Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 2, September 2026 – www.caeaccess.org

An Inclusive Diabetes Detection System using Machine Learning and Explainable Artificial Intelligence

Oluwaponmile H. Adefuwa
Department of Computer Science, Faculty of Science, Lagos State University, Ojo, Lagos State, Nigeria

Olatunde D. Akinrolabu
Department of Data Science and Artificial Intelligence, Faculty of Computing, Adekunle Ajasin University, Akungba Akoko, Ondo State, Nigeria

Ayodeji O. Ayeni
Department of Information and Communication Technology, Faculty of Computing, Adekunle Ajasin University, Akungba Akoko, Ondo State, Nigeria

ABSTRACT
Diabetes mellitus is a chronic metabolic disorder affecting over 537 million adults globally, with prevalence expected to reach 783 million by 2045. Even with advances in diagnostic technology, equitable and early detection remains a challenge, particularly across demographically diverse populations. This study presents the design and development of an inclusive diabetes detection system that combines machine-learning classification, Local Interpretable Model-agnostic Explanations (LIME)-based explainability, and interactive Streamlit web deployment. A publicly available synthetic dataset of 100,000 patient records was preprocessed through encoding, Min-Max normalisation, and SMOTE-based class balancing. Four machine learning models (XGBoost, Random Forest, Naive Bayes, and Long Short-Term Memory (LSTM)) were trained and evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. Hyperparameter tuning was conducted using Grid Search with 5-fold cross-validation. XGBoost achieved the best performance with 91.99% accuracy, precision of 1.0000, recall of 0.8665, F1-score of 0.9285, and AUC-ROC of 0.9438, clearly exceeding the research target of 70% accuracy. Demographic subgroup analysis across gender, ethnicity, age group, and income level confirmed equitable performance, with F1-score variation below 0.004 across most demographic dimensions. LIME generated clinically meaningful individual-level explanations, consistently identifying HbA1c and fasting glucose as the dominant predictive features. A broader evaluation was done, the trained XGBoost model was further tested on the Pima Indians Diabetes Dataset, a widely used benchmark with only 8 features, achieving 72.53% accuracy and AUC-ROC of 0.7969 using only 6 of 38 model features, demonstrating generalisation capability beyond the original training distribution. The complete system was deployed as a browser-based Streamlit application enabling real-time prediction and transparent explanation for clinical and non-clinical users.

General Terms
Machine Learning, Deep Learning, Artificial Intelligence, Explainable Artificial Intelligence

Keywords
Diabetes prediction, XGBoost, Random Forest, LSTM, Naive Bayes, LIME, Explainable AI, Streamlit, Demographic inclusivity, SMOTE, Hyperparameter tuning.

1. INTRODUCTION
Diabetes mellitus is a chronic metabolic disorder characterised by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both. It is currently one of the most rapidly expanding global health crises. According to the International Diabetes Federation, approximately 537 million adults were living with diabetes in 2021, a figure projected to rise to 783 million by 2045 if current trends persist [1]. The disease manifests in three primary forms: Type 1, an autoimmune condition; Type 2, driven largely by lifestyle factors and genetic predisposition; and gestational diabetes, occurring during pregnancy. A large proportion of individuals living with diabetes remain undiagnosed until severe complications such as retinopathy, nephropathy, neuropathy, and cardiovascular disease have already developed [2].

With advances in diagnostic technology, equitable and early detection of diabetes remains an unresolved challenge. Research has consistently demonstrated that diabetes prevalence, risk profiles, and clinical presentations vary significantly along demographic dimensions such as age, sex, ethnicity, socioeconomic status, and geographic setting. Existing screening approaches rely on fasting blood glucose tests, oral glucose tolerance tests, and HbA1c assays, supplemented by risk-scoring tools such as the Finnish Diabetes Risk Score (FINDRISC) and the American Diabetes Association (ADA) Risk Test. While effective at the population level, these tools are constrained by reliance on single or few biomarkers, resource intensity that limits accessibility in low- and middle-income settings, and validation predominantly on Western populations, introducing demographic bias when applied to broader populations [3].

The emergence of machine learning (ML) in healthcare has opened transformative possibilities for improving diagnostic accuracy. ML algorithms are capable of learning complex, high-dimensional patterns from diverse clinical datasets, potentially surpassing traditional risk calculators in predictive performance. Existing studies have demonstrated the effectiveness of ensemble methods such as Random Forest and XGBoost, probabilistic classifiers such as Naive Bayes, and deep learning architectures such as Long Short-Term Memory (LSTM) networks for diabetes prediction [4][5][6][7]. However, a persistent barrier to clinical adoption is the opacity of high-performing models, which function as black boxes and produce predictions without offering reasoning that clinicians or patients can interrogate [8].

Explainable Artificial Intelligence (XAI) addresses this problem by rendering model decisions transparent and interpretable. Local Interpretable Model-agnostic Explanations (LIME) generates locally faithful explanations for individual predictions by approximating complex model behaviour with interpretable surrogate models, making it well-suited for clinical deployment scenarios where personalised explanations

are valued [9]. Furthermore, a persistent deployment gap exists between models developed in research and tools that can be operationalised in clinical settings. Interactive frameworks such as Streamlit enable rapid development of browser-based web applications exposing model functionality to non-technical users such as clinicians, community health workers, and patients [10].

A critical gap in the existing literature is that most diabetes prediction studies stop at reporting classification accuracy without deploying models as usable tools, without generating individual explanations, and without systematically evaluating whether performance is equitable across demographically diverse subgroups [11][12]. This study addresses all three gaps simultaneously by proposing an inclusive diabetes detection system that integrates rigorous ML classification, LIME-based explainability, demographic subgroup fairness evaluation, and interactive Streamlit deployment.

2. LITERATURE REVIEW

A substantial body of research has explored machine learning approaches for diabetes prediction, using a variety of algorithms, datasets, and preprocessing strategies. Maniruzzaman et al. [4] developed an ML-based system using Logistic Regression (LR) for risk factor identification and four classifiers, namely Naive Bayes (NB), Decision Tree (DT), AdaBoost, and Random Forest (RF), for prediction. Using the National Health and Nutrition Examination Survey dataset of 6,561 respondents, the combination of LR-based feature selection and RF classification achieved an overall accuracy of 90.62% and AUC of 0.95 under 10-fold cross-validation. The study identified age, education, BMI, systolic and diastolic blood pressure, and cholesterol as significant diabetes risk factors.

Abnoosian et al. [5] proposed an innovative pipeline-based multi-classification framework using the imbalanced Iraqi Patient Dataset of Diabetes. Their framework incorporated duplicate removal, missing value imputation, data normalisation, k-fold cross-validation, and a weighted ensemble of k-NN, SVM, DT, RF, AdaBoost, and Gaussian NB, with AUC-based weight assignment. Grid search and Bayesian optimisation were used for hyperparameter tuning. The proposed ensemble achieved an average accuracy of 0.9887, an F1-score of 0.9851, and an AUC of 0.999, outperforming all individual classifiers. This work demonstrated the importance of addressing class imbalance and leveraging ensemble weighting in medical prediction tasks.

Hasan et al. [6] proposed a robust framework for diabetes prediction using the Pima Indian Diabetes Dataset, incorporating outlier rejection, missing value handling, data standardisation, feature selection, and K-fold cross-validation. Multiple ML classifiers, including KNN, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP) were compared, with the proposed AUC-weighted ensemble achieving an AUC of 0.950, outperforming prior state-of-the-art results by 2%.

