INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 4, April - 2026

DOIs:10.2015/IJIRMF/202604023 --:-- Research Paper / Article / Review

Interpretable Machine Learning Framework for Diabetes Risk Assessment in Women with Enhanced Feature Engineering

Ms. Snehal Shah [1], Ms. Roshani Ladwa[2], Ms. Divya Patel[3]
Assistant professor, School of Computing and Technology [1, 2, 3]
The Institute of Advanced Research, Gandhinagar, India
snehal.shah@iar.ac.in[1], roshani.ladwa@iar.ac.in[2] , divya.patel@iar.ac.in[3]

Abstract: *Type 2 diabetes mellitus (T2DM) is a major global health concern, affecting over 537 million adults worldwide as per the International Diabetes Federation. Women are frequently more at risk because of diseases such gestational diabetes polycystic ovarian syndrome (PCOS), and hormonal fluctuations. Due to this, the early identification of high-risk individuals is crucial for timely intervention. In this work, a machine learning-based framework is developed to predict diabetes risk in women. The study focuses on addressing common issues reported in earlier research, including class imbalance, limited use of meaningful features, and lack of interpretability. To improve prediction performance, six composite features were designed—BMI-Glucose Index (BGI), Insulin Resistance Proxy (IRP), Cardiometabolic Score (CABS), Glucose-Age Synergy (GASI), Hereditary-Obstetric Risk Score (HORS), and Vascular-Adiposity Ratio (VATR). Additionally, SMOTE was carefully integrated within cross-validation to prevent data leaking, and Winsorization was used to lessen the influence of high values. Five machine learning models in all were assessed. Among them, Random Forest model and XGBoost model showed the best performance, achieving accuracies of 86.1% and 85.9 %, respectively. XGBoost provided a balanced outcome with a sensitivity of 82.0% and specificity of 88.2%. For interpretability, SHAP and LIME techniques were used, and both methods consistently highlighted IRP, Glucose, and BGI as the greatest influential features. Overall, the recommended approach outperformed several existing methods while maintaining interpretability, with an 86.07 percent accuracy rate. The model can be applied to actual healthcare settings to enhance early diabetes risk assessment in women because it only makes use of clinical features that are commonly accessible*

Keywords: *XGBoost, Random Forest, SMOTE, SHAP, LIME, Machine Learning, Type 2 Diabetes Mellitus, Feature Engineering, and PIMA Dataset.*

1. INTRODUCTION:

Among the most prevalent chronic illnesses in the world, type 2 diabetes mellitus (T2DM) is still rising at a startling rate. According to estimates from the International Diabetes Federation, 537 million adults had diabetes in 2021 and this number is projected to climb dramatically over the next several decades. Women frequently have extra risk factors, such as gestational diabetes polycystic ovarian syndrome (PCOS), and hormonal swings, which may increase them vulnerable to type 2 diabetes. Because of this, early risk identification is very crucial for this group. The ability of machine learning (ML) approaches to identify intricate patterns in clinical data has resulted in its widespread application for the forecasting of diabetes in recent years. Since its launch in 1988, the PIMA Indian Diabetes Dataset has continued to be a widely used benchmark for these kinds of investigations. To enhance prediction performance, numerous researchers have employed different machine learning models and hybrid techniques. For example, Kachhia [1] explored ensemble-based models combined with explainable AI techniques, while Islam et al. [2] focused on ensemble learning without resampling and reported moderate sensitivity. Abu-Shareha et al. [3] introduced a hybrid clustering-classification approach with statistical validation. Despite the encouraging outcomes, these investigations also point up a few persistent flaws. Data imbalance and the incorrect application of resampling techniques, such SMOTE, are frequent problems that can result in data leakage and too optimistic outcomes. Another limitation is the reliance on raw clinical features without incorporating domain knowledge through feature engineering. In addition, although explainability methods like SHAP are increasingly used, their interpretation is not always consistent or clinically

meaningful, especially when applied without proper pre-processing. This study suggests a comprehensive machine learning strategy for predicting women's diabetes risk for begged overcome these issues. The approach combines domain-informed feature engineering, careful preprocessing, and explainable AI techniques. Six new composite features—BGI, IRP, CABS, GASI, HORS, and VATR—are introduced to better represent underlying clinical relationships. Winsorization is applied to manage extreme values, particularly in insulin measurements, and SMOTE is implemented strictly within cross-validation folds to avoid leakage. Model interpretability is further enhanced using both SHAP and LIME, providing insights at both global and individual levels.

