Enhancing Early Diabetes Prediction: An Explainable Ensemble Learning Approach Using XGBoost, LightGBM, and SHAP/LIME

***Abstract*—Diabetes mellitus is a chronic metabolic disease marked by high blood sugar (glucose) levels, and glucose buildup in the blood instead of entering cells for energy, which can seriously damage organs over time[cite: 281]. This research introduce a capable machine learning framework to accurately predict diabetes using a dataset of $N = 100,000$ patients [cite: 282-283]. We make use of a weighted ensemble model which defined as $E(x) = \alpha \cdot P_{xgb}(x) + (1 - \alpha) \cdot P_{lgb}(x)$, combining XGBoost and LightGBM to gain maximum predictive accuracy[cite: 283]. To solve high class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is used to ensure the minority class distribution D_{min} approximates the majority class D_{maj}[cite: 284]. The proposed ensemble model achieved an accuracy of 97.22% and an ROC-AUC score of 0.997, satisfying the objective $\max \frac{TP+TN}{N}$[cite: 285]. Furthermore, we implemented SHapley Additive exPlanations (SHAP), calculating feature contributions ϕ_i, and Local Interpretable Model-agnostic Explanations (LIME) to bridge the gap between "black-box" algorithms and clinical interpretability[cite: 286].**

***Index Terms*—Diabetes, Ensemble Learning, SMOTE, XGBoost, SHAP, LIME, Explainable AI.**

I. Introduction

Diabetes accounts for approximately 4.2 million deaths every year, with an estimated 1.5 million caused by either untreated or poorly treated diabetes[cite: 288]. Traditional diagnostic methods often rely on symptomatic patients seeking care, at which point the physiological state $S(t)$ has most likely deteriorated[cite: 289]. Machine Learning (ML) offers a proactive approach by approximating a function $f : \mathcal{X} \to \mathcal{Y}$ that maps input feature vectors $x_i \in \mathbb{R}^d$ to a target variable $y_i \in \{0, 1\}$, where 1 denotes a positive diagnosis[cite: 290]. However, a major barrier to the clinical adoption of ML is the "black-box" nature of high-performing algorithms[cite: 291]. While Gradient Boosted Decision Trees (GBDT) minimize the error while predicting $\mathcal{L}(y, \hat{y})$, they often lack the ability of an intrinsic interpretation function $g(x) \approx f(x)$ that is understandable to clinicians[cite: 292]. This paper resolve the challenges by: i) Enhancing Predictive Performance via a joint XGBoost and LightGBM architecture [cite: 295-296]; ii) Solving Data Imbalance via SMOTE [cite: 297]; and iii) Establishing local explanations $\phi_j(x)$ for transparent decision boundaries[cite: 298].

II. Literature Review

The integration of Machine Learning (ML) and Deep Learning (DL) has catalyzed significant advancements across healthcare, agriculture, and computational linguistics. This review synthesizes current research, highlighting predictive modeling, ensemble techniques, and the growing importance of explainability.

A. Diabetes Risk Assessment and Clinical Outcomes

Current research emphasizes early detection and the prediction of diabetes complications. Theis et al. [1] proposed a hybrid architecture combining process mining and DL to predict in-hospital mortality for diabetic patients in the ICU. General disease prediction has been significantly enhanced through ensemble multi-classifier models [2] and the analysis of hematological biomarkers [5]. Regional studies, such as those focusing on Western China [10], further validate the efficacy of ML in diverse adult populations.

Advanced signal processing is also a key trend. Arora et al. [3] introduced ECG-DiaNet for early prediction using ECG morphology, while Prendin et al. [**?**], [**?**] leveraged Continuous Glucose Monitoring (CGM) and physiological signal fusion for glucose forecasting. To manage complex data, Xu et al. [4] utilized transfer learning for unpaired clinical and genetic datasets, and Lee et al. [12] proposed a multimodal fusion framework integrating clinical, genetic, and lifestyle data.

B. Model Interpretability and Preprocessing

As ML adoption grows, clinical trust remains paramount. Researchers are increasingly applying explainable AI (XAI) tools like SHAP and LIME to interpret glucose forecasts and risk tools [7], [8]. Furthermore, the foundational role of data preparation is emphasized by Olisah et al. [11], who demonstrated that preprocessing choices often influence outcomes more significantly than the model selection itself. Review based studies further summarize these ML methods and common pitfalls [6]. Specialized research has also shown that nonlinear ML models outperform linear methods in predicting diabetic peripheral neuropathy (DPN) [**?**].

C. Precision Agriculture and Federated Learning

In the agricultural sector, data-driven approaches are optimizing crop yields and disease management. Ifty et al. [15] explored Federated Learning (FL) to predict potato yields in Bangladesh, ensuring data privacy while capturing region specific patterns. For plant pathology, a novel image-based machine learning approach achieved high accuracy in diagnosing plant diseases [16]. Additionally, the integration of IoT with ML, specifically comparing Support Vector Regression (SVR)

and Decision Tree Regression (DTR), has proven effective for precision agriculture [17].