Zhou et al. [7] proposed a diabetes prediction model based on an enhanced deep neural network (DLPD) using dropout regularisation and binary cross-entropy loss. Experiments on the Pima Indians Diabetes Dataset achieved 99.41% training accuracy, demonstrating the capacity of deep learning architectures to extract non-linear patterns in clinical data.

Rastogi and Bansal [3] evaluated diabetes prediction using data mining techniques including SVM, Naive Bayes, Regression, and Random Forest, emphasising that proper preprocessing and model selection are critical for prediction accuracy. Febrian et al. [11] compared KNN and Naive Bayes algorithms for diabetes prediction using supervised machine learning, finding that Naive Bayes outperformed KNN in accuracy when evaluated using a confusion matrix framework.

Bukhari et al. [13] presented an improved artificial neural network (ANN) model using a dataset comprising female diabetic patients. The study highlighted the role of ANN architecture optimisation in improving prediction performance and demonstrated competitive accuracy for clinical diabetic detection.

Ihnaini et al. [14] proposed a smart healthcare recommendation system (SHRS-M3DP) using deep ensemble learning with IoMT data fusion for multidisciplinary diabetes patients. The system fused multiple datasets and applied an ensemble DML architecture, achieving 99.64% accuracy after data fusion, significantly outperforming individual classifiers.

Mahajan et al. [8] conducted a comprehensive review of ensemble learning methods for disease prediction across five major chronic diseases. The review found that boosting achieved the best results 40.5% of the time across studies, followed by bagging, providing systematic evidence for the superiority of ensemble approaches in medical AI.

Krishnamoorthi et al. [15] developed a novel diabetes healthcare disease prediction framework using machine learning techniques. Their study evaluated multiple algorithms and emphasised the need for systematic framework development guided by a rigorous literature review of existing prediction models.

Akther et al. [10] developed a Streamlit-based AI system for multi-disease prediction that integrated ML, deep learning, and ensemble methods with LIME and SHAP explainability tools. Their system covered diabetes, heart disease, breast cancer, and liver disease prediction within a single interactive interface, achieving competitive performance and demonstrating that XAI tools significantly improve model clarity and clinical trust.

Singh et al. [16] presented a comparative analysis of ML algorithms for multiple disease prediction with optimised scalable deployment, testing Logistic Regression, Random Forest, and SVM with preprocessing for handling missing data and feature scaling. Beghriche et al. [17] developed an efficient prediction system for diabetes using deep neural networks, evaluating performance across multiple metrics on the Pima Indian Diabetes Dataset.

Arumugam et al. [12] implemented multiple disease predictions using ML algorithms, including Naive Bayes, SVM, and Decision Trees, with performance evaluated on accuracy and error rate. Sharma et al. [18] predicted diabetes using ML models with preprocessing and comparative algorithm evaluation, finding high accuracy using ensemble approaches on the Pima Indian dataset.

All these studies confirm the strong potential of machine learning for diabetes prediction. However, most work evaluates performance only at the overall level without assessing equity across demographic subgroups, and very few studies integrate XAI with interactive deployment in a unified system. The present study addresses these gaps through LIME-based explanations, demographic subgroup analysis, and model deployment.

3. METHODOLOGY

3.1 Research Strategy

This study followed a quantitative, applied design methodology using a cross-sectional approach. A publicly available synthetic diabetes dataset was obtained from Kaggle, preprocessed for machine learning, and used to train and evaluate four ML classifiers. The best-performing model was integrated with LIME for explainability and deployed as a Streamlit web application.

3.2 Dataset

The dataset used was the Diabetes Health Indicators Dataset, a synthetic dataset of 100,000 patient records sourced from Kaggle (https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset). It contained 35 columns spanning demographic, lifestyle, and clinical domains. The dataset was synthetically generated using statistical distributions derived from real-world medical research, ensuring clinically realistic patterns while preserving patient privacy. The target variable, diagnosed_diabetes, was binary (0 = Non-Diabetic, 1 = Diabetic). Table 1 summarises the dataset variables.

Table 1: Dataset Variables and Categories

Feature(s)	Category	Description
age, gender, ethnicity	Demographic	Patient background and identity attributes for subgroup analysis
education_level, income_level, employment_status	Socioeconomic	Measures of social and economic position
smoking_status, alcohol_consumption_per_week	Lifestyle	Substance use behaviour indicators
physical_activity_minutes_per_week, diet_score, sleep_hours_per_day	Lifestyle	Modifiable behavioural health factors
family_history_diabetes, hypertension_history, cardiovascular_history	Medical History	Binary flags for prior conditions (0=No, 1=Yes)
bmi, waist_to_hip_ratio	Anthropometric	Body composition measurements
systolic_bp, diastolic_bp, heart_rate	Vital Signs	Cardiovascular and haemodynamic indicators
cholesterol_total, hdl_cholesterol, ldl_cholesterol, triglycerides	Lipid Panel	Blood lipid measurements (mg/dL)
glucose_fasting, glucose_postprandial	Glycaemic	Fasting and post-meal blood glucose levels (mg/dL)
insulin_level, hba1c	Glycaemic	Insulin (mcU/mL) and glycated haemoglobin (%)
diagnosed_diabetes	Target Variable	Binary outcome: 0 = Non-Diabetic, 1 = Diabetic

Table 2: Summary of Machine Learning Models Used

Model	Type	Key Strength	Limitation
XGBoost	Ensemble (Boosting)	High accuracy, handles imbalanced data	Computationally intensive
Random Forest	Ensemble (Bagging)	Robust, provides feature importance	Less interpretable
Naive Bayes	Probabilistic	Fast, low computational cost	Assumes feature independence
LSTM	Deep Learning (RNN)	Captures complex non-linear patterns	Requires larger dataset, slower

Table 3: Grid Search Hyperparameter Tuning Results

Model	Best CV F1	Best Parameters
XGBoost	0.9299	learning_rate=0.1, max_depth=4, n_estimators=100, subsample=0.8, colsample_bytree=0.8
Random Forest	0.9299	max_depth=20, n_estimators=200
Naive Bayes	0.8653	var_smoothing=1e-11
LSTM	N/A	Early stopping on validation loss (patience=5); 2 LSTM layers, Dropout=0.3

The class distribution showed 59,998 Diabetic (60%) and 40,002 Non-Diabetic (40%) records. No missing values were found in any column.

3.3 Data Preprocessing

The preprocessing pipeline comprised five sequential steps. First, two columns, diabetes_risk_score and diabetes_stage, were removed to prevent data leakage, as both were derived from the same underlying information as the target variable. Second, categorical variables were encoded. Label Encoding was applied to binary categorical variables (gender, smoking_status), and One-Hot Encoding was applied to multi-class variables (ethnicity, education_level, income_level, employment_status), expanding the feature matrix to 38 columns. Third, all numerical features were scaled using Min-Max Normalisation to the range [0, 1]:

$$X\_scaled = (X - X\_min) / (X\_max - X\_min$$

The scaler was fitted on training data only to prevent information leakage. Fourth, class imbalance was addressed by applying the Synthetic Minority Over-sampling Technique (SMOTE) to the training set only, balancing both classes at 47,998 records each, yielding a total training set of 95,996 records. Fifth, the dataset was split into training (80%) and testing (20%) sets using stratified sampling:

$$N\_train = 0.80 \times N, \quad N\_test = 0.20 \times N$$

3.4 Model Development and Optimisation

Four machine learning models were developed: XGBoost, Random Forest, Naive Bayes, and LSTM. Table 2 summarises their types and characteristics.

