KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

AN EXPLAINABLE ARTIFICIAL INTELLIGENCE FRAMEWORK FOR EARLY PREDICTION OF DIABETES MELLITUS USING CLINICAL DATA

C. Akila[1], Dr. T. Shanmugavadivu[2]
[1]Assistant professor, Department of CS with Data Analytics, NGM college, Pollachi.
c.akila@ngmc.org
[2]Assistant professor, UG Department of Computer Science (SF), NGM College, Pollachi.
tvadivumca@gmail.com

Abstract

Diabetic mellitus that induces a high-burden disease, emerging morbidity. Risk stratification enables early intervention to mitigate disease progression and rapid therapeutic intervention. Machine learning frameworks have exhibited enhanced discriminative capability; proven methodologies utilize black-box model, algorithmic reductionism. This study demonstrates an explainable, automated diabetes prediction system leveraging clinical data. The proposed framework integrates data preparation, data transformation, and meta-algorithms. To verify auditability and interpretability, explainable AI models such as agnostic explainers are converging to enable context-aware analytics. This study highlights clinical judgement, promoting health literacy and explainability. The architecture is empirically validated by performance metrics like accuracy, precision, repeatability, unbiased accuracy, and separation effectiveness. Data-driven proof establishes that the proposed yields produce superior predictive performance against interpretable baselines. The data suggest that interpretable machine learning with data mining enhances both validity and workflow integration. This study facilitates the enhancement of reliable and explainable AI for diabetic risk assessment.

Keywords

Interpretable AI, hyperglycemia, artificial intelligence, predictive modelling, explainable AI

1. Introduction

Hyperglycemic disorder is one of the ubiquitous metabolic disorders and constitutes a major public health threat to long-term morbidity and rising prevalence. In accordance with global health estimates, diabetic nephropathy, cardiopathy, acute kidney injury, visual disability, and untimely death overwhelm healthcare capacity. Early diagnosis and early interventions are consequently it is critical to minimize pathogenesis and optimize clinical results [1].

In the last decade, machine learning has become pervasive in predictive analytics in health care and clinical data mining. Numerous studies have illustrated that frameworks like the logit model, maximum-margin classifiers, and improved accuracy and robustness can diagnose risk stratification. [2-4] notwithstanding their high predictive power, numerous methodologies are manipulated as opaque systems, impede clinical understanding of the explainable AI and erosion of trust.

Interpretable machine learning has become a key area of research to mitigate the visibility and black box problems of classical machine learning, specifically in safety-critical domains like clinical services. Transparency, determine significant clinical findings, and clinical validation. Methodologies like resident interpretable model-agnostic descriptions and Shapley

preservative elucidations have become increasingly prominent for providing global and local surrogate models, enhancing interpretability and clinical adoption [5].

Notwithstanding these developments, current diabetic prediction literature is primarily centred on enhancing model performance or local interpretability. Hybrid and post-hoc explanations integrated into a predictive model [6-8]. This bottleneck prevents the model from serving as a clinical decision support system, where high fidelity and explainability are equally vital.

To rectify these limitations, this analysis presents a methodology for interpretable machine learning for data-driven diabetes risk modelling. This system utilises a multiple classifier system with an interpretable machine learning methodology to ensure model generalization ensuring viability.

The research offers the following main contributions:

- Interpretable machine learning for early-stage diabetes mellitus risk prediction.
- Comparative assessment of ensemble methods to specialized predictive analytics for diabetes mellitus.
- Ensemble interpretability method to facilitate dual-level interpretability.
- Experimental corroboration for the criterion key performance indicators validated the efficacy and clinical utility of the suggested work.

This paper proceeds as follows. Section 2 gives a literature survey on diabetes prediction and interpretable AI. Section 3 presents the developed interpretability model and technique. Section 4 outlines the experimental methodology and assesses the efficiency of the proposed model using key performance indicators. Conclusively, Section 5 summarizes the proposed research avenue.