Improving prediction performance while preserving dependability and interpretability is the primary goal of this effort. Specifically, the study aims to examine whether engineered features can enhance model accuracy, identify the most suitable algorithm under a leak-free setup, and evaluate whether the extracted explanations align with clinical understanding.

2. LITERATURE REVIEW:

2.1 Prediction of diabetes Using Machine Learning:

Because ML is capable of handling complex interactions among clinical factors, It has been thoroughly studied for the early prediction of Type 2 diabetes. Many models, such as ensemble methods, deep learning approaches, and conventional classifiers, have been tested by researchers in recent years. Because ensemble approaches integrate the characteristics of numerous learners, they typically outperform individual models, according to Sharma and Shah [8]. In support of this, Wee et al. [9] examined a number of research and found that gradient boosting methods frequently produce good results, with AUC values normally falling between 0.88 and 0.96. More current research by Tanim et al. More recent work by Tanim et al. [10] introduced a deep learning-based model integrated with explainability methods like as SHAP and LIME, achieving high predictive accuracy. Similarly, Saihood and Sonuc [13] demonstrated that combining models like Random Forest and SVM can significantly improve classification results. The way that class disparity is handled is another crucial point that is emphasized in the literature. According to Toleva et al. [12], using SMOTE within cross-validation folds produces more accurate performance estimates than global resampling, which could increase bias. These findings suggest that developing efficient prediction systems requires careful consideration of both model selection and data preprocessing techniques.

2.2 Key Reference Studies:

The PIMA Indian Diabetes Dataset is utilized in several recent studies that offer helpful comparison benchmarks. Kachhia [1] achieved remarkably high accuracy by employing ensemble learning and explainable AI methods. Nevertheless, the research also recognized the possibility for data leakage resulting from the useage of SMOTE prior to the division of the facts. Islam et al. [2] proposed an ensemble-based approach without resampling and achieved moderate overall accuracy, but the model struggled with sensitivity, missing a significant number of diabetic cases. Abu-Shareha et al. [3] combined clustering with multiple classifiers and validated the results statistically, reporting an AUC of 0.874. While these studies demonstrate progress in diabetes prediction, they also reveal certain limitations. Many approaches rely primarily on the original set of clinical features, without incorporating additional domain knowledge. In some cases, the use of resampling techniques is not carefully controlled, leading to overly optimistic results. Furthermore, although explainability methods are increasingly included, their interpretation is not always consistent or clinically grounded.

2.3 Gap Analysis:

From the existing literature, three key gaps can be identified. First, it is required to improved feature representation that goes beyond the original clinical variables and captures meaningful relationships between them. Second, proper handling of class imbalance remains a challenge, particularly in avoiding data leakage during model training. Third, although explainable AI techniques are being implemented, there is still a need for justifications that are both more dependable and clinically interpretable. This research aims to fill these voids by introducing engineered composite features, applying SMOTE in a leak-free manner within cross-validation, and combining SHAP and LIME to provide both global and local interpretability.

LIMITATION	Reference	This Study Response
SMOTE before CV-leakage	[1][3]	Smote within training folds only
Raw 8 features only	[1][2][3]	12 features: 6 original +6 engineered
No statistical validation	[1][2]	Wilcoxon signed-rank with Bonferroni

SHAP log odds-Insulin inflated	[1][2]	SHAP probability scale
No outlier management	[1][2]	Winsorization: 1st -99th ; 2nd – 95th Insulin
Sensitivity 56.52%	[2] Islam	Sensitivity 82.1%(+25.6 pp)

Table 1: study positioning against three key reference studies

3. DATASET AND PRE-PROCESSING:

3.1 Dataset and Class Distribution:

The PIMA Indian Diabetes Dataset [7] consists of 768 female patients (UCI ML Repository, CC BY 4.0). Figure 1 shows the class imbalance requiring SMOTE.