XGBoost minimises the objective function:

$$L(\varphi) = \Sigma\, l(y\_i, \hat{y}\_i) + \Sigma\, \Omega(f\_k) \quad where\ \Omega(f) = \gamma T + ½\lambda||w||^2$$

Random Forest aggregates predictions from B trees by majority voting:

$$\hat{y} = mode\{h\_1(x), h\_2(x), \ldots, h\_B(x)\}$$

Naive Bayes computes the posterior using Bayes' theorem:

$$P(C\_k|x) \propto P(C\_k) \cdot \prod P(x\_i|C\_k)$$

The LSTM network used two LSTM layers with dropout (rate = 0.3) and a sigmoid output neuron, implemented in TensorFlow/Keras. All models were tuned using Grid Search with 5-fold Cross-Validation. Table 3 summarises the best parameters found.

3.5 Model Evaluation

Each model was evaluated on the 20,000-record held-out test set using four standard metrics:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

AUC-ROC was also computed to measure discriminative ability across all thresholds. In addition, subgroup analysis was performed by stratifying the test set by gender, ethnicity, age group (under 40, 40–60, over 60), and income level, with performance metrics computed separately for each subgroup.

3.6 LIME Explainability

LIME was integrated with the best-performing model to generate individual-level prediction explanations. The LIME explainer was initialised from 500 background training samples. For each patient, LIME perturbs the input, passes variations through the model, and fits a weighted local linear surrogate model, solving:

$$x = argmin\_\{g \in G\}\, [L(f, g, \pi\_x) + \Omega(g)]$$

Where f is the original model, g is a simple interpretable surrogate, π_x is a proximity measure, L is the fidelity loss, and Ω(g) is the complexity penalty. Feature weights of the surrogate represent each feature's contribution to the prediction.

3.7 Streamlit Deployment

The complete system was deployed as a browser-based interactive web application using Streamlit. The application included a patient input sidebar organised into six sections (Demographics, Lifestyle, Medical History, Anthropometric, Vital Signs, Laboratory Results), a prediction output panel showing the predicted class with probability score, and a LIME explanation panel with a horizontal bar chart of the top 12 feature contributions. Three models were available for selection: XGBoost, Random Forest, and Naive Bayes. The LIME explainer was rebuilt at runtime from saved background data due to serialisation limitations of internal lambda functions.

3.8 Cross-Dataset Generalisation Testing

XGBoost model was tested on the Pima Indians Diabetes Dataset, a widely used benchmark containing 768 female patient records with 8 clinical features. Because the trained model expects 38 encoded features, the 6 available Pima features were mapped to their closest equivalents in the model's feature space, with the remaining 32 columns filled with zeros. The same MinMaxScaler fitted on the original training data was applied to maintain consistent scaling. Performance was evaluated using the same five metrics used throughout the study. The feature mapping applied was (glucose → glucose_fasting, blood_pressure → diastolic_bp, insulin → insulin_level, bmi → bmi, age → age, and diabetes_pedigree_function → family_history_diabetes). The features pregnancies and skin_thickness had no direct equivalent in the model and were set to zero.

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

Initial analysis confirmed that HbA1c, fasting glucose, post-meal glucose, and insulin level were the features most strongly positively correlated with the target variable, consistent with their established medical roles in diabetes diagnosis. A comparison of mean feature values between diabetic and non-diabetic groups showed noticeably higher values for fasting glucose, HbA1c, BMI, and systolic blood pressure in the diabetic group. Figure 1 shows the class distribution and Figure 2 shows the correlation matrix.

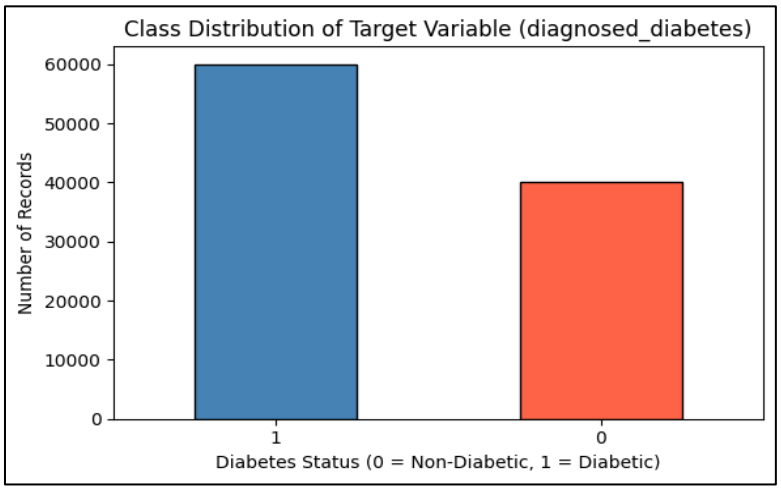

Figure 1: Class Distribution of the Target Variable (diagnosed_diabetes)

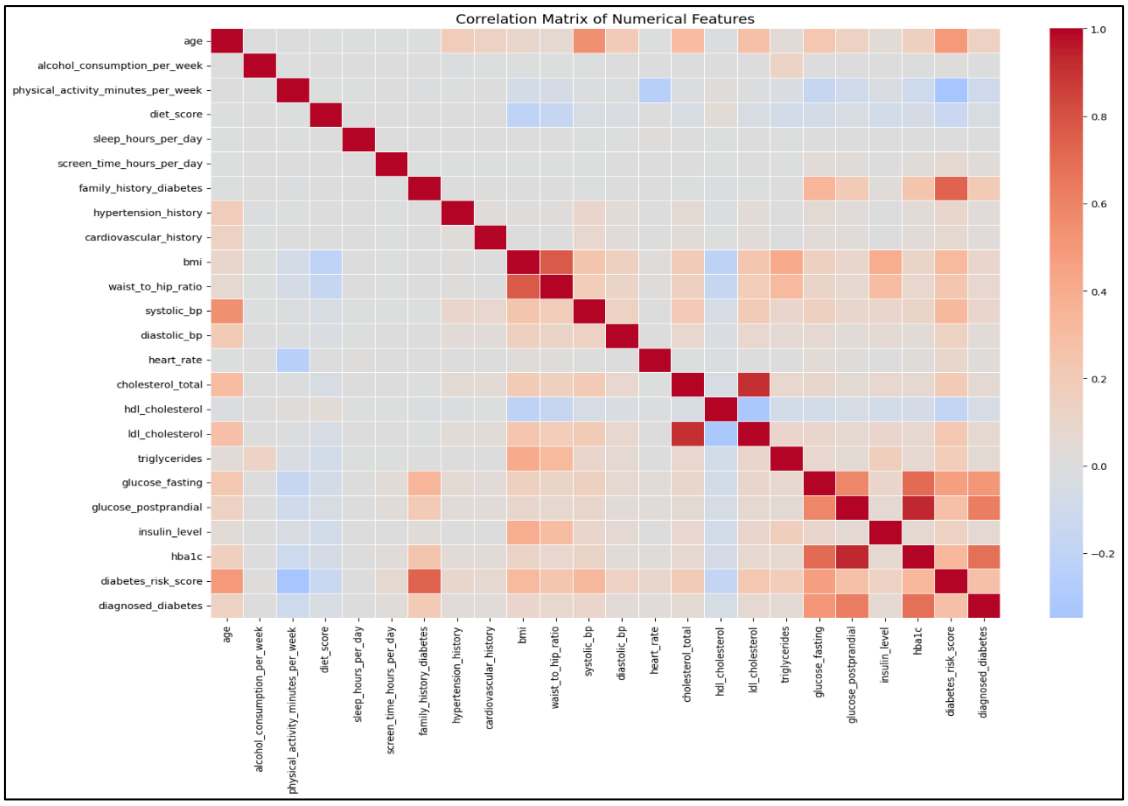

Figure 2: Correlation Matrix of Numerical Features

4.2 Overall Model Performance

Table 4 presents the full performance results for all four models on the 20,000-record held-out test set.