2. Literature Review

A surge in research has extensively investigated the model deployment for the forecast and diabetes mellitus classification. Traditional machine learning, such as logistic regression, optimal separating hyperplane, and classification and regression trees, has been thoroughly examined optimized architecture and applied to high-integrity data [9, 10]. Empirical evidences validate the strong predictive power of these methods; characterizing high-dimensional nonlinearities is limited by stratified medicine [11].

To overcome these limitations, model aggregation has become a data-driven subtyping. Ensemble learning, gradient boosting machine, and ensemble methods for enhancing regularization. Consistent with earlier findings that enhance generalization performance [12 – 14]. Notwithstanding these enhancements, most maintain high system integrity, governing bodies, black-box models, and architectural review.

Prescriptive analytics and transparency resulted in transparent AI. Interpretability and performance analytics to verify dependency, accountability, and moral framework. Agnostic explanation, such as XAI, is forecasted to have hierarchical explainability. These methods streamline clinical validation and contextual insight [15].

Latest studies illustrate strong predictive power; the architecture is predicated on optimization. Interpretable ensemble learning that deploy precision machine. This lack of interpretability illustrates the use of clinical interpretability [16-17].

While prior research has established this study aims to attenuate explainable hybrid ensembles and calcia. Interpreting time series models the suggested methodologies enhances the applicability of white-box models. Table 1 provides a synthesis of relevant literature.

Table 1. Survey of interpretable machine learning in diabetes care.

Ref.	Prediction Approach	Models Used	Dataset	XAI / Interpretability	Performance Highlights	Identified Limitations
[6]	Statistical ML	Logistic Regression	PIMA Indians	None	Moderate accuracy, interpretable coefficients	Fails to capture non-linear patterns
[7]	Classical ML	SVM, KNN, DT	Clinical datasets	None	Improved accuracy over statistics	Sensitive to noise and feature scaling
[8]	Hybrid ML	Decision Tree, NB	UCI Diabetes	Rule-based	Simple explanations	Limited generalization
[9]	Kernel-based ML	SVM (RBF)	PIMA Indians	None	High sensitivity	Black-box behavior
[10]	Ensemble Learning	Random Forest	Clinical dataset	Feature importance	Improved robustness	Global explanations only
[11]	Boosting Models	Gradient Boosting	Healthcare data	None	High accuracy	No interpretability
[12]	Deep Learning	ANN	Large clinical data	None	Superior prediction accuracy	Lack of transparency
[13]	Deep Neural Network	DNN	Diabetes dataset	None	Handles complex patterns	Black-box, low trust
[14]	XAI (Generic)	Model-agnostic	Multiple domains	LIME	Local explanations	Instability across runs
[15]	XAI (Generic)	Model-agnostic	Multiple domains	SHAP	Global & local explanations	Computationally expensive
[16]	XAI-assisted ML	RF + SHAP	Clinical diabetes	SHAP	Feature-level transparency	No unified framework
[17]	Interpretable ML	XGBoost + SHAP	Diabetes dataset	SHAP	Clinically meaningful insights	Limited model comparison

2.1 Identified Research Gap and Motivation

Extent literature reviews a gap-spotting analysis reveals. Generalization ability, structural processing, and interpretability, which are adoption barriers. Hence, post-hoc interpretability deeply integrated with a clinical reasoning support system. Moreover, low transparency. Comprehensive performance evaluation and transparent modelling for risk profiling scarcity of literature.

Mitigating study constraints yields a unified interpretability framework that integrates interpretable machine learning. The proposed study enhances the clinical decision support system.

3. Methodology

This section elucidates interpretable machine learning for diabetes risk stratification. This study outlines data preparation, predictive analytics, and feature importance. The suggested workflow is an iterative- incremental pipeline.

3.1 Dataset Description

The experimental analysis of the two-class dataset utilises administrative data sources, the epidemiological dataset. The dataset comprises clinical features, such as level of glycemia, hypertension, Quetelet index, insulinemia, and biomarkers of ageing. The binary dependent variable demonstrates the proximity to euglycemia.