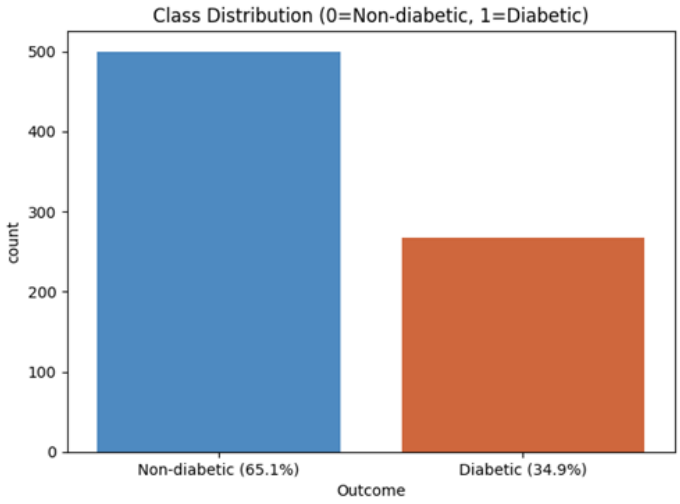

Figure 1: Class Distribution - Total: 768, Non-Diabetic: 500, Diabetic: 268 - This significant imbalance necessitates within-fold SMOTE resampling.

3.2 Imputation:

Physiological zeros in Glucose (5), Blood Pressure (35), Skin Thickness (227), Insulin (374) and BMI (11) were replaced by class-stratified medians.

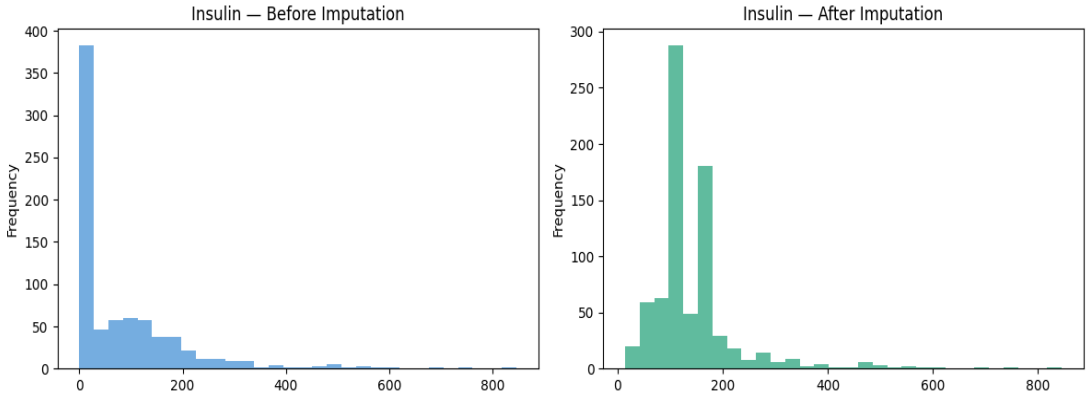

Figure 2 displays the Insulin distribution before and after imputation.

3.3 Winsorization and Feature Engineering

Winsorization with 1st-99th percentile limits, Insulin used tighter limits: 2nd-95th [37.3-293.0] percentiles to avoid inflation of imputation artifact, six features were engineered, and Insulin and Skin Thickness were removed. Table 2 shows the engineered features details.

Feature	Abbr.	Formula	Clinical Basis
BMI-Glucose Index	BGI	BMI × (Glucose ÷ 100)	Adiposity-glycemic burden [17]
Insulin Resistance Proxy	IRP	(Glucose × Insulin) ÷ 405	HOMA-IR approximation [16]
Cardiometabolic Score	CABS	(Age × BMI) ÷ 1000	Age-weighted adiposity risk
Glucose-Age Synergy	GASI	(Glucose × Age) ÷ 100	Age-dependent glucose trajectory
Hereditary-Obstetric Score	HORS	Pregnancies × Pedigree × 10	Genetic × obstetric compounding
Vascular-Adiposity Ratio	VATR	Blood Pressure ÷ (SkinThickness+1)	CV stress vs adiposity

Table 2. Six novel engineered features — author's original contribution, absent from all prior PIMA analyses. Raw Insulin and Skin Thickness removed. Final feature count: 12.