Table 4: Model Performance on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
XGBoost	0.9199	1.0000	0.8665	0.9285	0.9434
Random Forest	0.9198	0.9997	0.8665	0.9284	0.9411
Naive Bayes	0.8542	0.9069	0.8437	0.8742	0.9206
LSTM	0.8638	0.9702	0.7975	0.8754	0.9290

XGBoost achieved the best overall performance, with an accuracy of 91.99%, precision of 1.0000, recall of 0.8665, F1-score of 0.9285, and AUC-ROC of 0.9434. A precision score of 1.0000 means the model made no false-positive predictions, so every patient identified as diabetic was actually diabetic. The recall score of 0.8665 shows that the model correctly identified 86.65% of all diabetic patients. In addition, all four machine learning models achieved accuracy scores above the 70% target set for this study.

Random Forest performed almost identically to XGBoost, with differences of less than 0.0002 across all metrics, confirming that both ensemble methods performed at a comparable level. Naive Bayes achieved 85.42% accuracy and an AUC-ROC of 0.9206, consistent with the findings of Maniruzzaman et al. [4] and Febrian et al. [11], who also found Naive Bayes to be competitive but inferior to ensemble methods on clinical datasets. LSTM achieved 86.38% accuracy but had the lowest recall at 0.7975, consistent with findings in the literature that deep learning models do not necessarily outperform ensemble methods on structured tabular clinical data [7][17]. Figure 3 presents the confusion matrices for all four models, and Figure 4 shows the ROC curves and the model comparison bar chart.