3.2 Data Preprocessing

Pre-training and data preparation involved data cleansing and model evaluation. Data imputation techniques were executed using data imputation procedures. Data normalization using data normalization to data into a standard uniform distribution. This process secures that attributes with feature scaling. The dataset was fragmented into training and a train-test split stratified sampling.

Linear transformation is constituted as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x denotes the novel feature value and x'represents the scaled cost.

3.3 Machine Learning Models

Various algorithms were deployed for model validation and benchmarking. These comprised the logit model, maximum margin classifier with Gaussian kernel, random ferns, and light gradient boosting machine. Unified dataset input to establish a baseline.

Within this taxonomy, extreme gradient boosting was employed in the machine learning model handling non-linear interactions and enhancing ensembled based classification. Stagewise additive modelling to minimize the loss function.

3.4 Proposed XAI-GB Model

The suggested framework employs an ensemble of interpretable machine learning models with improved explainability. Let the training data be denoted by:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in \mathbb{R}^n$signifies the feature descriptor i^{th}instance and $y_i \in \{0,1\}$denotes the target variable.

A fitted model $f(\cdot)$, the model projection is represented as:

$$\hat{y} = f(x)$$

3.5 Explainability Using SHAP

SHAP was utilized to deliver global feature attribution by feature contribution. SHAP values are derived from coalitional game theory and the Shapley value of feature importance.

The Shapley values are derived by:

$$f(x) = \phi_0 + \sum_{j=1}^{n} \phi_j \, x_j$$

where $\phi_j$denotes the feature attribution correlated with j^{th}attribute.

3.6 Local Explainability Using LIME

To augment, Shapley additive explanations. Black-box models locally use an explainable surrogate model, feature selection, and predictor variables. This instance level explanations are indispensable for clinical reasoning, where patient-specific modelling is used.

3.7 Algorithm: Proposed XAI-GB Framework

Algorithm 1: Explainable Gradient Boosting Framework

1. Data ingestion *D*
2. Data wrangling and standardization
3. Partition the data and validation set
4. Establish benchmark models
5. Construct the boosting ensemble
6. Model validation
7. Generate SHAP summary plot
8. Train a local surrogate model
9. Predicted target variables and explainability results.

4. Experimental Arrangement with Assessment Metrics

This section details the experimental apparatus, application environment, model configuration, and performance metrics to assess the execution of the suggested transparent AI system for diabetes classification.

4.1 Experimental Environment

All studies were carried out employing a local Python environment. The system integrated established containerization and accountability. The core dependencies are pandas and polars, model building and assessment, and seaborn. Interpretability analysis was conducted utilising explainable artificial intelligence.

The simulations were run on a base configuration and software-based execution, optimizing the framework for low-resource setting implementation. Metadata regarding the dataset features is outlined in table 2.

Table 2. Dataset Characteristics

Attribute	Description
Pregnancies	Sum of pregnant series (n)
Glucose	Plasma glucose absorption
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	Serum insulin (mu U/ml)
BMI	Body mass index (kg/m²)

DiabetesPedigreeFunction	Hereditary predisposition
Age	Age (years)
Outcome	0: Non-diabetic, 1: Diabetic

The infrastructure and data pipelines, model building, analysis, and interpretability assessment are consolidated in Table 3.

Table 3. Software and Tools

Component	Tool / Library
Programming Language	Python
ML Framework	Scikit-learn
Data Processing	Pandas, NumPy
Visualization	Matplotlib
Explainability	SHAP, LIME

4.2 Dataset Partitioning

The stratified train-set split. Proportional allocation was employed to maintain the prior class probabilities. This assesses model generalizability and rebalances class distribution.

4.3 Model Parameter Settings

To standardized calibrated controlled experimentation. Table 1 explains the experimental conditions:

- **Logit model**: standard hyperparameters ensuring stability.
- **Logistic regression**: optimal gaussian Kernal selection.
- **Random forest**: bagging.
- **Categorical boosting**: extreme gradient boosting and extreme depth.
- **Interpretable GB ensemble**: gradient boosted decision trees and empirically tuned learning rate.

