ISSN: 2096-3246
Volume 58, Issue 2, 2026

COMPARATIVE EVALUATION OF SHAP AND LIME FOR EXPLAINABLE TYPE 2 DIABETES PREDICTION MODELS

Devinder Kumar
PhD Scholar, Gandhinagar Institute of Computer Science And Applications, Gandhinagar University.
Email - devinder09999@gmail.com
Dr. Angira A. Patel
Associate Professor, Gandhinagar Institute of Computer Science And Applications, Gandhinagar University.
Email - angira.it@gmail.com

Abstract

Type 2 Diabetes Mellitus (T2DM) is a globally escalating chronic metabolic disorder affecting over 537 million adults worldwide as of 2021, with projections surpassing 780 million by 2045. Early and accurate prediction of T2DM through machine learning (ML) models offers significant clinical promise; however, the deployment of opaque, black-box predictive systems in healthcare raises substantial concerns regarding interpretability, accountability, and patient trust. Explainable Artificial Intelligence (XAI) methodologies have emerged as pivotal solutions to address these challenges. This paper presents a rigorous comparative evaluation of two leading XAI frameworks — SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) — applied to multiple state-of-the-art ML classifiers for T2DM prediction, including Random Forest (RF), Gradient Boosting (XGBoost), LightGBM, Support Vector Machine (SVM), and Logistic Regression (LR). Experiments were conducted on the benchmark PIMA Indians Diabetes Dataset (PIDD) and an augmented clinical cohort. Key evaluation dimensions include predictive accuracy, explanation fidelity, computational efficiency, stability, and clinical interpretability. XGBoost with SHAP yielded the highest classification accuracy (AUC = 0.947), while LIME demonstrated superior explanation locality and model-agnostic adaptability. SHAP consistently produced more stable, globally coherent feature attributions compared to LIME's locally bounded, computationally intensive explanations. The findings provide actionable recommendations for clinical AI practitioners selecting XAI tools and highlight avenues for hybrid XAI architectures in diabetes management systems.

Keywords: Explainable Artificial Intelligence (XAI), SHAP, LIME, Type 2 Diabetes Mellitus, Machine Learning, Feature Importance, Random Forest, XGBoost, Clinical Decision Support, Health Informatics

1. Introduction

Type 2 Diabetes Mellitus (T2DM) represents one of the most significant global public health crises of the 21st century. According to the International Diabetes Federation (IDF), approximately 537 million adults (20–79 years) were living with diabetes in 2021, with T2DM accounting for approximately 90–95% of all cases [1]. This figure is projected to rise to 643 million by 2030 and 783 million by 2045, imposing enormous healthcare burdens, particularly in low- and middle-income countries (LMICs) including India, which hosts the second-largest diabetic population globally [2].

Machine learning (ML) has demonstrated considerable promise in early diabetes prediction by identifying complex nonlinear patterns in clinical, demographic, and biochemical data. Algorithms including Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines have achieved high predictive accuracy in numerous studies [3,4]. However, the inherently opaque nature of such models — commonly termed 'black boxes' — creates significant barriers to clinical adoption. Clinicians require not merely accurate predictions but transparent, interpretable reasoning that aligns with established medical knowledge, supports regulatory compliance under frameworks such as the EU AI Act, and sustains patient trust [5].

Explainable Artificial Intelligence (XAI) has therefore emerged as an essential interdisciplinary field bridging machine learning performance and clinical interpretability. Among the most prominent XAI methodologies, SHapley Additive exPlanations (SHAP) [6] and Local Interpretable Model-agnostic Explanations (LIME) [7] have garnered substantial adoption in medical AI research. While both frameworks aim to illuminate model predictions, they differ fundamentally in theoretical foundations, scope (global versus local explanation), computational complexity, and suitability across different deployment scenarios.

Despite the growing body of literature applying SHAP or LIME individually to diabetes prediction tasks, a comprehensive, unified comparative evaluation of both frameworks across multiple ML models, using consistent datasets, standardized metrics, and clinical relevance criteria, remains notably absent. This gap motivates the present study.

1.1 Research Objectives

The primary objectives of this study are:

(i) To build and evaluate multiple ML classifiers for Type 2 Diabetes prediction on benchmark and augmented datasets.

(ii) To apply SHAP and LIME explainability frameworks to each trained model.

(iii) To compare SHAP and LIME across key dimensions: explanation fidelity, stability, computational cost, and clinical interpretability.

(iv) To identify the optimal SHAP/LIME configuration for clinical decision support in diabetes management.

(v) To propose a hybrid XAI architecture leveraging complementary strengths of both frameworks.

1.2 Contributions

The key contributions of this work are: (1) a systematic, multi-model comparative evaluation of SHAP and LIME under unified experimental conditions; (2) a novel six-dimensional XAI evaluation metric framework extending beyond accuracy; (3) a proposed hybrid XAI pipeline for clinical diabetes AI systems; and (4) empirically grounded clinical recommendations for XAI tool selection in health informatics.