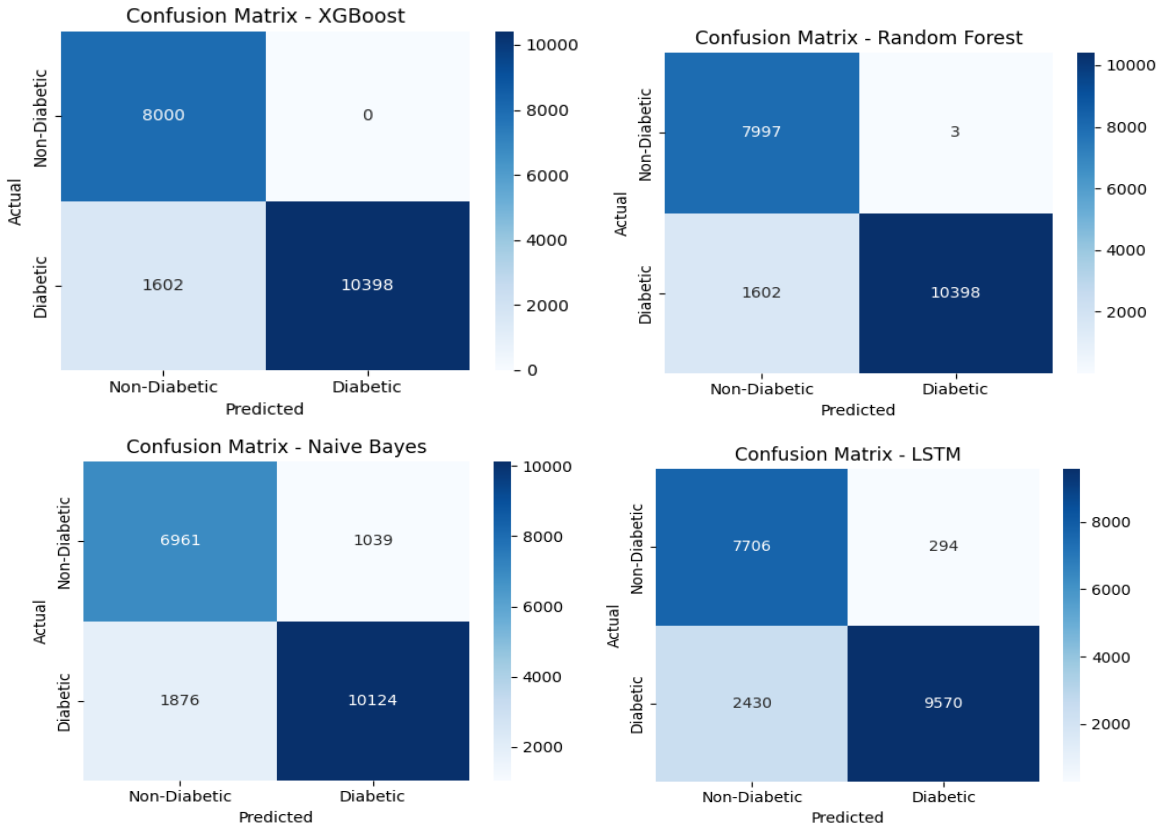

Figure 3: Confusion Matrices

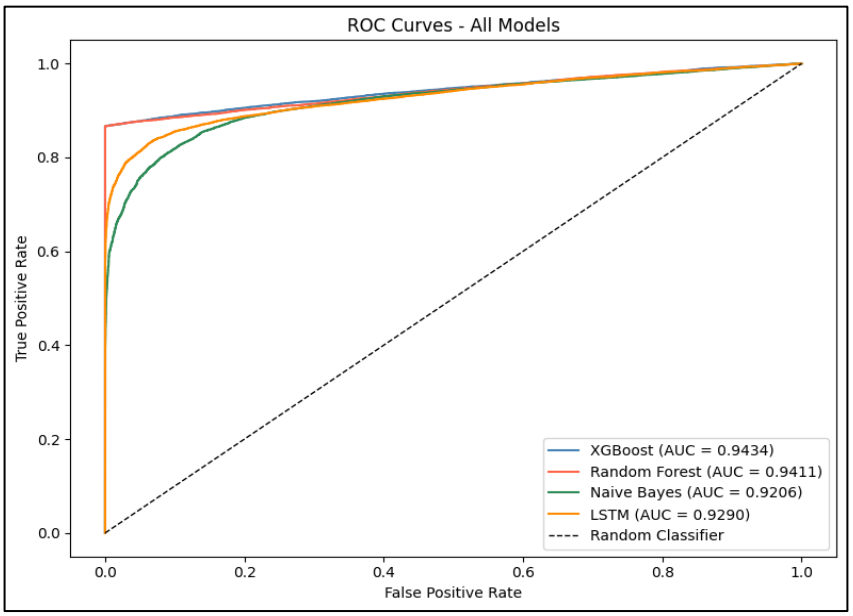

Figure 4: ROC Curves for All Four Models

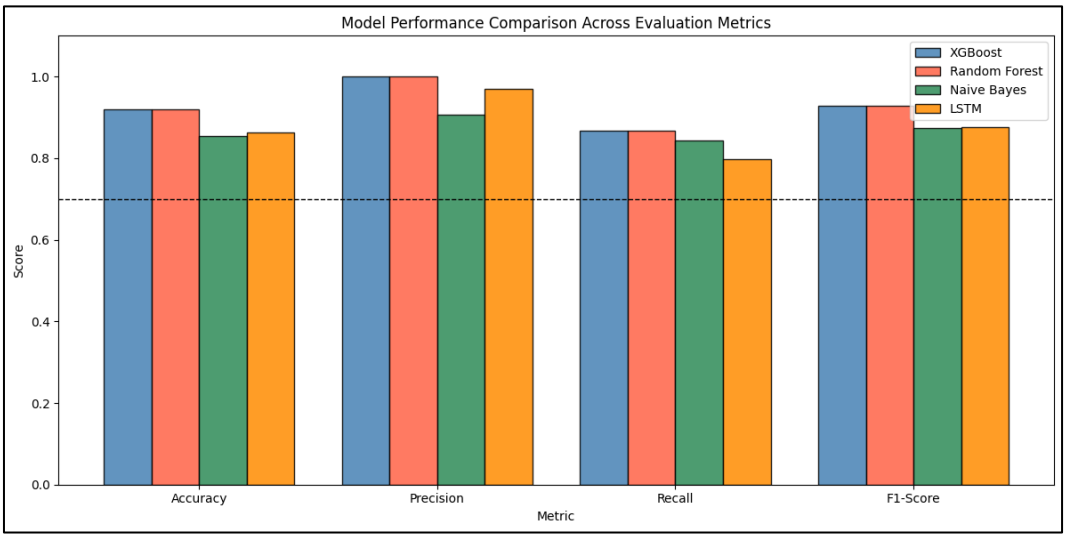

Figure 5: Model Performance Comparison Across All Evaluation Metrics.

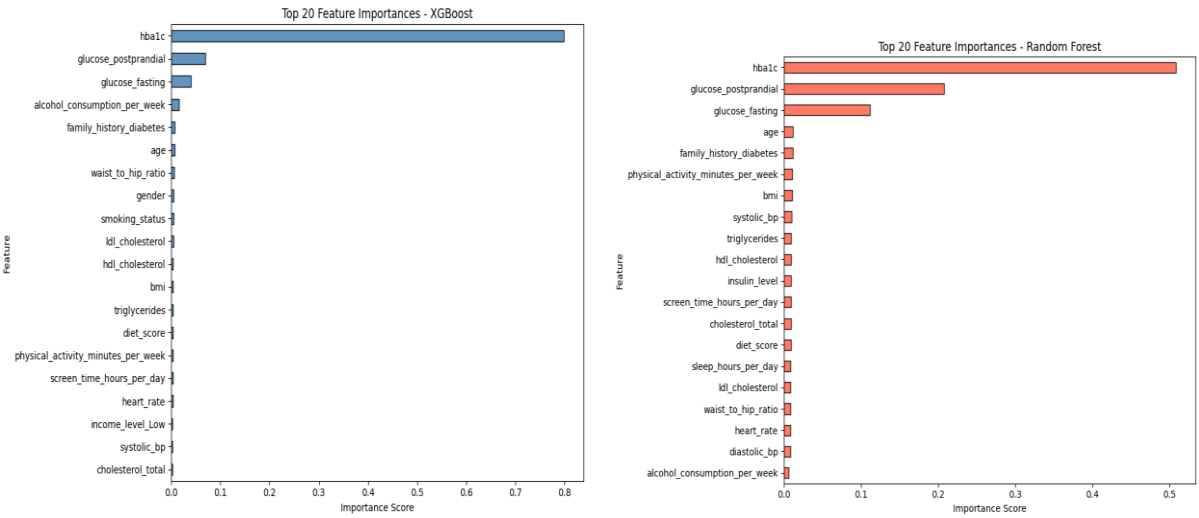

Figure 6: Top 20 Feature Importances between XGBoost (left) and Random Forest (right).

4.3 Feature Importance

Figure 6 shows the top 20 feature importances for XGBoost and Random Forest. Both models consistently identified HbA1c, glucose_fasting, glucose_postprandial, and insulin_level as the four most important features, confirming the primacy of glycaemic biomarkers in diabetes prediction. This consistency across two independent ensemble methods provides strong face validity for the models. Age and BMI also appeared in the top 20, consistent with established clinical risk factors for Type 2 diabetes.

4.4 Demographic Subgroup Analysis

Table 5 presents the XGBoost performance across all demographic subgroups. Performance was highly consistent across all gender groups (F1 range 0.0029), all ethnic groups (F1 range 0.0037), and all income levels (F1 range 0.0064). The most notable variation was in the age group dimension, where patients over 60 achieved the highest F1-score of 0.9420 while patients under 40 achieved the lowest at 0.9116, a gap of 0.0304. This pattern is clinically understandable as diabetes biomarkers are generally less pronounced in younger populations. Notably, precision was 1.0000 for all subgroups, meaning that the model did not make any false positive predictions for any subgroup

4.5 LIME Explainability Results

Figure 8 shows the LIME explanations for a correctly predicted Diabetic patient (probability 0.9916) and a Non-Diabetic patient (probability 0.1757). Table 6 lists the top 10 feature contributions for the diabetic prediction.

HbA1c (weight +0.3703) and fasting glucose (+0.3134) dominated the Diabetic prediction, accounting for the majority of the positive prediction signal. Physical activity (−0.0421) and LDL cholesterol (−0.0306) reduced the predicted probability. The Non-Diabetic patient showed the opposite pattern: low HbA1c and low fasting glucose were the dominant risk-reducing features (see Table 6). These explanations align directly with established clinical guidelines for diabetes diagnosis, providing strong face validity for the model. Figure 9 shows the aggregate LIME importance across 200 randomly sampled test patients, confirming that HbA1c, glucose_fasting, and glucose_postprandial were the most consistently influential features across the population, while demographic features appeared at the bottom of the ranking.

Table 5: XGBoost Demographic Subgroup Performance

Subgroup	Count	Accuracy	Precision	Recall	F1-Score
Gender - Male	9,616	0.9209	1.0000	0.8690	0.9299
Gender - Female	10,034	0.9192	1.0000	0.8639	0.9270
Gender - Other	350	0.9143	1.0000	0.8690	0.9299
Ethnicity - White	9,079	0.9195	1.0000	0.8665	0.9285
Ethnicity - Black	3,519	0.9199	1.0000	0.8685	0.9296
Ethnicity - Hispanic	4,110	0.9221	1.0000	0.8670	0.9288
Ethnicity - Asian	2,326	0.9175	1.0000	0.8643	0.9272
Age - Over 60	5,195	0.9251	1.0000	0.8903	0.9420
Age - 40 to 60	9,283	0.9204	1.0000	0.8665	0.9285
Age - Under 40	5,522	0.9142	1.0000	0.8376	0.9116
Income - Low	2,960	0.9233	1.0000	0.8712	0.9312
Income - Middle	7,030	0.9215	1.0000	0.8678	0.9292
Income - High	1,001	0.9181	1.0000	0.8601	0.9248