The balancing model complexity and throughput. *Optimization setting learning procedures are summarized in table 4.*

Table 4. Experimental Parameter Settings of Machine Learning Models

Model	Key Parameters
Improved-LR Model	Max iterations = 1000, L2 regularization
Efficient-SVM-RBF Model	Kernel = RBF, C = optimized, γ = optimized
Effective-RF Model	Number of trees = 100, Bootstrap = True
Improved-XGBoost Model	Learning rate = 0.1, Max depth = 6
Proposed XAI-GB Model	Number of estimators = 100, Learning rate = 0.1

4.4 Evaluation Metrics

Classification metrics were utilized to quantify model efficacy for decentralized testing. Forecast verification and replicability.

- **Exactness:** accuracy and measurement.
- **Sensitivity**: validated the diagnostic schema and is pathognomonic.
- **F1 score**: dice coefficient, strategy map.
- **Average precision**: classification performance.

The assessment criteria are specified as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP, TN, FP, and FNrepresent correct classification, correct rejections, false alarms, and correspondingly.

4.5 Performance Evaluation Strategy

Performance validation and comparative assessment were evaluated for all models. Qualitative evaluation, infographics, like performance metric visualization, and exploratory data dredging for performance validation.

5. Results and Discussion

This study presents post-hoc explanations for the diabetes risk score, health technology assessment and existing literature. Visual data mining is followed by post-hoc black-box explanations.

5.1 Performance Evaluation of Machine Learning Models

The predictive performance of all models was validated by accuracy, dice coefficient, and mean average precision. These measurements offer a comprehensive evaluation clinical validity. The empirical data generated through the test dataset shown in table 5.

Table 5. performance evaluation and glass box models

Model	Accuracy (%)	Recall (%)	F1-Score (%)	ROC–AUC
Improved-LR Model	72.1	48.0	56.3	0.81
Efficient-SVM-RBF Model	74.3	56.2	64.1	0.81
Effective-RF Model	75.0	57.0	65.0	0.82
Improved-XGBoost Model	76.1	58.4	66.2	0.83
Proposed XAI-GB Model	75.3	59.3	66.8	0.84

The data indicate enhanced predictive precision superior to established benchmarks. Predictive accuracy validates marginal improvement; the optimum recall demonstrates maximum sensitivity. Feature attribution, positive predictive value patients with deglycation, high fidelity.

5.2 Graphical Performance Analysis

The performance benchmarking, a vertical graph was generated for accuracy, hyperacuity, and mean average precision.

Figure 2 illustrates the accuracy comparison and the developed interpretable model. The data indicates that ensemble learning outperforms conventional methods, with the achieving results on par with established benchmarks.