2. Literature Review

2.1 Machine Learning for Diabetes Prediction

Research into ML-based diabetes prediction has intensified considerably since 2022, driven by the proliferation of electronic health records and wearable biosensors. Rashid et al. (2022) conducted a systematic review of 47 studies employing ML models for T2DM prediction, identifying Random Forest and XGBoost as the most frequently outperforming algorithms with average AUC values of 0.87–0.93 [8]. Ahmad et al. (2023) proposed a stacking ensemble integrating LightGBM, CatBoost, and ExtraTreesClassifier, achieving an AUC of 0.941 on the PIMA Indians Diabetes Dataset, underscoring the superiority of ensemble approaches over single classifiers [9].

Sarwar et al. (2023) explored deep learning architectures including bidirectional LSTMs for longitudinal diabetes prediction in electronic health record (EHR) datasets, achieving AUC = 0.952 but noting the critical limitation of interpretability deficiency [10]. Similarly, Kumar and Arora (2024) benchmarked fifteen ML algorithms across four diabetes datasets, concluding that while gradient boosting variants dominated accuracy metrics, logistic regression maintained superior interpretability-performance trade-offs in resource-constrained clinical environments [11].

2.2 Explainable AI in Healthcare

The application of XAI in healthcare has undergone rapid maturation since the landmark contributions by Lundberg and Lee (2017) for SHAP and Ribeiro et al. (2016) for LIME. In the post-2022 period, regulatory developments including the EU Artificial Intelligence Act (2023) and the WHO's guidance on ethics and governance of AI for health (2023) have intensified demand for explainable clinical AI systems [12].

Ghassemi et al. (2022) published a critical review in The Lancet Digital Health emphasizing that XAI for clinical models must satisfy not merely technical accuracy but also epistemic alignment with domain knowledge, causal coherence, and actionability for clinicians [13]. Tjoa and Guan (2022) provided a survey of XAI methods specifically applied to medical imaging and tabular clinical data, identifying SHAP as the most theoretically grounded method while noting LIME's superior adaptability to complex, non-differentiable model architectures [14].

2.3 SHAP Applications in Diabetes Research

SHAP has been applied extensively in diabetes research. Islam et al. (2023) applied SHAP to an XGBoost diabetes classifier, revealing glucose concentration, BMI, age, and diabetes pedigree function as the four most significant predictors — findings corroborated by endocrinological literature [15]. Ogunleye and Qing-Guo (2023) demonstrated that SHAP interaction values could identify synergistic feature interactions in T2DM prediction (e.g., BMI × Age interaction effects), providing mechanistic insights beyond single-feature attributions [16].

Notably, Nohara et al. (2022) conducted a pioneering clinical validation of SHAP explanations for diabetes risk models in a Japanese cohort study, demonstrating statistically significant concordance between SHAP-identified risk factors and established clinical risk markers including HbA1c, waist circumference, and family history [17]. Jabbar et al. (2024) extended SHAP to multi-modal diabetes datasets incorporating continuous glucose monitoring (CGM) data, demonstrating that SHAP TreeExplainer outperformed KernelExplainer in both accuracy and computational efficiency for tree-based models [18].

2.4 LIME Applications in Diabetes Research

LIME has found application in diabetes prediction primarily through its model-agnostic perturbation sampling approach. Sharma and Mittal (2022) applied LIME to a neural network diabetes classifier, demonstrating that while LIME provided useful local explanations for individual predictions, explanation instability across repeated samplings (variance coefficient > 0.18) posed reliability concerns in clinical contexts [19]. Mishra et al. (2023) proposed LIME-Stabilizer, an extension of LIME with kernel bandwidth optimization, reducing explanation variance by 34.7% compared to standard LIME while preserving locality [20].

A comparative study by Velmurugan et al. (2024) evaluated SHAP and LIME on diabetes risk prediction for a South Asian cohort, concluding that SHAP produced more clinically consistent explanations aligned with physician reasoning, whereas LIME offered superior explanatory granularity for borderline or ambiguous predictions [21]. This observation supports the hypothesis that SHAP and LIME possess complementary clinical utilities rather than one being universally superior.

2.5 Research Gap

Despite the growing literature, several critical gaps persist. First, most studies evaluate either SHAP or LIME in isolation rather than conducting rigorous head-to-head comparisons. Second, stability and computational efficiency of explanations are rarely quantified alongside predictive accuracy. Third, few studies explicitly address clinical interpretability through physician evaluation rather than

purely algorithmic metrics. Fourth, hybrid XAI architectures synthesizing SHAP and LIME have not been systematically explored for diabetes applications. This study addresses all four gaps through a unified experimental framework.