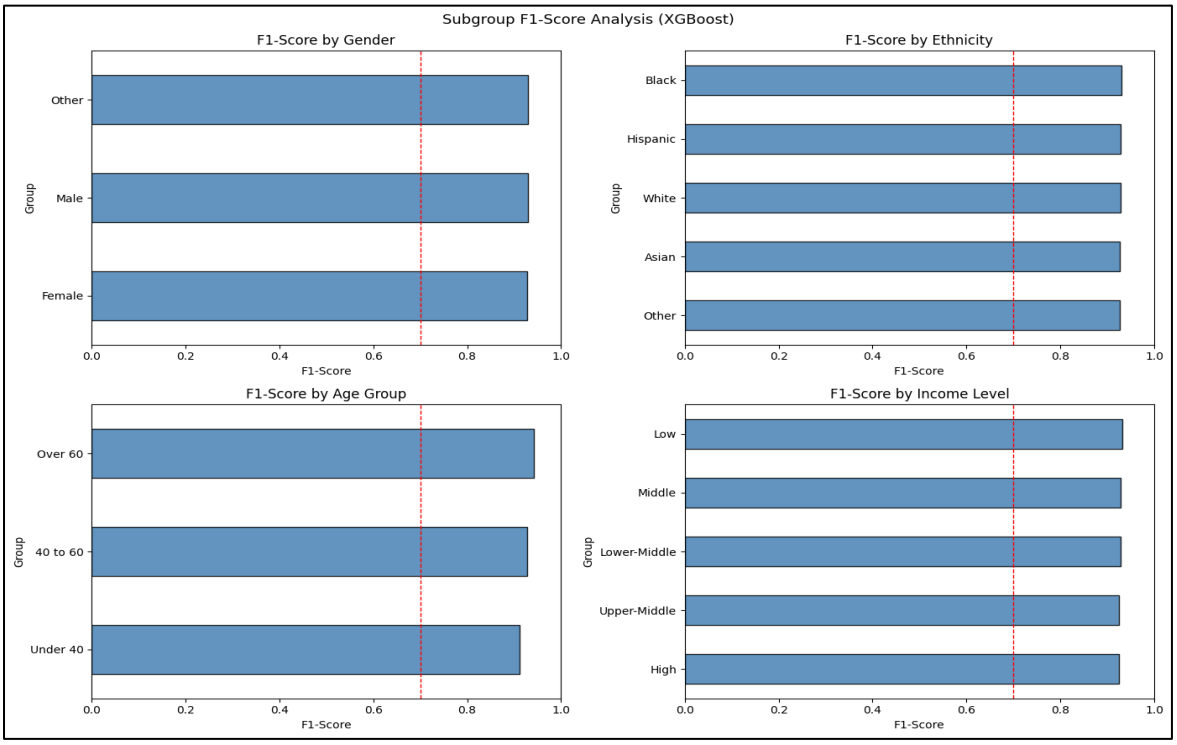

Figure 7: Subgroup F1-Score Analysis Across All Demographic Dimensions.

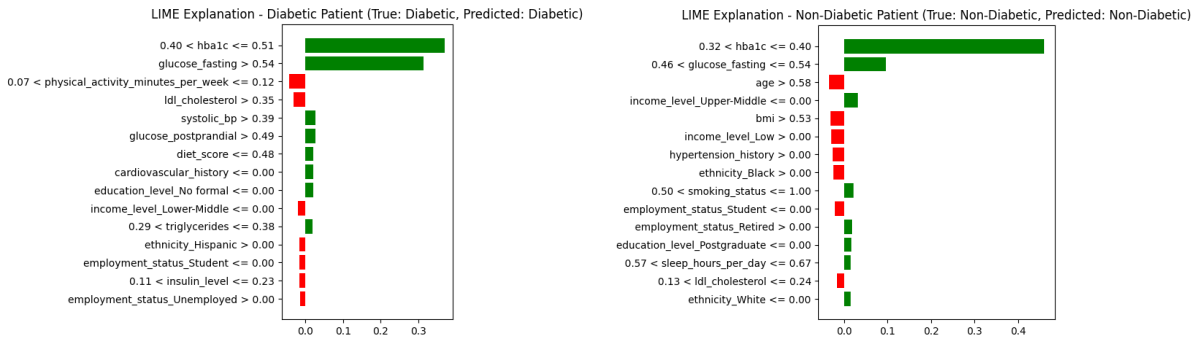

Figure 8: LIME Explanations (Diabetic patient vs Non-Diabetic patient)

Table 6: Top 10 LIME Feature Contributions for a Diabetic Patient (XGBoost)

Feature Condition	LIME Weight	Effect on Prediction
0.40 < hba1c <= 0.51	+0.3703	Increases diabetes risk (dominant feature)
glucose_fasting > 0.54	+0.3134	Increases diabetes risk
0.07 < physical_activity <= 0.12	−0.0421	Decreases diabetes risk
ldl_cholesterol > 0.35	−0.0306	Decreases diabetes risk
systolic_bp > 0.39	+0.0280	Increases diabetes risk
glucose_postprandial > 0.49	+0.0275	Increases diabetes risk
diet_score <= 0.48	+0.0215	Increases diabetes risk
cardiovascular_history <= 0.00	+0.0214	Increases diabetes risk
education_level_No formal <= 0.00	+0.0204	Increases diabetes risk
income_level_Lower-Middle <= 0.00	−0.0198	Decreases diabetes risk

