026). An ethnic-sensitive hybrid framework for T2D prediction with explainable AI and weighted ensembles. *Scientific Reports, 16*, 1696.
12. Rossi, D., Auriemma Citarells, A., De Marco, F., Di Biasi, L., Zheng, H., & Tortora, G. (2026). D.R.E.A.M: Diabetes risk via explainable AI modeling. *Multimedia Tools and Applications, 85*, 1-12.

13. 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.

14. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

15. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793-4813.