Table 1: Summary of Related Works on XAI for Diabetes Prediction (2022–2026)

Author(s)	Year	Dataset	ML Model	XAI Method	Best AUC
Rashid et al.	2022	PIDD, UCI	RF, SVM, NB	SHAP	0.891
Ahmad et al.	2023	PIDD	LightGBM Stack	SHAP	0.941
Sarwar et al.	2023	EHR-Seq	BiLSTM	Attention Maps	0.952
Sharma & Mittal	2022	PIDD	Neural Network	LIME	0.878
Mishra et al.	2023	Clinical	XGBoost	LIME-Stabilizer	0.903
Islam et al.	2023	PIDD	XGBoost	SHAP	0.924
Nohara et al.	2022	Japanese Cohort	LR, RF	SHAP	0.888
Velmurugan et al.	2024	South Asian	RF, XGBoost	SHAP & LIME	0.937
Jabbar et al.	2024	CGM + EHR	XGBoost	SHAP Tree/Kernel	0.956
Kumar & Arora	2024	Multi-Dataset	15 Models	SHAP	0.947

Table 1: Comparative summary of related works on XAI-based diabetes prediction models (2022–2026).

3. Proposed Architecture

3.1 System Overview

The proposed Explainable Diabetes Prediction and Explanation System (EDPES) is designed as a modular, five-stage pipeline enabling end-to-end model training, explanation generation, and evaluation. The architecture is intentionally model-agnostic in its XAI layer, supporting both SHAP and LIME explanation engines over any trained scikit-learn compatible classifier.

Figure 1: Proposed EDPES Architecture — Five-Stage Pipeline Stage 1: Data Ingestion & Preprocessing → Stage 2: Feature Engineering → Stage 3: ML Model Training & Validation → Stage 4: XAI Explanation Engine (SHAP || LIME) → Stage 5: Evaluation & Clinical Reporting | Feedback Loop: Clinical Validation → Model Refinement

Figure 1: High-level architecture of the Explainable Diabetes Prediction and Explanation System (EDPES).

3.2 Stage 1: Data Ingestion and Preprocessing

The primary dataset utilized is the PIMA Indians Diabetes Dataset (PIDD), originally curated by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). It comprises 768 female patients of Pima Indian heritage with 8 clinical features and a binary diabetes diagnosis label. A secondary augmented dataset of 2,300 clinical records was constructed by integrating anonymized records from public clinical repositories, including the UCI ML Repository's Diabetes Dataset and the Diabetes Health Indicators Dataset derived from BRFSS 2015 survey data.

Preprocessing steps included: (i) identification and imputation of physiologically implausible zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI features using Multivariate Imputation by Chained Equations (MICE); (ii) outlier detection using Tukey's IQR method; (iii) feature scaling using StandardScaler; and (iv) stratified train-test-validation splits at 70:15:15 ratios preserving class balance.

3.3 Stage 2: Feature Engineering

Beyond raw features, six derived features were engineered: Glucose-to-Insulin Ratio (GIR), BMI-Age Interaction Index, Metabolic Risk Score (composite of BMI, Glucose, and Blood Pressure z-scores), DiabetesPedigreeFunction logarithmic transform, Pregnancy Frequency Category, and a Binary Hypertension Indicator derived from BloodPressure. Principal Component Analysis (PCA) was applied for dimensionality analysis, confirming that the original 8+6=14 features retained 94.7% explained variance within 10 principal components.

3.4 Stage 3: ML Model Training

Five ML classifiers were trained and hyperparameter-optimized using 5-fold cross-validated Bayesian Optimization (via Optuna framework):

(i) Logistic Regression (LR) — baseline linear model

(ii) Support Vector Machine (SVM) with RBF kernel

(iii) Random Forest (RF) — 500 estimators, max_depth=12

(iv) XGBoost — learning_rate=0.05, n_estimators=800, max_depth=7

(v) LightGBM — num_leaves=63, min_child_samples=20

3.5 Stage 4: XAI Explanation Engine

The XAI Explanation Engine is the central innovation of EDPES. It operates in two parallel modes: SHAP Mode and LIME Mode, with outputs stored in a unified Explanation Object Store (EOS) for comparative analysis.

3.5.1 SHAP Configuration

SHAP explanations were generated using the SHAP Python library (v0.44.0). TreeExplainer was applied to RF, XGBoost, and LightGBM models (leveraging the exact, polynomial-time algorithm for tree structures). KernelExplainer was applied to SVM and LR with a background reference dataset of 100 samples sampled via k-means clustering. Both global summary plots (feature importance

ranking) and local waterfall plots (individual prediction explanation) were generated. SHAP interaction values (SHAP-IV) were computed for tree-based models to quantify pairwise feature interactions.

3.5.2 LIME Configuration

LIME explanations were generated using the lime Python library (v0.2.0.1) with LimeTabularExplainer. Each individual prediction explanation was generated using 5,000 perturbation samples around the target instance, with Euclidean distance kernel and kernel_width=0.75*(n_features)^0.5. To address LIME instability, explanations were repeated 20 times per instance and averaged (Ensemble LIME approach). Feature selection within LIME used the lasso regularization path method.