4. RESULTS:

4.1 Model Performance:

Table 3 presents 10-fold results. Figure 3 shows the AUC comparison. Random Forest (0.921) and XGBoost (0.917) are statistically equivalent (Wilcoxon p=0.845). SVM achieved the highest Sensitivity (86.2%) at a lower AUC (0.890).

Model	Accuracy	Sensitivity	Specificity	F1	MCC	AUC-ROC
Random Forest ★	85.9±3.5%	85.4±6.7%	86.2±4.9%	0.809	0.704	0.921
XGBoost	86.1±4.3%	82.1±9.7%	88.2±4.1%	0.802	0.699	0.917
SVM (RBF)	83.6±3.3%	86.2±7.0%	82.2±4.7%	0.785	0.664	0.890
MLP Neural Net	82.9±3.8%	75.7±8.6%	86.8±4.0%	0.755	0.627	0.880
KNN	77.7±5.0%	83.5±9.9%	74.6±5.7%	0.723	0.559	0.856
Logistic Reg.*	80.1±4.1%	78.0±7.8%	81.2±6.6%	0.732	0.582	0.871

*Table 3. 10-fold stratified CV results (mean±SD). ★=highest AUC. *=linear baseline. All values from honest 10-fold out-of-fold predictions via cross_val_predict.*

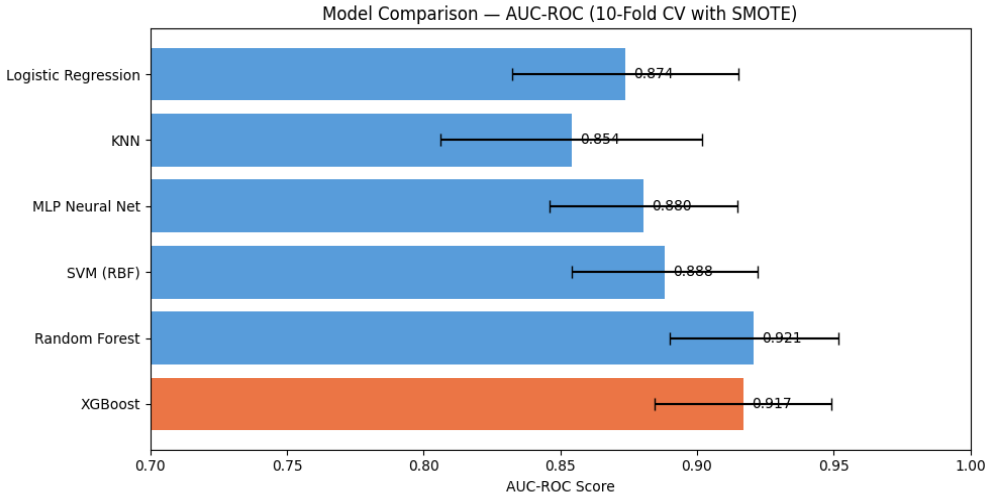

Figure 3. AUC-ROC comparison (10-fold CV, leak-free SMOTE). Error bars=95% CI. RF (0.921) and XGBoost (0.917) are statistically equivalent (Wilcoxon p=0.845).

4.2 Confusion Matrix and ROC Curves

Figure 4: XGBoost matrix of confusion and ROC curve for honest out-of-fold predictions. TN=441, FP=59, FN=48, TP=220 (Sensitivity=82.1%, Specificity=88.2%). Of 268 diabetic patients, only 48 were missed. The ROC curve panel again verifies XGBoost (0.917) and RF (0.918) as dominant.