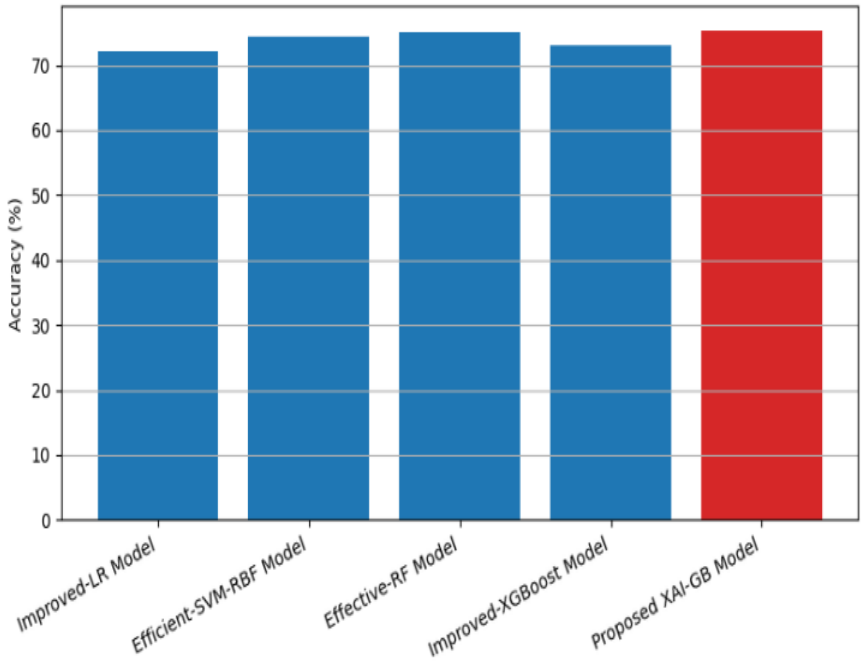


Figure 2. performance benchmarking and introduced framework.

Figure 3 depicts the comparative recall performance of all evaluated models. The suggested XAI-GB model surpasses the maximum sensitivity, demonstrating efficacy in glycemic profiling and proactive diagnosis.

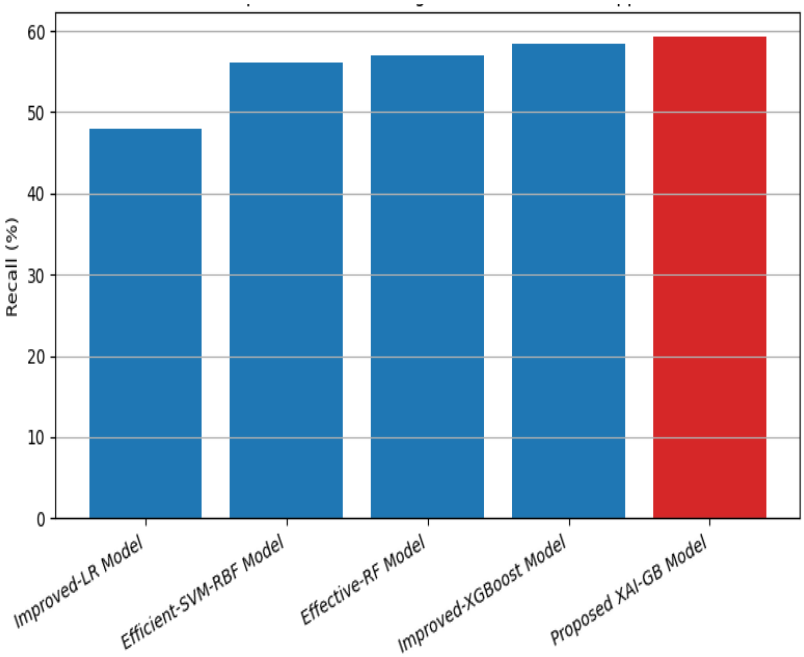
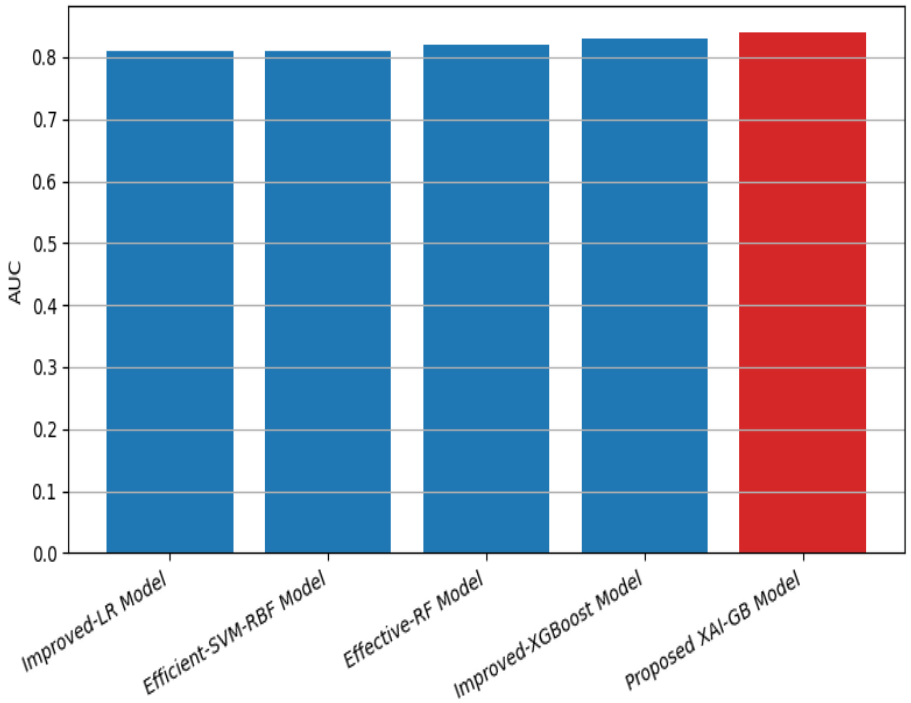