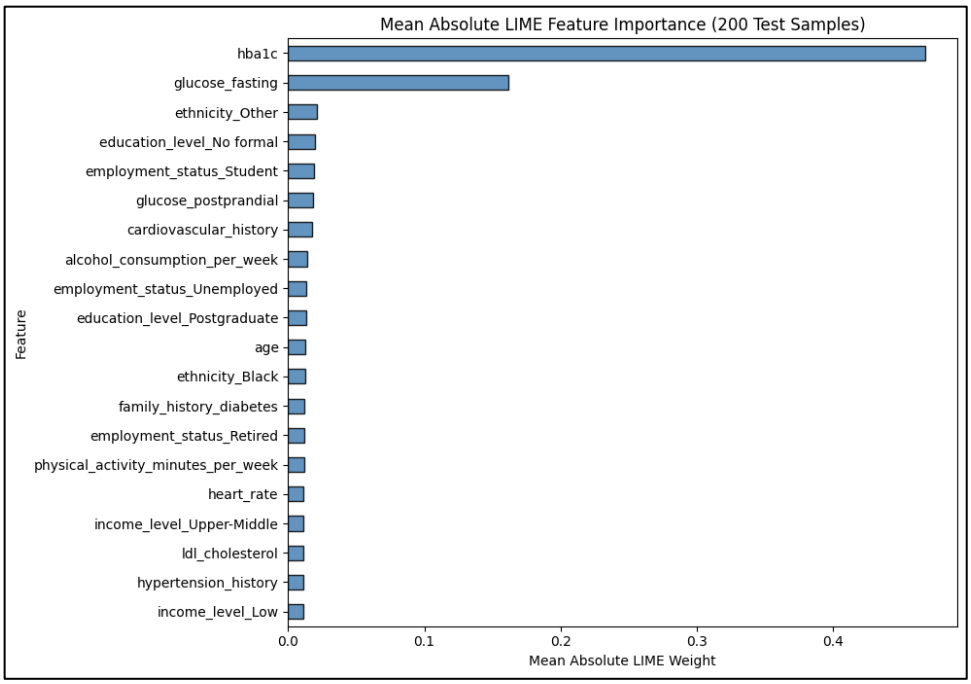

Figure 9: Mean Absolute LIME Feature Importance Across 200 Test Patients

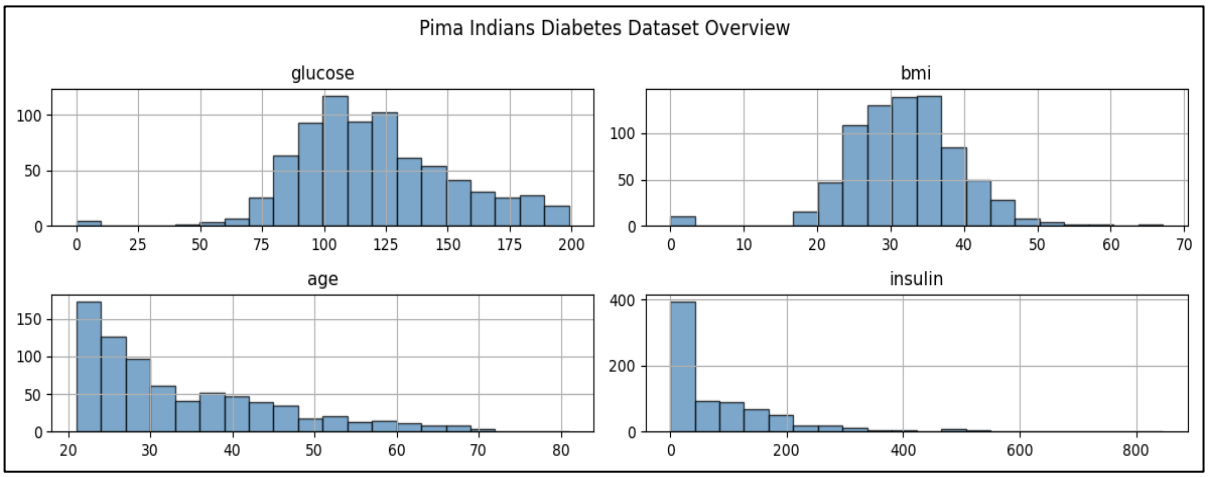

Figure 10: Pima Indians Dataset

4.6. Extended Evaluation

4.6.1. Cross-Dataset Generalisation Test

A more comprehensive evaluation was done across multiple datasets and scenarios, the trained XGBoost model was tested on the Pima Indians Diabetes Dataset, which is one one of the most widely used benchmark datasets in diabetes prediction research, containing 768 records with 8 clinical features. The model was trained exclusively on the Diabetes Health Indicators Dataset and was never exposed to any Pima data during training, making this a genuine out-of-sample generalisation test. Figure 10 shows the Pima dataset class distribution and feature distributions. The dataset contains 500 non-diabetic (65.1%) and 268 Diabetic (34.9%) records, representing a class imbalance different from the original training distribution. Only 6 of the model's 38 encoded features could be mapped from the Pima dataset. The remaining 32 feature columns were set to zero. The same MinMaxScaler fitted on the original training data was applied without refitting. Table 7 presents the side-by-side performance comparison between the original dataset evaluation and the Pima cross-dataset test.

Table 7: Cross-Dataset Generalisation Comparison (XGBoost)

Dataset	Records	Features	Accuracy	F1-Score	AUC-ROC
Diabetes Health Indicators (Original)	20,000	38	0.9199	0.9285	0.9438
Pima Indians (Cross-Dataset Test)	768	6 mapped	0.7253	0.6279	0.7969

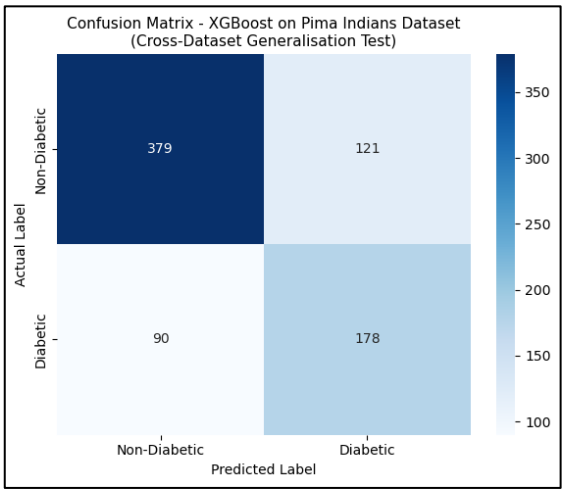

Figure 11: Confusion Matrix on Pima Indians Dataset

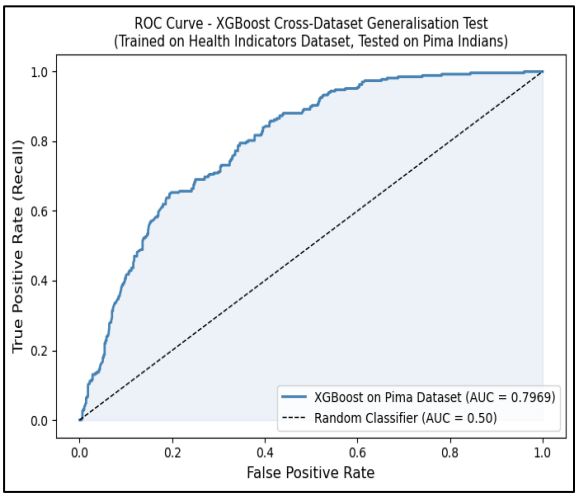

Figure 12: ROC for XGBoost (cross-dataset generalization test)

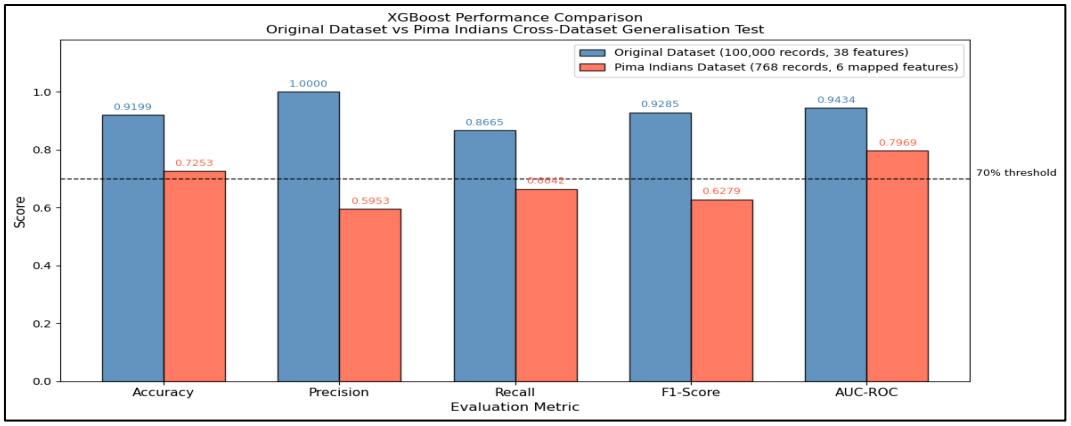

Figure 13: XGBoost Performance Comparison (Original Dataset vs Pima Indians Cross-Dataset Test)

The model achieved 72.53% accuracy and an AUC-ROC of 0.7969 on the Pima dataset. While these results are lower than the original dataset performance, three important observations support the validity of this outcome. First, the model had access to only 6 of its 38 trained features during this test; 32 feature columns were absent and filled with zeros. This represents a severe reduction in available information, making the retained performance level notable rather than disappointing. Second, the AUC-ROC of 0.7969 is substantially above the 0.50 baseline of a random classifier, confirming that the model retained genuine discriminative ability between diabetic and non-diabetic patients even under these constrained conditions. Third, the accuracy of 72.53% still exceeds the 70% research target established for this study, demonstrating that the core predictive signal generalises across datasets. These results demonstrate that the clinical features the model learned-glucose levels, BMI, insulin level, and age carry a transferable predictive signal that extends beyond the specific dataset it was trained on. This finding strengthens the argument for the model's clinical utility and partially addresses the concern that performance on a single synthetic dataset may not reflect real-world generalisability.