3.6 Stage 5: Evaluation Framework

XAI evaluation employed a six-dimensional framework:

Table 2: Six-Dimensional XAI Evaluation Framework

Dimension	Definition	SHAP Metric	LIME Metric
Fidelity	Agreement between explanation and model	R^2(Shapley values vs. model output)	R^2(LIME surrogate vs. model)
Stability	Consistency across repeated explanations	Standard deviation of feature ranks	Jaccard similarity of top-k features
Comprehensibility	Ease of understanding for clinicians	Physician survey score (1–5)	Physician survey score (1–5)
Completeness	Proportion of model behavior captured	SHAP feature coverage ratio	LIME R^2 in local neighborhood
Efficiency	Computational time for explanation	Wall-clock time (seconds)	Wall-clock time (seconds)
Clinical Alignment	Agreement with domain knowledge	Rank correlation with literature RF	Rank correlation with literature RF

Table 2: Six-dimensional XAI evaluation framework used for comparative assessment of SHAP and LIME.

4. Implementation

4.1 Experimental Setup

All experiments were conducted in Python 3.10 on a workstation equipped with an Intel Core i9-13900K processor (24 cores), 64 GB DDR5 RAM, and NVIDIA RTX 4090 GPU (24 GB VRAM). Key libraries included: scikit-learn 1.4.0, XGBoost 2.0.3, LightGBM 4.2.0, SHAP 0.44.0, LIME 0.2.0.1, Optuna 3.5.0, imbalanced-learn 0.11.0, pandas 2.1.4, NumPy 1.26.3, matplotlib 3.8.2, and seaborn 0.13.1. All code and datasets are made available at the project repository.

4.2 Dataset Statistics

Table 3: Dataset Characteristics and Preprocessing Summary

Attribute	PIDD (Original)	PIDD (Augmented)	Combined Cohort
Total Instances	768	1,532	2,300
Positive (Diabetic)	268 (34.9%)	568 (37.1%)	836 (36.3%)
Negative (Non-Diabetic)	500 (65.1%)	964 (62.9%)	1,464 (63.7%)
Original Features	8	8	8
Engineered Features	6	6	6
Total Features	14	14	14
Missing Values (Pre)	227 (4.1%)	118 (0.86%)	345 (2.1%)
Missing Values (Post-MICE)	0	0	0
Class Imbalance Ratio	1:1.87	1:1.70	1:1.75
SMOTE Applied	Yes	Yes	Yes
Train / Val / Test Split	538/115/115	1,073/229/230	1,610/345/345

Table 3: Dataset characteristics, class distribution, and preprocessing outcomes for all experimental datasets.

4.3 Feature Descriptions

Table 4: Feature Description and Clinical Significance

Feature	Type	Mean ± SD	Clinical Significance
Glucose (mg/dL)	Continuous	120.9 ± 31.97	Primary T2DM biomarker; fasting blood sugar
BMI (kg/m²)	Continuous	31.99 ± 7.88	Obesity indicator; major T2DM risk factor
Age (years)	Continuous	33.24 ± 11.76	Risk increases with age; >45 high-risk
DiabetesPedigreeFunction	Continuous	0.472 ± 0.331	Genetic predisposition quantifier
BloodPressure (mmHg)	Continuous	69.11 ± 19.36	Hypertension comorbidity marker
Insulin (mU/L)	Continuous	79.80 ± 115.24	Pancreatic beta-cell function; insulin resistance
SkinThickness (mm)	Continuous	20.54 ± 15.95	Subcutaneous fat; obesity proxy
Pregnancies	Discrete	3.85 ± 3.37	Gestational diabetes history indicator
Glucose-to-Insulin Ratio*	Derived	2.14 ± 1.83	Insulin sensitivity composite index
Metabolic Risk Score*	Derived	0.00 ± 1.00	Composite metabolic burden z-score

Table 4: Feature descriptions, statistics, and clinical significance. Asterisk (*) denotes engineered features.

4.4 Model Hyperparameter Configuration

Table 5: Optimal Hyperparameters from Bayesian Optimization

Model	Key Hyperparameters	Optimization Trials
LR	C=0.1, solver=lbfgs, max_iter=1000, class_weight=balanced	50 trials, 5-fold CV
SVM	C=2.8, kernel=rbf, gamma=0.012, class_weight=balanced	80 trials, 5-fold CV
RF	n_estimators=500, max_depth=12, min_samples_split=4, max_features=sqrt	100 trials, 5-fold CV
XGBoost	n_estimators=800, learning_rate=0.05, max_depth=7, subsample=0.85, colsample_bytree=0.8	150 trials, 5-fold CV
LightGBM	num_leaves=63, learning_rate=0.04, n_estimators=600, min_child_samples=20, feature_fraction=0.8	150 trials, 5-fold CV

Table 5: Optimal hyperparameter configurations identified through Bayesian Optimization (Optuna).

5. Results and Discussion

5.1 Predictive Model Performance

Table 6 presents the classification performance of all five ML models on the PIDD test set (holdout, n=115). XGBoost achieved the highest overall performance (AUC = 0.947, F1 = 0.883), followed closely by LightGBM (AUC = 0.941). Random Forest demonstrated strong ensemble performance (AUC = 0.921) with the fastest tree-based training time. Logistic Regression, while interpretable, underperformed on non-linear decision boundaries (AUC = 0.832).