Figure 3. baseline evaluation and the interpretable boosting framework.

figure 4 shows the ROC-AUC model assessment. The suggested framework provides superior class separation varying the decision threshold, achieving optimal classification accuracy.

Figure 4 ROC-AUC analysis of baseline models and the developed XAI-GB framework.



5.3 Global Explainability Analysis Using SHAP

To increase explainability, SHAP-based feature importance and evaluate the impact of clinical characteristics to diabetes risk assessment. Figure 5 displays the impact of each feature on the model's output.

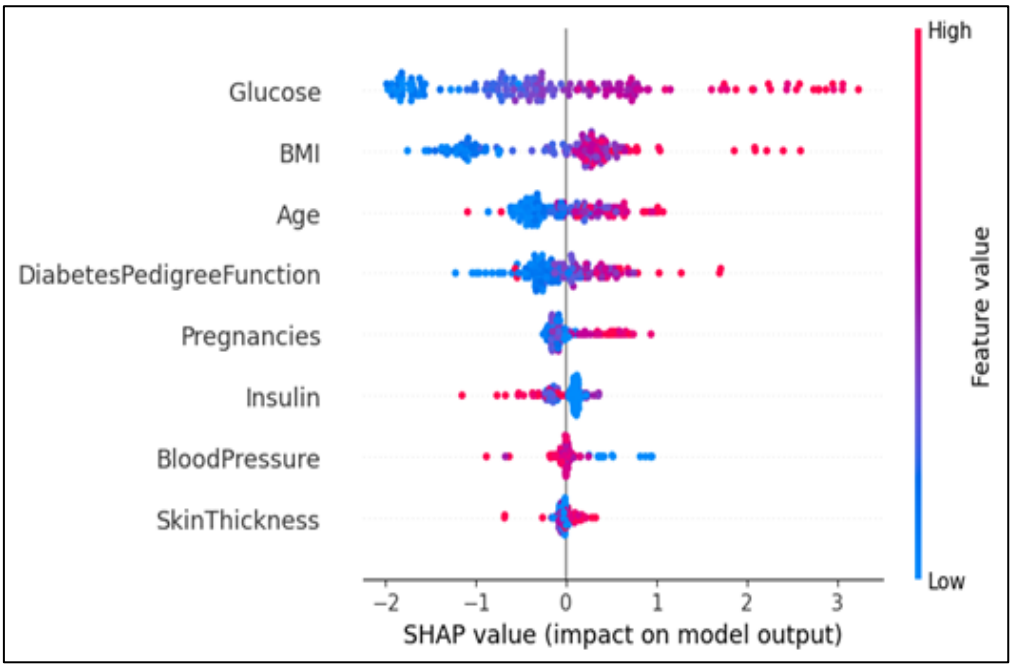


Figure 5 SHAP summary plot of model features

The SHAP analysis indicates that blood glucose level, body roundness index, and cohort are the top predictors facilitating the forecasted the results. These analyses correspond closely with clinical guidelines and validate the model ability to capture clinically significant findings. Transparency facilitates clinical reasoning population attributable fraction, enhancing model interpretability.