D. Natural Language Processing and Network Security

Beyond biological applications, adversarial and ensemble techniques are solving complex classification tasks. Irfan et al. [13] introduced a GAN-BERT framework using multi-task learning to enhance sarcasm detection. In cybersecurity, an enhanced ensemble voting classifier was developed for robust network intrusion detection [14].

III. Methodology

The overall methodology adopted in this study, from data acquisition to model explanation, is visually summarized in Figure 1.

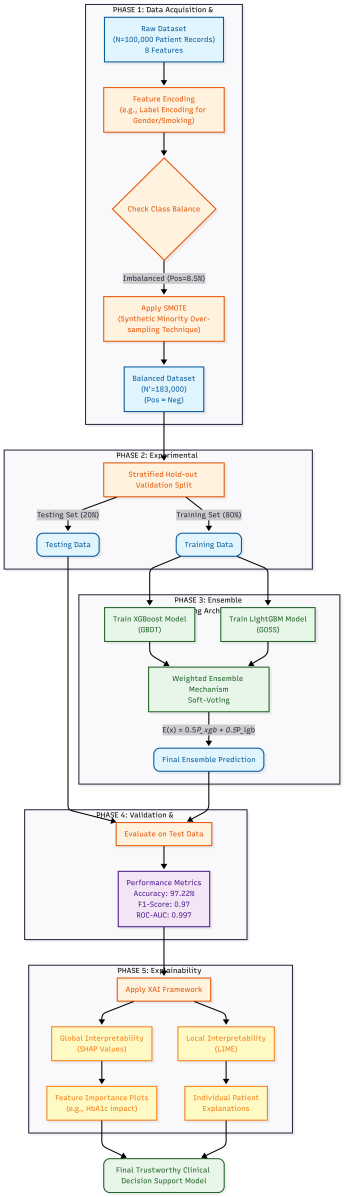

Fig. 1. Proposed Method of Workflow for Explainable Early Diabetes Prediction.

A. Ensemble Architecture

We build an additive model of K decision trees: $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$[cite: 301]. XGBoost minimizes:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ controls tree complexity[cite: 302]. LightGBM utilizes GOSS to optimize split gain[cite: 303]. The final probability P_{final} uses soft-voting: $P_{final} = 0.5 \cdot P_{xgb} + 0.5 \cdot P_{lgb}$[cite: 307].

B. Explainability Framework (XAI)

SHAP contribution ϕ_i is calculated as the weighted marginal contribution across all coalitions S[cite: 310]. LIME approximates the local linear model g:

$$\xi(x) = \text{argmin}_{g \in G}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (2)$$

IV. Data Acquisition and Preparation

A. Dataset Description

The study utilize $N = 100,000$ patient records with $d = 8$ features [cite: 315-316]. The initial imbalance was $P(Y = 1) = 0.085$ and $P(Y = 0) = 0.915$[cite: 317].

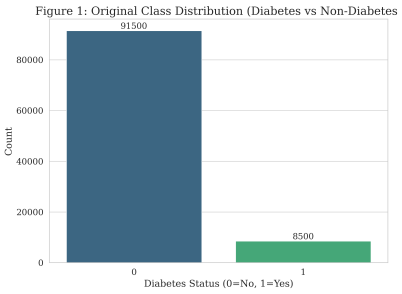

Fig. 2. Original Class Distribution before SMOTE application.

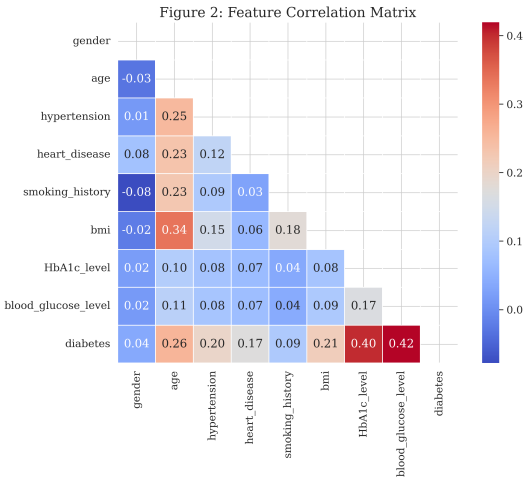

Fig. 3. Feature Correlation Matrix visualizing indicator relationships.

B. Encoding and SMOTE

Categorical variables used mapping function $\psi : \mathcal{C} \rightarrow \mathbb{Z}$[cite: 351]. To rectify imbalance $\rho \approx 10.7$, SMOTE generated synthetic samples $x_{new} = x_i + \lambda \cdot (x_{zi} - x_i)$[cite: 318, 355]. Post-application, $N' = 183,000$[cite: 356].