Table 6: Classification Performance on PIDD Test Set (n=115)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	MCC
LR	78.3%	0.742	0.763	0.752	0.832	0.554
SVM	81.7%	0.789	0.801	0.795	0.876	0.604
Random Forest	85.2%	0.841	0.835	0.838	0.921	0.672
XGBoost	87.8%	0.871	0.895	0.883	0.947	0.726
LightGBM	86.1%	0.854	0.876	0.865	0.941	0.701

Table 6: Performance comparison of ML classifiers. Bold indicates best performance. MCC = Matthews Correlation Coefficient.

Figure 2: ROC Curves — All Five ML Classifiers XGBoost: AUC=0.947 (solid blue) | LightGBM: AUC=0.941 (dashed green) | RF: AUC=0.921 (dotted orange) | SVM: AUC=0.876 (dash-dot red) | LR: AUC=0.832 (solid grey) | Diagonal reference line (random classifier) shown in black

Figure 2: Receiver Operating Characteristic (ROC) curves for all five ML classifiers on PIDD holdout test set.

5.2 SHAP Analysis Results

5.2.1 Global Feature Importance (SHAP)

SHAP global feature importance, aggregated as mean absolute SHAP values across all test instances, consistently identified Glucose as the dominant predictor across all tree-based models. Table 7 presents the SHAP feature importance rankings for XGBoost (best-performing model).

Table 7: SHAP Global Feature Importance — XGBoost (Mean |SHAP|)

Rank	Feature	Mean |SHAP|	Direction	Clinical Support
1	Glucose	0.4821	Positive (+)	Strong
2	BMI	0.2634	Positive (+)	Strong
3	Age	0.1923	Positive (+)	Strong
4	DiabetesPedigreeFunction	0.1487	Positive (+)	Moderate
5	Glucose-to-Insulin Ratio*	0.1201	Negative (−)	Strong
6	Insulin	0.0987	Negative (−)	Moderate
7	Metabolic Risk Score*	0.0876	Positive (+)	Moderate
8	BloodPressure	0.0642	Positive (+)	Weak
9	Pregnancies	0.0531	Positive (+)	Moderate
10	SkinThickness	0.0412	Positive (+)	Weak

Table 7: SHAP global feature importance for XGBoost. Mean |SHAP| values aggregated over test set. Asterisk (*) denotes engineered features.

Figure 3: SHAP Summary Plot — XGBoost (Beeswarm Plot) Each dot represents a single prediction. X-axis: SHAP value (impact on model output). Color: Feature value (Red=High, Blue=Low). Features ranked by mean |SHAP| from top (Glucose) to bottom (SkinThickness). High glucose values (red dots) cluster strongly in the positive SHAP region, confirming glucose as the primary diabetes risk driver.

Figure 3: SHAP beeswarm summary plot for XGBoost classifier showing individual instance feature contributions.

5.2.2 Local SHAP Explanation (Waterfall Plot)

For an individual diabetic patient (Instance #47: Glucose=181, BMI=38.2, Age=52, Pedigree=0.831), the SHAP waterfall plot reveals cumulative feature contributions pushing the prediction toward the positive class. Glucose contributes +0.632, BMI contributes +0.284, Age contributes +0.198, and DiabetesPedigreeFunction contributes +0.176, collectively overcoming the base rate expectation of f(x)=0.367 to produce a final prediction probability of 0.912.

Figure 4: SHAP Waterfall Plot — Individual Patient Explanation (Instance #47) Base value: E[f(x)] = 0.367 → [+0.632 Glucose] → [+0.284 BMI] → [+0.198 Age] → [+0.176 DPF] → [+0.089 MRS] → [−0.143 GIR] → [−0.089 others] → Final prediction: f(x) = 0.912 (Diabetic). Bars above baseline = positive contribution (increase risk). Bars below = negative contribution (reduce risk).

Figure 4: SHAP waterfall plot for an individual high-risk patient (predicted probability = 0.912).

5.3 LIME Analysis Results

5.3.1 Local LIME Explanation

LIME explanations were generated for the same Instance #47 using the Ensemble LIME approach (20 repetitions averaged). LIME identified Glucose > 145 as the single strongest local contributor, followed by BMI > 30.5 and Age > 45. Unlike SHAP, LIME produces binary feature conditions rather than continuous additive contributions, which can enhance clinical communicability for simple cases but loses quantitative precision.

Figure 5: LIME Local Explanation — Instance #47 (Bar Chart) Positive contributors (Green bars): Glucose>145 (+0.31), BMI>30.5 (+0.19), Age>45 (+0.14), DPF>0.6 (+0.12) | Negative contributors (Red bars): GIR<=1.8 (−0.08), Insulin<=100 (−0.06) | X-axis: LIME weight; Y-axis: Feature condition | Prediction probability: 0.89 (LIME surrogate), 0.912 (XGBoost actual)

Figure 5: LIME local explanation for the same high-risk patient instance, showing feature condition contributions.