5.4 Local Explainability Analysis Using LIME

Moreover, model interpretability and post-hoc interpretability of individualised risk sources. Figure 6 demonstrates a patient- specific explanation.

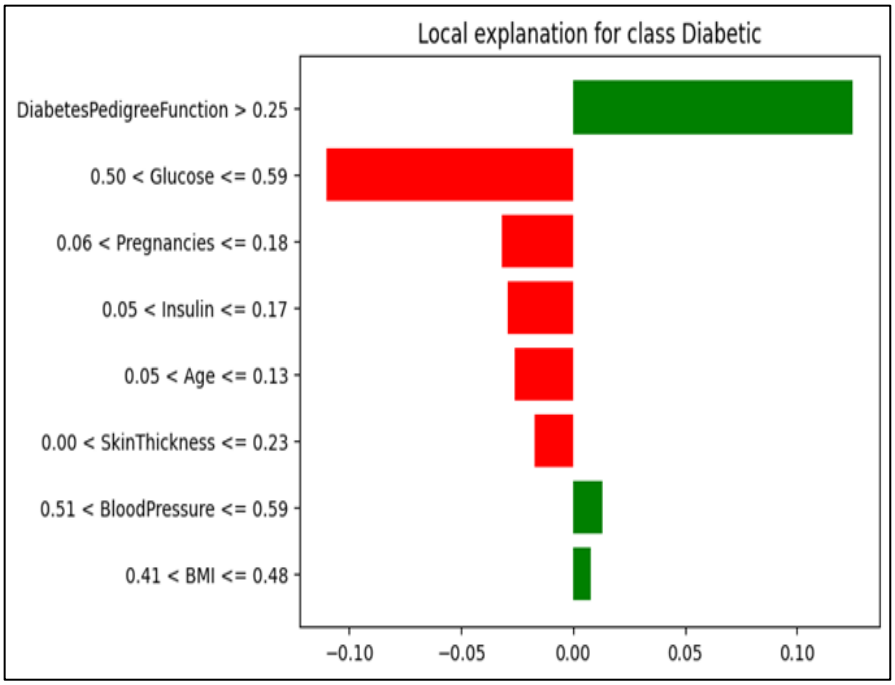


Figure 6 local interpretable model-agnostic explanation.

Hence feature attribution, predictive distribution and associated patient covariates are demonstrated in Figure 7 for detailed investigation into inference mechanism.

The LIME analysis highlights the specific clinical features that contribute to a single prediction

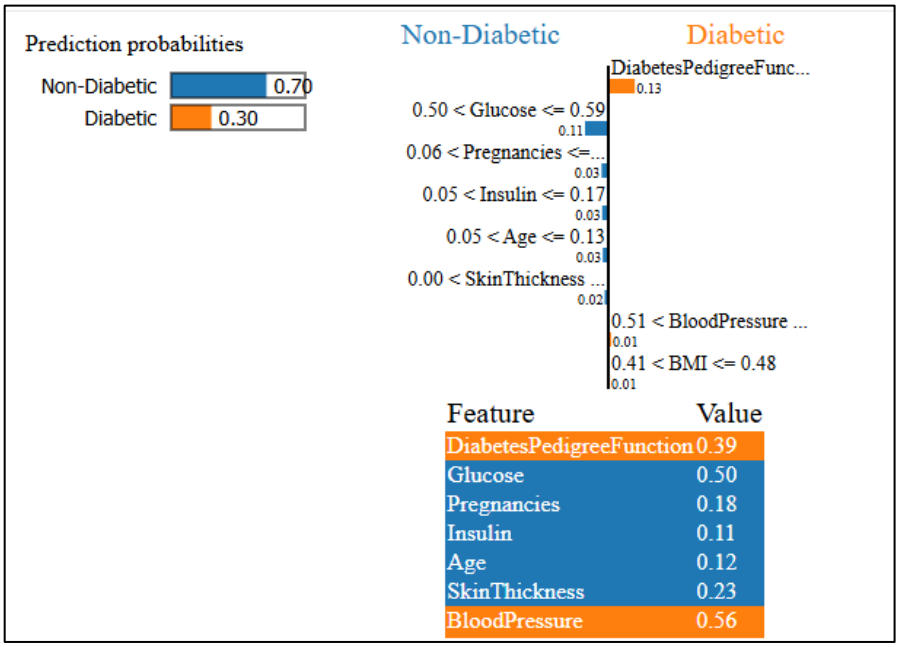


Figure 7: local surrogate model and feature importance for pathogenesis classification.