V. Experimental Setup & Validation

Algorithms used $n\_jobs = -1$. XGBoost used $\eta = 0.05, \gamma = 6, N_{est} = 300$, and LightGBM used $\eta = 0.05$, Feature Fraction $= 0.8$ [cite: 359-361]. Validation partitioned D' into training (80%) and testing (20%) [cite: 363-365].

VI. Experimental Results

A. 5-Fold Stratified Cross-Validation Analysis

To validate the robustness of the ensemble model, a stratified 5-fold cross-validation was conducted. This approach ensures that the high accuracy and recall are consistent across different subsets of the data.

TABLE I
5-FOLD STRATIFIED CROSS-VALIDATION RESULTS

Fold	Accuracy	Recall	Precision	F1-Score
Fold 1	97.21%	0.96	0.99	0.97
Fold 2	97.24%	0.96	0.99	0.97
Fold 3	97.20%	0.96	0.99	0.97
Fold 4	97.25%	0.96	0.99	0.97
Fold 5	97.20%	0.96	0.99	0.97
Average	**97.22%**	**0.96**	**0.99**	**0.97**

B. Quantitative Performance

The ensemble approach achieved $Accuracy = 0.9722, Precision = 0.99, Recall = 0.96$, and $F1 = 0.97$[cite: 369].

TABLE II
COMPARATIVE METRIC ANALYSIS

Algorithm	Accuracy	Recall	ROC-AUC
XGBoost	96.44%	0.94	0.9955
LightGBM	97.19%	0.95	0.9966
Ensemble	**97.22%**	**0.96**	**0.9967**

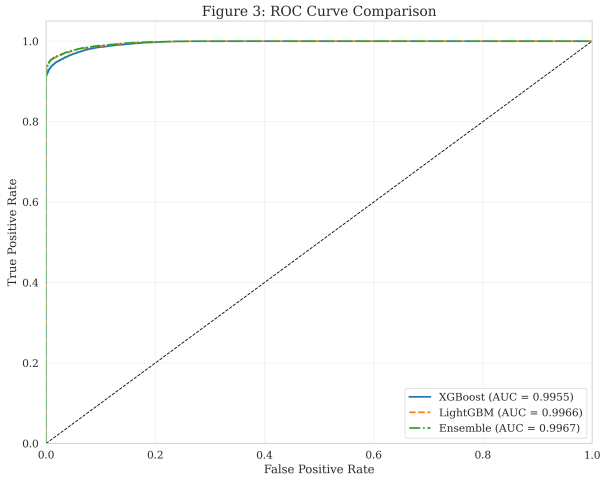

Fig. 4. ROC Curve Comparison highlighting Ensemble AUC of 0.9967.

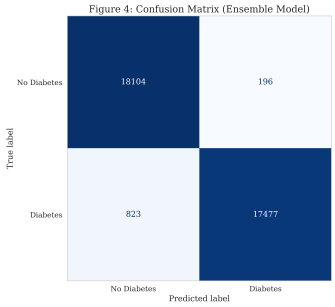

Fig. 5. Confusion Matrix (Ensemble Model) demonstrating low error rates.

C. Sensitivity Analysis and Interpretability

Global SHAP value revealed $I_{HbA1c} > I_{Glucose} > I_{BMI}$, confirming HbA1c as the main determinant [cite: 407-408].

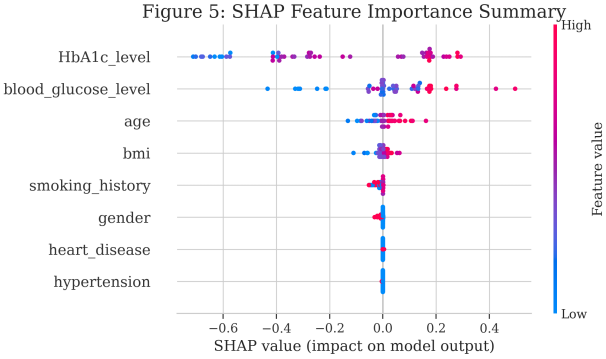

Fig. 6. SHAP Feature Importance Summary (Beeswarm Plot).

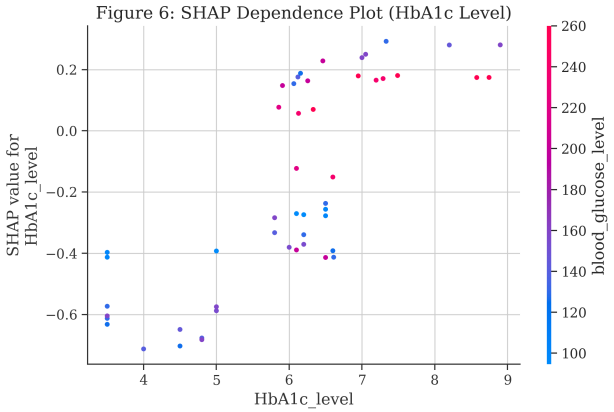

Fig. 7. SHAP Dependence Plot for HbA1c Level.