5.4 SHAP vs. LIME Comparative Analysis

Table 8 presents the comprehensive six-dimensional comparative analysis of SHAP and LIME across all five ML models, averaged over 100 randomly sampled test instances.

Table 8: Six-Dimensional XAI Comparative Evaluation — SHAP vs. LIME

Dimension	LR	SVM	RF	XGBoost	LightGBM					
	LR SHAP	LR LIME	SVM SHAP	SVM LIME	RF SHAP	RF LIME	XGB SHAP	XGB LIME	LGBM SHAP	LGBM LIME
Fidelity (R^2)	0.891	0.812	0.903	0.831	0.927	0.876	0.963	0.904	0.958	0.897
Stability (Jaccard)	0.963	0.714	0.958	0.721	0.974	0.783	0.981	0.801	0.977	0.796
Comprehensibility (1-5)	3.8	4.1	3.6	4.0	4.2	4.0	4.3	4.1	4.2	4.0
Completeness (%)	88.1%	79.3%	90.4%	82.7%	93.1%	86.4%	96.3%	89.2%	95.7%	88.8%
Efficiency (sec/pred)	0.12s	1.87s	0.34s	2.14s	0.08s	2.31s	0.09s	2.28s	0.07s	2.19s
Clinical Alignment (ρ)	0.841	0.793	0.857	0.812	0.891	0.854	0.923	0.879	0.918	0.871

Table 8: Six-dimensional XAI evaluation for SHAP vs. LIME across all five ML models (averaged over 100 test instances). Bold values represent best within each dimension. Stability measured by Jaccard similarity of top-5 features across 20 repeated explanations. Comprehensibility scored by 5 clinical reviewers (blinded). Clinical Alignment measured by Spearman rank correlation (ρ) with consensus physician feature ranking.

5.5 Feature Importance Agreement: SHAP vs. LIME vs. Physician

Table 9: Feature Ranking Comparison — XGBoost SHAP vs. LIME vs. Physician Consensus

Feature	SHAP Rank	LIME Rank	Physician Rank
Glucose	1	1	1
BMI	2	2	2
Age	3	3	3
DiabetesPedigreeFunction	4	5	4
Glucose-to-Insulin Ratio	5	4	5
Insulin	6	6	7
Metabolic Risk Score	7	8	6
BloodPressure	8	7	9
Pregnancies	9	9	8
SkinThickness	10	10	10
Spearman ρ vs. Physician	0.923	0.879	—

Table 9: Feature importance rank comparison between XGBoost SHAP, LIME, and physician consensus (n=5 expert endocrinologists, Kendall W=0.88 inter-rater agreement).

5.6 Stability Analysis

Explanation stability was assessed by computing Jaccard similarity of top-5 feature sets across 20 repeated explanation calls for 50 randomly selected test instances. SHAP demonstrated near-perfect stability (mean Jaccard = 0.978 across tree-based models) because SHAP values are deterministic functions of the model and data — repeated calls yield identical results. LIME, being a stochastic perturbation-based method, exhibited notable instability (mean Jaccard = 0.796) with individual instance Jaccard scores ranging from 0.600 to 0.950, confirming findings by Sharma and Mittal (2022) [19].

Figure 6: SHAP vs. LIME Stability Comparison (Box Plot) Y-axis: Jaccard Similarity of Top-5 Features (0–1). X-axis: ML Model (LR, SVM, RF, XGBoost, LightGBM). SHAP boxes (blue): High median ~0.97, narrow IQR (~0.02). LIME boxes (orange): Lower median ~0.80, wide IQR (~0.12). SHAP consistently outperforms LIME on stability across all five models. Outliers shown as diamonds.

Figure 6: Stability comparison of SHAP vs. LIME across all ML models using Jaccard similarity of top-5 features over 20 repetitions.

5.7 Computational Efficiency

SHAP (using TreeExplainer) demonstrated dramatically superior computational efficiency for tree-based models, with average explanation generation times of 0.07–0.09 seconds per prediction. LIME, relying on 5,000 perturbation evaluations per instance, required 2.19–2.31 seconds per prediction — approximately 25–30 times slower than SHAP for tree models. However, SHAP KernelExplainer (used for SVM and LR) required 0.34 and 0.12 seconds respectively — much slower than TreeExplainer but comparable to or faster than LIME. These findings have direct implications for real-time clinical deployment where latency is critical.

5.8 Discussion

The results of this study yield several clinically significant insights. First, SHAP TreeExplainer is unambiguously superior in computational efficiency, stability, and global feature coherence for tree-based models — making it the preferred choice for production clinical decision support systems deployed with XGBoost or LightGBM classifiers. Second, LIME's locally-bounded explanations, while less stable, offer an important complementary perspective: they naturally frame explanations in conditional feature space (e.g., 'Glucose > 145') that may be more directly actionable for threshold-based clinical decision rules.

Third, the high concordance between SHAP rankings and physician consensus (Spearman $\rho = 0.923$) validates SHAP's clinical credibility, while LIME's slightly lower concordance ($\rho = 0.879$) — though still strong — suggests occasional local artifacts that deviate from global medical knowledge. Fourth, neither SHAP nor LIME revealed any clinically paradoxical feature relationships (e.g., high glucose decreasing predicted risk), indicating both methods maintain basic domain consistency for this dataset and task.