4.7 Discussion

The results demonstrate that ensemble ML models, particularly XGBoost and Random Forest, achieve high predictive accuracy for diabetes detection when trained on well-prepared, class-balanced data. XGBoost's F1-score of 0.9285 and AUC-ROC of 0.9434 outperform reported results from several prior studies, including Maniruzzaman et al. [4] (90.62% accuracy, 0.95 AUC) and are comparable to the ensemble results of Abnoosian et al. [5] (0.9887 accuracy) and Hasan et al. [6] (AUC 0.950), while operating on a substantially larger and more demographically diverse dataset.

The precision of XGBoost (1.0000) is notable, with the model producing no false positive predictions. As this reflects strong discriminative ability, it comes at the cost of a recall of 0.8665, meaning 13.35% of actual diabetic cases were missed. In clinical screening contexts where missing a true diabetic patient carries serious health consequences, a higher recall threshold may be appropriate, suggesting that adjusting the classification threshold toward a lower value could increase recall at the cost of some precision.

The LSTM model underperformed relative to ensemble methods, confirming the finding reported by Zhou et al. [7] and Beghriche et al. [17] that deep learning does not always outperform well-tuned ensemble models on structured tabular data, particularly when the data lacks the sequential or temporal structure that LSTM is architecturally designed to exploit.

The subgroup analysis results support the inclusivity objective of this study. Variation across gender, ethnicity, and income dimensions was below 0.007 in all cases. The larger variation in the age group analysis (0.0304) is clinically expected rather

than indicative of algorithmic bias. The consistent precision of 1.0 across all subgroups ensures that no demographic group was subjected to false alarms, an important equity consideration in clinical screening contexts.

The LIME explanations consistently identified HbA1c and fasting glucose as the dominant predictors of diabetes, aligning with established diagnostic criteria. This alignment between XAI outputs and clinical knowledge constitutes face validity for the model and supports clinician trust. The integration of LIME with the Streamlit interface bridges the deployment gap identified in the literature [10], providing a practical tool that non-technical users can operate without specialised expertise.

The cross-dataset generalisation test on the Pima Indians Diabetes Dataset provided additional evidence that the model's learned representations extend beyond its training distribution. Achieving 72.53% accuracy and AUC-ROC of 0.7969 using only 6 of 38 features confirms that the core clinical predictors (glucose, BMI, insulin, and age) carry transferable discriminative signal that is not unique to the synthetic training dataset.

5. CONCLUSION

This study presented the design and development of an inclusive diabetes detection system that combines machine-learning classification, LIME-based explainability, and Streamlit deployment. A 100,000-record synthetic dataset was preprocessed using encoding, Min-Max normalisation, SMOTE, and stratified splitting. Four models XGBoost, Random Forest, Naive Bayes, and LSTM, were trained and tuned using Grid Search with 5-fold cross-validation. XGBoost achieved the best overall performance with 91.99% accuracy, an F1-score of 0.9285, and an AUC-ROC of 0.9434, clearly exceeding the 70% research target.

Demographic subgroup analysis confirmed equitable performance across gender, ethnicity, age group, and income level, with minimal variation across most dimensions. LIME generated clinically meaningful individual-level explanations, consistently identifying HbA1c and fasting glucose as the dominant predictors of diabetes, aligned with established clinical knowledge. The Streamlit web application successfully deployed all system components in an accessible, user-facing interface providing real-time predictions and transparent feature-level explanations.

Cross-dataset generalisation testing on the Pima Indians Diabetes Dataset confirmed that the model retains meaningful predictive ability 72.53% accuracy and AUC-ROC of 0.7969 even when restricted to only 6 of its 38 trained features, demonstrating transferable clinical signal beyond the original training distribution.

The key contributions of this study are: (i) a comprehensive, inclusive ML system evaluated across four demographic dimensions; (ii) LIME explainability integrated and validated for clinical meaningfulness; (iii) interactive Streamlit deployment bridging the research-to-practice gap; and (iv) a demonstration that high predictive accuracy, transparent explanation, and demographic equity can be achieved simultaneously using open-source tools.

Future work should evaluate the system on real patient datasets from diverse clinical settings, explore fairness-constrained modelling to formally address demographic disparities, and investigate multiclass prediction extending beyond binary classification to Type 1, Type 2, and gestational diabetes subtypes. Integration of the system with electronic health records and regulatory validation for clinical deployment are also important directions.

6. ACKNOWLEDGMENT

The authors acknowledge the Almighty God for the wisdom and strength to complete this research. Sincere appreciation is also extended to the various academic repositories, journals, and research platforms that offered valuable knowledge and studies, which greatly influenced the development and direction of this work.

7. REFERENCES

[1] International Diabetes Federation. IDF Diabetes Atlas, 10th edition. Brussels, Belgium: IDF; 2021. Available at: https://diabetesatlas.org

[2] Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), p.7.

[3] Rastogi, R. and Bansal, M., 2023. Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, p.100605.

[4] Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), p.7.

[5] Abnoosian, K., Farnoosh, R. and Behzadi, M.H., 2023. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics, 24(1), p.337.

[6] Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516–76531.

[7] Zhou, H., Myrzashova, R. and Zheng, R., 2020. Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking, 2020(1), p.148.

[8] Mahajan, P., Uddin, S., Hajati, F. and Moni, M.A., 2023. Ensemble learning for disease prediction: A review. Healthcare, 11(12), p.1808. MDPI.

[9] Cheungpasitporn, W., Athavale, A., Ghazi, L., Kashani, K.B., Colicchio, T., Koyner, J.L., Chen, J., Ix, J.H., Nadkarni, G. and Neyra, J.A., 2026. Transforming Nephrology Through Artificial Intelligence: A State-of-the-Art Roadmap for Clinical Integration. *Clinical Kidney Journal*, p.sfag004.

[10] Akther, F., Begum, M., Mahmud, T., Boltayev, A., Hanip, A. and Hossain, M.S., 2025. Streamlit-based AI for multi-disease prediction. In 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), pp.1622–1628. IEEE.

[11] Febrian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M. and Yunanda, R., 2023. Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, pp.21–30.

[12] Arumugam, K., Naved, M., Shinde, P.P., Leiva-Chauca, O., Huaman-Osorio, A. and Gonzales-Yanac, T., 2023.

Multiple disease prediction using machine learning algorithms. Materials Today: Proceedings, 80, pp.3682–3685.

[13] Bukhari, M.M., Alkhamees, B.F., Hussain, S., Gumaei, A., Assiri, A. and Ullah, S.S., 2021. An improved artificial neural network model for effective diabetes prediction. Complexity, 2021(1), p.5525271.

[14] Ihnaini, B., Khan, M.A., Khan, T.A., Abbas, S., Daoud, M.S., Ahmad, M. and Khan, M.A., 2021. A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning. Computational Intelligence and Neuroscience, 2021(1), p.4243700.

[15] Krishnamoorthi, R., Joshi, S., Almarzouki, H.Z., Shukla, P.K., Rizwan, A., Kalpana, C. and Tiwari, B., 2022. A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, 2022(1), p.1684017.

[16] Singh, S., Agarwal, S., Singh, V., Hazela, B., Dubey, V. and Singh, P., 2025. Comparative analysis of machine learning algorithms for multiple disease prediction model with an optimized scalable deployment. In 2025 International Conference on Electronics, AI and Computing (EAIC), pp.1–6. IEEE.

[17] Beghriche, T., Djerioui, M., Brik, Y., Attallah, B. and Belhaouari, S.B., 2021. An efficient prediction system for diabetes disease based on deep neural network. Complexity, 2021(1), p.6053824.

[18] Sharma, A., Guleria, K. and Goyal, N., 2021. Prediction of diabetes disease using machine learning model. In International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020, pp.683–692. Singapore: Springer Singapore.