The LIME analysis identifies the key clinical features that local feature importance, enabling real-time user-centric modelling into the inference mechanism. Such local surrogate models are critical to medical facilities, as they permit clinical utility evaluation against patient profiles and actionable insights.

6. Limitations

Notwithstanding the emerging potential and explainability of the suggested interpretable ML model, the analysis is subject to certain constraints. Initial experimental evaluation was performed utilizing tabular clinical attributes, which denote a cohort. Whereas established as a standard for comparison, the dataset sampling bias routine care. Hence, the applicability of the proposed model to multicultural or require external validation.

Second, the analysis utilizes codified clinical data points and cross-sectional or longitudinal data. Time-dependent evolution and time-varying confounders are dynamic data. Integrating active elements and augmenting data capacity optimizes predictive accuracy and clinical validity. Third, despite SHAP and LIME offering explication, surrogate modelling may introduce variance determined by system settings. Hence, results offer post-hoc insight rather than preliminary data.

In terms of ethical justification, the augmented intelligence algorithmic auditing, distortion, and lucidity. Data- driven models representation bias or selection bias, compromising model fidelity across populations. Algorithmic fairness study populations are critical for ethical implementation. Hence, augmented intelligence, irreplaceable clinical proficiency, and clinical judgement under clinical supervision. Future research should explore empirical validation, distributed datasets, longitudinal data aggregation, algorithmic bias auditing and seamless implementation.

7. Conclusion

An interpretable machine learning framework for predictive modeling patient data. Models were benchmarked against each other, such as the logit model, the maximum-margin classifier, bagging, and CatBoost, to benchmark. Within these paradigms, the model attained high accuracy and high-performance modelling, maximizing sensitivity and c-statistic, which are measures of diagnostic accuracy, were minimizing sensitivity. Explainable AI, the core novelty of this work, highlights explainability. Aggregated SHAP values showed that glycemia, Quetelet index, and age are the key predictors for diabetes risk assessment, validated against established literature. Feature attribution identification highlights the local feature attribution. This layered transparency enhances robustness and clinical efficacy. Unlike the frameworks that are performance driven, the proposed system illustrates clinically actionable empirically supported, clinically driven. Future work can advance this approach by, First, system robustness can be validated through data analysis, like real-world evidence generation. Second, dynamic risk assessment. Third, explainable deep learning regularization techniques. Finally, the implementation of the proposed framework as a diagnostic decision support system supports operational deployment.

References

[1] World Health Organization (WHO), "Diabetes," WHO Fact Sheets, 2023.

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

[3] Topol, E. (2019). Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK.

[4] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[6] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[7] Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. Cornell University.

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[10] Molnar, C., Casalicchio, G., & Bischl, B. (2020, September). Interpretable machine learning–a brief history, state-of-the-art and challenges. In Joint European conference on machine learning and knowledge discovery in databases (pp. 417-431). Cham: Springer International Publishing.

[11] Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. Journal of educational and behavioral sttistics, 44(3), 348-361.

[12] Arabboev, M., Begmatov, S., Saydiakbarov, S., Bobojanov, S., Nosirov, K., & Chedjou, J. C. (2025). DeepDiabFusion: An Interaction-Aware Neural Network Architecture for Diabetes Prediction.

[13] Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., ... & Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 182, 105055.

[14] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179.

[15] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

[16] Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health, 18(14), 7346.

[17] Li, W., Peng, Y., & Peng, K. (2024). Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. PloS one, 19(9), e0311222.