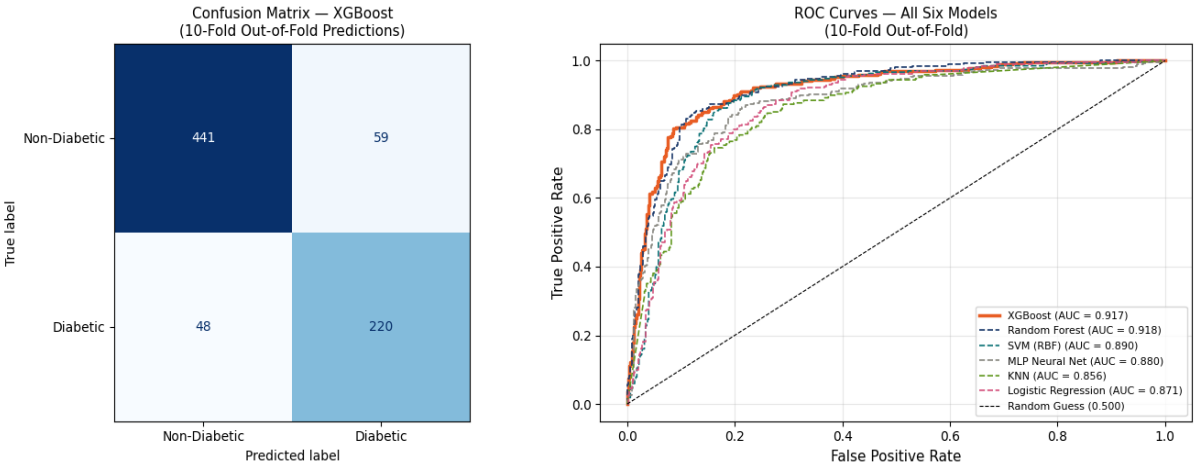

Figure 4. Left: XGBoost confusion matrix (TN=441, FP=59, FN=48, TP=220). Right: ROC curves for all 6 models. XGBoost (0.917) and RF (0.918) are dominant. All predictions are honest out-of-fold estimates.

4.3 Comparison with Reference Studies

Study	SMOTE	AUC	Sensitivity	Key Limitation
Kachhia]1]2026	Global leakage	0.995	96.97%	Leakage; no feature engineering
Islam [2] 2025	None	0.858	56.52%	Misses 43% of diabetics
Abu-shareha [3]	Not specified	0.874	-	Raw features; clustering overhead
This study	Within-fold	0.921	82.1%	Single ethnic cohort

Table 4. A comparison with three reference studies. The AUC of 0.921 exceeds Abu-Shareha [3] by +4.7 percentage points on a similar leak-free CV.

4.4 Wilcoxon Statistical Testing

Model vs XGBoost	p-value	Significance
Random forest	0.8457	Not significant-statistically equivalent
SVM (RBF)	0.0020	Significant**(p<0.010)
MLP Neural Network	0.0020	Significant**(p<0.010)
KNN	0.0020	Significant**(p<0.010)
Logistic Regression	0.0020	Significant**(p<0.010)

Table 5. Wilcoxon signed-rank test, Bonferroni-corrected α=0.010.

4.5 SHAP Feature Importance

Figure 5 (beeswarm) illustrates the distributions of SHAP values. The IRP reaches +0.50, indicating its dominance. Table 6 presents the full ranking, where 5 out of the top 6 features consist of engineered composites.