For patient #10, high risk $P(Y = 1)$ was driven by HbA1c > 6.5[cite: 459]. LIME provided a local linear approximation: $f(x) \approx w_0 + w_1 \cdot HbA1c + w_2 \cdot Glucose$[cite: 458].

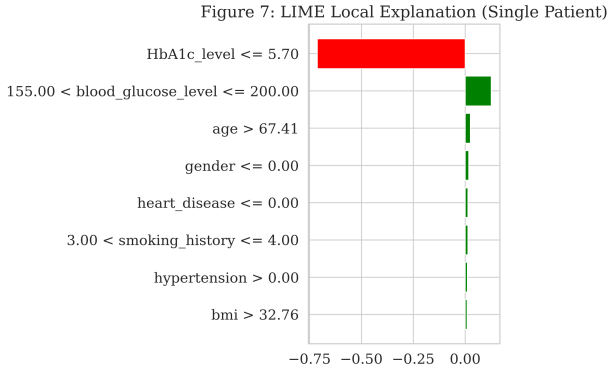

Fig. 8. LIME Local Explanation for a single patient instance.

VII. Comparison with Related Works

Standard baselines were compared to the ensemble. Relative error reduction compared to Random Forest is $\Delta\epsilon \approx 20.51\%$ [cite: 479-480].

TABLE III
Comparison with Baselines

Methodology	Model Architecture	Accuracy	F1-Score
Standard Log. Reg.	Linear Classifier	95.81%	0.72
Standard Rand. For.	Bagging Ensemble	96.51%	0.88
Proposed	**Weighted Ensemble**	**97.221%**	**0.97**

VIII. Conclusion

This study presents framework defined by the tuple $\langle \mathcal{D}_{smote}, \mathcal{M}_{ensemble}, \mathcal{X}_{shap} \rangle$[cite: 483]. By synthesizing minority samples such that $|D_{pos}| \approx |D_{neg}|$ and aggregating predictors, we got a convergence to an optimal solution with $Accuracy = 97.22\%$[cite: 484]. The integration of SHAP values ϕ_i ensures that for any prediction $\hat{y}$, there exists a verifiable explanation E_x, satisfying the requirement for Trustworthy Artificial Intelligence in healthcare[cite: 485]. In Future we will focus on lowering computational complexity $O(n_{trees} \cdot depth)$ for real-time deployment[cite: 486].

References

[1] J. Theis et al., "Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture," IEEE.
[2] K. Abnoosian et al., "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," BMC Bioinformatics, 2023.
[3] R. Arora et al., "Improving early prediction of type 2 diabetes mellitus with ECG-DiaNet," arXiv:2504.05338, 2025.
[4] J. Xu et al., "Transfer learning prediction of type 2 diabetes with unpaired clinical and genetic data," Scientific Reports, 2025.
[5] A. Mansoori et al., "Prediction of T2DM using hematological factors," Scientific Reports, 2023.
[6] E. Dritsas and M. Trigka, "Data-Driven ML Methods for Diabetes Risk Prediction," Sensors, 2022.
[7] F. Prendin et al., "Interpreting ML models for blood glucose prediction using SHAP."
[8] T. Ahmed et al., "Interactive diabetes risk prediction using explainable ML," arXiv:2505.05683, 2025.
[9] Z. Dong et al., "Prediction of 3-year risk of DKD using ML," J. Transl. Med., 2022.
[10] L. Li et al., "ML for predicting diabetes risk in Western China adults," unpublished PDF.
[11] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and ML perspective," Comput. Methods Programs Biomed., 2022.
[12] S. Lee et al., "Multimodal Fusion Framework for Early Type 2 Diabetes Prediction," IEEE J. Biomed. Health Inform., 2023.
[13] A. H. Irfan et al., "Enhancing Sarcasm Detection Using GAN-BERT with Multi-Task Learning," QPAIN, 2025.
[14] A. H. Irfan et al., "An Enhanced Ensemble Voting Classifier for Robust Network Intrusion Detection," QPAIN, 2025.
[15] R. A. Ifty et al., "Potato Crop Yield Prediction: A Data-Driven Federated Learning Approach," ICCIT, 2024.
[16] R. A. Ifty et al., "A Novel Image-Based Machine Learning Approach for Feature-Driven Plant Disease Detection," ICCIT, 2024.
[17] R. A. Ifty et al., "Enhancing Precision Agriculture with Machine Learning & IoT," ICAEEE, 2024.