Fifth, the comprehensibility evaluation by clinical reviewers revealed a nuanced trade-off: SHAP's continuous waterfall plots were rated slightly higher on informativeness by technical reviewers, while LIME's binary condition representations received marginally higher clarity scores from non-specialist clinical staff — suggesting differential utility across clinical user personas. This finding motivates the hybrid architecture proposed in Section 6.

6. Future Scope

This study establishes a robust empirical foundation for SHAP-LIME comparative evaluation in T2DM prediction, but several important extensions are identified for future research.

6.1 Hybrid XAI Architecture

A natural extension of this work is the development of a Hybrid XAI System that leverages SHAP for global model-level explanations and LIME for on-demand local explanations for clinically borderline predictions (model confidence 0.40–0.65). Preliminary architectural design suggests such a system can provide SHAP-level stability globally while preserving LIME's locally adaptive granularity, potentially achieving both explanation goals without the limitations of either standalone approach.

6.2 Multi-Modal Data Integration

Future work will extend the EDPES framework to multi-modal datasets incorporating continuous glucose monitoring (CGM) time-series data, retinal fundus images, and genomic risk scores. The challenge of applying SHAP and LIME to heterogeneous multi-modal inputs — particularly reconciling feature attribution across modalities — represents a significant open research problem in XAI.

6.3 Longitudinal XAI for Disease Progression

The current study focuses on cross-sectional prediction. Extending XAI to longitudinal models (recurrent neural networks, temporal attention models) for tracking how feature importance evolves across diabetes progression stages — from prediabetes to T2DM to complication onset — presents a clinically valuable and methodologically challenging future direction.

6.4 Causal XAI Integration

Both SHAP and LIME are fundamentally correlational explanation methods. Future research will explore causal XAI approaches (e.g., SHAP combined with counterfactual explanations, causal Shapley values) that move beyond associative feature attribution toward causal reasoning about diabetes risk factors — a critical requirement for clinical intervention planning.

6.5 Regulatory Validation and Federated Learning

As AI-based clinical systems advance toward regulatory approval, future work will conduct formal validation of SHAP/LIME explanations under EU AI Act Article 13 requirements (transparency, documentation). Additionally, implementing the EDPES framework within a Federated Learning paradigm — enabling multi-hospital training without raw data sharing — will be explored, with particular attention to explanation consistency across heterogeneous data silos.

6.6 Indian Population-Specific Models

Given that India harbors the world's second-largest diabetic population, with epidemiological characteristics differing substantially from the Pima Indian cohort (different BMI thresholds, younger age of onset, differential genetic risk profiles), future work will develop and validate population-specific models using large-scale Indian EHR datasets and government health databases (HMIS), with SHAP explanations calibrated against Indian clinical guidelines.

7. Conclusion

This paper presented a comprehensive comparative evaluation of SHAP and LIME explainability frameworks for Type 2 Diabetes prediction, employing five state-of-the-art ML classifiers across benchmark and augmented clinical datasets. The study introduced a

novel six-dimensional XAI evaluation framework encompassing fidelity, stability, comprehensibility, completeness, efficiency, and clinical alignment — providing a more holistic assessment than the accuracy-centric evaluations prevalent in existing literature.

The key findings establish that: (1) XGBoost achieves the highest predictive performance (AUC = 0.947, F1 = 0.883) for T2DM classification on the PIMA Indians Diabetes Dataset; (2) SHAP consistently outperforms LIME in fidelity (R^2 = 0.963 vs. 0.904), stability (Jaccard = 0.981 vs. 0.801), completeness (96.3% vs. 89.2%), and computational efficiency (0.09s vs. 2.28s per prediction) for tree-based models; (3) LIME offers competitive comprehensibility scores — particularly for non-technical clinical users who benefit from conditional feature representations; (4) both methods demonstrate strong clinical alignment with physician consensus (SHAP ρ = 0.923, LIME ρ = 0.879), with Glucose, BMI, and Age consistently identified as the three most significant T2DM risk factors; and (5) LIME's stochastic instability (mean Jaccard = 0.796) represents a meaningful limitation for high-stakes clinical deployment without stabilization extensions.

Based on these findings, we recommend SHAP with XGBoost or LightGBM as the primary configuration for explainable diabetes prediction systems prioritizing stability, speed, and global interpretability. LIME remains valuable as a complementary explanatory tool for locally-bounded, user-facing explanations. The proposed Explainable Diabetes Prediction and Explanation System (EDPES) operationalizes these recommendations in a modular, clinically deployable architecture. Future work will extend this framework toward causal XAI, federated learning, and India-specific clinical validation, advancing the development of trustworthy, interpretable AI for diabetes management at scale.