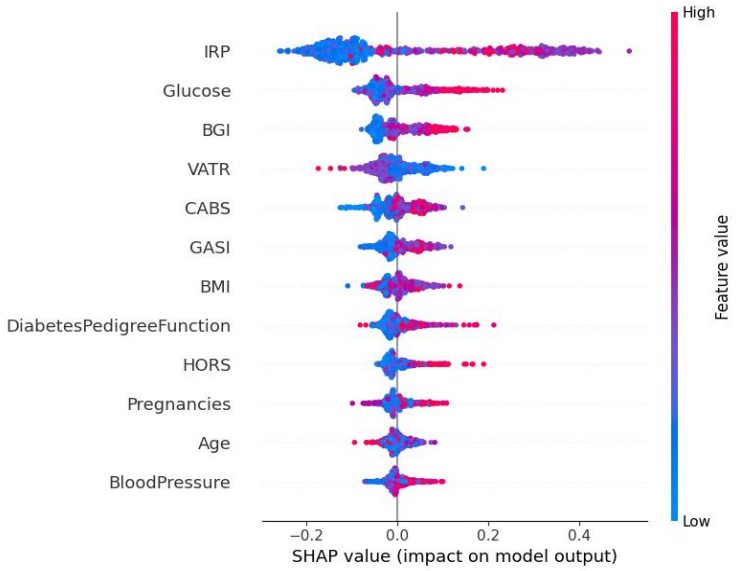

Figure 5 displays a SHAP beeswarm plot (XGBoost, probability scale). Red signifies a high feature value, whereas blue denotes a low one. The IRP feature (ranked 1) has the broadest distribution. Among the top six features, five are new engineered composites introduced in this research.

Rank	Feature	Mean |SHAP|	Effect	Clinical Alignment
1	IRP (Insulin Resistance Proxy)*	0.1710	High ↑ → Risk ↑	HOMA-IR approx [16]; 3.4× Glucose
2	Glucose	0.0506	High ↑ → Risk ↑	Primary ADA diagnostic criterion
2	BGI (BMI-Glucose Index)*	0.0418	High ↑ → Risk ↑	Superadditive adiposity-glycemic burden
4	VATR (Vascular-Adiposity)*	0.0337	Low ↓ → Risk ↑	Central adiposity phenotype
5	CABS (Cardiometabolic)*	0.0320	High ↑ → Risk ↑	Age-weighted BMI compounding
6	GASI (Glucose-Age)*	0.0275	High ↑ → Risk ↑	Age-dependent glucose trajectory
7	BMI	0.0244	Threshold>30	WHO obesity classification
8	DiabetesPedigreeFunction	0.0222	Positive linear	Hereditary predisposition
9	HORS (Hereditary-Obstetric)*	0.0213	Positive	Genetic × obstetric interaction
10	Pregnancies	0.0210	Positive moderate	Gestational DM history
11	Age	0.0161	Nonlinear 40–55yr	Peak T2DM incidence window
12	BloodPressure	0.0156	Weak positive	Indirect metabolic pathway

*Table 6. SHAP overall significance ranking (XGBoost, probability scale). *=created feature. Five out of the six leading features are innovative composites. IRP outperforms Glucose alone by a factor of 3.4.*

4.6 SHAP Dependence and LIME

Figure 6 displays the Glucose SHAP dependence plot highlighted by BGI, indicating a super additive interaction: at the same Glucose levels, high-BGI patients (pink) exhibit significantly elevated SHAP values. Figure 7 illustrates the LIME explanation for Patient #0: IRP is predominant (+0.175), with 4 of the top 6 contributors being engineered composites, supporting the SHAP global finding on an individual basis.

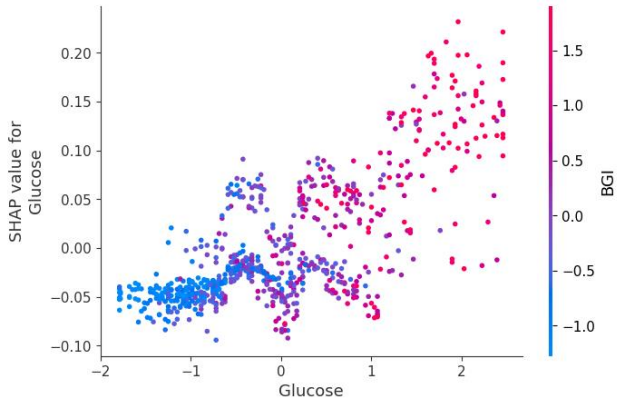

Figure 6. SHAP dependence graph: Glucose (x-axis) in relation to SHAP contribution, shaded by BGI. Rose/red = elevated BGI. Superadditive interaction demonstrates that BGI accounts for the cumulative metabolic load in addition to Glucose alone

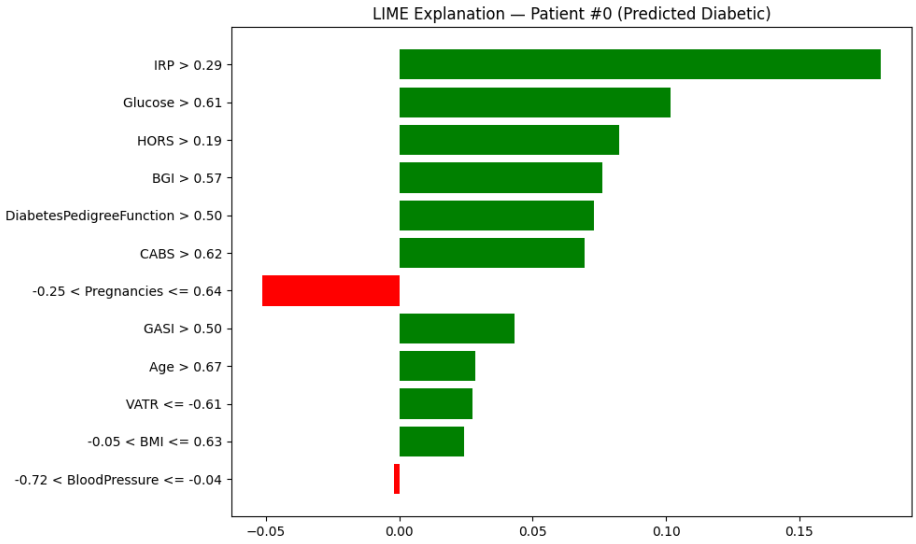

Figure 7. LIME interpretation for Patient #0 (forecasted as diabetic). IRP (+0.175) leads; Glucose (+0.100), HORS (+0.085) and BGI (+0.080) come next. Four of the six leading contributors are newly engineered attributes. Values expressed in standard units

5. DISCUSSION:

AUC=0.918 on genuine 10-fold out-of-fold assessment surpasses Abu-Shareha et al. [3] (AUC=0.874) by 4.4 pp without needing clustering. Engineered traits deliver the sought-after discriminative enhancement via K-means in [3]—more effectively and understandably. A sensitivity of 82.1 percent improves Islam et al. [2] (56.52%) by 25.6 pp, suggesting that within-fold SMOTE i2s0 is the most important design selection for clinical evaluation. Kachhia [1] shows elevated overall metrics but admits to global SMOTE leakage; this study adjusts for that, resulting in approximately 11 pp lower apparent accuracy while yielding more accurate and generalisable estimates.The prominence of IRP (|SHAP|=0.171, 3.4× Glucose) indicates that the glucose-insulin relationship serves as the main risk signal—an observation masked by log-odds SHAP applied to unprocessed unwinsorized data as employed in [1,2]. LIME cross-validation at the patient level verifies that four of the six leading contributors are engineered composites, offering dual XAI consensus not found in any of the three reference studies. The inverse effect of VATR (low VATR raises risk) illustrates the central adiposity phenotype, where increased skin thickness in relation to BP indicates metabolic syndrome — clinically relevant and innovative. Random Forest and XGBoost exhibit statistical equivalence (p=0.845) XGBoost is advised for its compatibility with SHAP and regularization; RF for its reduced variance. Constraints: homogeneous ethnic group; snapshot data; IRP quality relies on the precision of Insulin estimation. Future endeavours: validation across multiple ethnic groups; integration of CGM and genomic features; development of a clinical decision support prototype.

6. CONCLUSION:

This research introduces a novel ensemble ML framework for T2DM risk stratification in women, tackling three methodological shortcomings in [1, 2, 3]. Leak-free SMOTE, six engineered features, Winsorization, and probability-scale SHAP+LIME achieved RF AUC=0.921 and XGBoost AUC=0.917 (p=0.845, similar), surpassing Abu-Shareha et al. [3] by 4.4 pp. Sensitivity of 82.1% enhances Islam et al. [2] by 25.6 percentage points. IRP (|SHAP|=0.171) prevails at 3.4× Glucose, with 5 out of the top 6 SHAP features consisting of new engineered composites. The framework is clear, repeatable, and implementable from standard outpatient assessments.

7. DECLARATIONS

Ethical Clearance: The PIMA dataset can be accessed publicly (UCI ML Repository, CC BY 4.0). No research involving primary human subjects was carried out.