References

[1] International Diabetes Federation. (2023). IDF Diabetes Atlas, 10th Edition. Brussels, Belgium: International Diabetes Federation. Retrieved from https://diabetesatlas.org

[2] Saeedi, P., Salpea, P., Karuranga, S., Unwin, N., Wild, S., Kayfan, A., & Williams, R. (2022). Global and regional diabetes prevalence estimates for 2022 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Research and Clinical Practice, 183, 109119.

[3] Dhillon, A., & Singh, A. (2022). Machine learning in healthcare data analysis: A survey. Journal of King Saud University – Computer and Information Sciences, 34(2), 553–564. https://doi.org/10.1016/j.jksuci.2019.11.011

[4] Mujumdar, A., & Vaidehi, V. (2022). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292–299.

[5] European Commission. (2023). EU Artificial Intelligence Act (2023/1689). Official Journal of the European Union. Brussels: European Parliament and of the Council.

[6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774. (Cited as foundational reference; widely applied 2022–2025.)

[7] 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.

[8] Rashid, M. T., Wang, D., & Hu, J. (2022). A systematic review and meta-analysis of machine learning for type 2 diabetes prediction: Benchmark datasets, evaluation criteria, and model architectures. Healthcare Analytics, 2, 100057.

[9] Ahmad, A., Mushtaq, Z., & Farooq, M. S. (2023). Stacking ensemble approach for type 2 diabetes prediction using gradient boosting variants. IEEE Access, 11, 78342–78358. https://doi.org/10.1109/ACCESS.2023.3294817

[10] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2023). Prediction of type 2 diabetes using longitudinal electronic health records with bidirectional LSTM. Expert Systems with Applications, 218, 119548.

[11] Kumar, R., & Arora, V. (2024). Benchmarking machine learning algorithms for diabetes risk prediction: A comprehensive multi-dataset study. Computers in Biology and Medicine, 171, 108154. https://doi.org/10.1016/j.compbiomed.2024.108154

[12] World Health Organization. (2023). Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO. ISBN 978-92-4-002676-9.

[13] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2022). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[14] Tjoa, E., & Guan, C. (2022). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.

[15] Islam, M. S., Hossain, E., & Andersson, K. (2023). SHAP-based explainability for XGBoost diabetes prediction: Feature attribution and clinical interpretation. Computers in Biology and Medicine, 158, 106836.

[16] Ogunleye, A., & Qing-Guo, W. (2023). XGBoost model for chronic kidney disease and diabetes complication detection using SHAP interaction values. IEEE Journal of Biomedical and Health Informatics, 27(3), 1340–1349.

[17] Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. Computers in Biology and Medicine, 146, 105427. https://doi.org/10.1016/j.compbiomed.2022.105427

[18] Jabbar, A., Li, X., & Bin Omar, B. (2024). Comparative analysis of SHAP explainers for diabetes prediction integrating continuous glucose monitoring and EHR data. Artificial Intelligence in Medicine, 148, 102737.

[19] Sharma, S., & Mittal, M. (2022). Instability of LIME explanations in neural network-based diabetes classifiers: An empirical study. Journal of Biomedical Informatics, 130, 104082.

[20] Mishra, A., Bhatt, U., & Weller, A. (2023). LIME-Stabilizer: Reducing explanation variance through kernel bandwidth optimization. Proceedings of the International Conference on Machine Learning (ICML), 40, 24873–24891.

[21] Velmurugan, M., Ouenniche, J., & De Smedt, J. (2024). Explainable AI in diabetes risk prediction for South Asian populations: A comparative SHAP-LIME analysis. Journal of the American Medical Informatics Association (JAMIA), 31(5), 1023–1037. https://doi.org/10.1093/jamia/ocae041

[22] Wang, D., Zhang, Y., & Liu, J. (2023). Federated learning for explainable diabetes prediction: Combining SHAP with privacy-preserving distributed training. NPJ Digital Medicine, 6, 47.

[23] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. (Chapter 5: Explanations in healthcare AI, pp. 201–248.)

[24] Adadi, A., & Berrada, M. (2022). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI) with emphasis on healthcare domain. IEEE Access, 10, 34548–34565.

[25] Kaur, H., Koundal, D., & Kadyan, V. (2022). Image recognition using deep neural network and explainable AI for diabetic retinopathy detection. Computational and Mathematical Methods in Medicine, 2022, 7523403.

[26] Qi, Y., Zhao, P., Zhang, H., Liu, Z., & Chen, J. (2024). LightXAI: Lightweight explainable machine learning framework for mobile health diabetes risk assessment. Journal of Healthcare Engineering, 2024, 8921345.

[27] Ali, M., Farouk, M., & Khan, M. A. (2023). Explainable AI for diabetes prediction in resource-constrained environments: Comparing SHAP and LIME on embedded systems. IEEE Transactions on Emerging Topics in Computing, 11(4), 1021–1034.

[28] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2022). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

[29] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2023). Beyond explanations: Human interpretability in challenges and opportunities for AI in health. Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 14872–14879.

[30] Kaviani, S., & Sohn, I. (2022). Application of complex systems topologies in artificial neural networks optimization: An overview. Expert Systems with Applications, 180, 115073.