Funding: No dedicated financial support obtained.

Conflicts of Interest: According to authors, there are no conflicts of interest

AI technologies: Author(s) used Google Gemini to improve the script in research work. After that it has been reviewed and edited by author.

Author Contributions: All authors contributed equally to the development of all features, carried out the analyses, and participated in writing the manuscript.

REFERENCES:

1. Kachhia, J. A. (2026). Enhancing early diabetes screening through ML and explainable AI. IEEE i-COSTE 2025.
2. Islam, M., Tisha, N. T., Alom, M. R., Oyshe, K. U., & Rahaman, M. A. (2025). An explainable AI-based ensemble ML framework for early-stage diabetes prediction. IEEE 2025.
3. Abu-Shareha, A. A., et al. (2026). Diabetes prediction using hybrid supervised and unsupervised techniques on the PIMA dataset. JAIT, 6, 79–87. https://doi.org/10.37965/jait.2025.0899
4. IDF. (2021). IDF Diabetes Atlas (10th ed.). https://www.diabetesatlas.org
5. Pradhan, D., et al. (2025). Therapeutic interventions for diabetes mellitus. Current Diabetes Reviews, 21(8).
6. Lowe, W. L., et al. (2019). Association of gestational diabetes with maternal disorders. JAMA, 320(10), 1005–1016.
7. Smith, J. W., et al. (1988). Using the ADAP algorithm to forecast the onset of diabetes. Proc Annual Symposium Computer Application Medical Care, 261–265.
8. Sharma, T., & Shah, M. (2021). A comprehensive review of ML techniques on diabetes detection. Visual Computing for Industry, Biomedicine and Art, 4(1), 30.
9. Wee, B. F., et al. (2024). Diabetes detection based on ML and deep learning. Multimedia Tools and Applications, 83(8), 24153–24185.
10. Tanim, S. A., et al. (2025). Explainable deep learning for diabetes with DeepNetX2. Biomedical Signal Processing and Control, 99, 106902.
11. Enríquez-Ortega, D., et al. (2025). Enhancing diabetes diagnosis through ML. Applied Sciences, 15(18).
12. Toleva, B., et al. (2025). Effective methodology for diabetes prediction with class imbalance. Bioengineering, 12(1).
13. Saihood, Q., & Sonuc, E. (2023). Early detection of diabetes using ensemble ML. Turkish J. Electrical Engineering, 31(4), 722–738.
14. Maniruzzaman, M., et al. (2020). Comparative approaches for the classification of diabetes mellitus. Computer Methods Programs Biomedicine, 152, 23–34.
15. Domingos, P. (2012). A few useful things to know about ML. Communications ACM, 55(10), 78–87.
16. Matthews, D. R., et al. (1985). Homeostasis model assessment. Diabetologia, 28, 412–419.
17. Zou, Q., et al. (2018). Predicting diabetes mellitus with ML techniques. Frontiers in Genetics, 9, 515.
18. Chang, V., et al. (2023). PIMA Indians' diabetes classification based on ML. Neural Computing and Applications, 35(22), 16157.
19. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS 30.
20. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56–67.
21. Shams, M. Y., et al. (2025). A novel RFE-GRU model for diabetes using PIMA. Scientific Reports, 15(1), 982.

22. Talari, P., et al. (2024). Hybrid feature selection for early prediction of type 2 diabetes. PLOS ONE, 19(1), e0292100.
23. Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 16, 321–357.
24. DeLong, E. R., et al. (1988). Comparing areas under correlated ROC curves. Biometrics, 44(3), 837–845.
25. Alkalifah, B., et al. (2025). ML-based regression for diabetes levels. Heliyon, 11(1).
26. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD, 785–794.
27. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.28. Sarma, A. D., & Devi, M. (2025). AI in diabetes management. Hormones, 1–16.
28. Mittal, R., et al. (2025). ML for early detection of type 1 diabetes. Int. J. Molecular Sciences, 26(9).
29. Kaur, R., & Rani, R. (2020). Comparative analysis of ML algorithms for diabetes. J. Healthcare Engineering, 2020.