Effective Diabetes Prediction: Integrating Ensemble Learning with LIME for Robust Results

Syeda Aqsa Habib Gilani
Institute of Information Technology
Quaid-i-Azam University
Islamabad, Pakistan
aqsagilani@iit.qau.edu.pk

Madiha Haider Syed
Institute of Information Technology
Quaid-i-Azam University
Islamabad, Pakistan
madiha@qau.edu.pk

Adeel Anjum
Institute of Information Technology
Quaid-i-Azam University
Islamabad, Pakistan
aanjum@qau.edu.pk

2024 International Conference on Frontiers of Information Technology (FIT) | 979-8-3315-1050-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/FIT63703.2024.10838461

***Abstract*—In this study, we developed a robust framework for predicting diabetes using a combination of Machine Learning (ML) algorithms, including Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, Decision Tree (DT), Neural Networks, and K-means clustering. We utilized a comprehensive dataset encompassing medical and demographic features. Our methodology included meticulous data preprocessing, extensive exploratory data analysis (EDA), and the creation of interaction features to capture complex relationships. We applied Recursive Feature Elimination (RFE) to select significant features and employed ensemble learning techniques using a Voting Classifier to enhance model performance. Our Neural Network model demonstrated superior predictive capability, achieving an accuracy of 97.21% and an ROC-AUC score of 0.9769. To ensure model interpretability, we implemented Local Interpretable Model-agnostic Explanations (LIME), which provided clear insights into individual feature contributions. This approach not only ensures high accuracy in diabetes prediction but also enhances transparency and trustworthiness, making the results valuable for clinical decision-making.**

***Index Terms*—Diabetes Prediction, Machine Learning, Explainable AI, Model Interpretability, LIME**

I. Introduction

Diabetes mellitus is a chronic condition that significantly impairs the body's ability to produce or effectively utilize insulin, resulting in elevated blood glucose levels. This condition leads to severe health complications, including heart disease, kidney failure, and nerve damage. The prevalence of diabetes is alarmingly high, with the International Diabetes Federation estimating that 537 million individuals lived with diabetes in 2021, contributing to 6.7 million deaths globally [1].

Diabetes is caused by many risk factors, which include but are not limited to obesity, hypertension, genetic causes, increase in age, and engagement in lifestyles. The following paper will discuss and collect some of the risk factors on how people develop diabetes [2]. According to, early diagnosis and intervention make the difference in grim health consequences that come with prediabetes [3]. Machine learning in the prediction of diabetes In a nutshell caught on for promising enhanced diagnostic early detection and individualized treatment.

Kumar et al. [4] used a Random Forest algorithm for diabetes prediction with 90% accuracy, highlighting ML's potential. Sneha et al. [5] showed effective feature selection, achieving 79.8% accuracy with a Random Forest model. Hennebelle et al. [6] applied IoT-edge-cloud computing with ML for type 2 diabetes diagnosis but did not detail model predictability. Tasin et al. [7] developed a diabetes prediction system using class imbalance management, domain adaptation, and explainable AI (LIME and SHAP) with XGBoost and ADASYN, achieving 81% accuracy.

In this study, we aim to advance the predictive capabilities of diabetes diagnosis by employing a comprehensive ML framework and explainable AI techniques. We utilize the diabetes prediction dataset, which includes medical and demographic features such as age, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

Our methodology integrates several ML algorithms, including RF, SVM, Naive Bayes, DT, and Neural Networks (NN), with the application of **LIME** for model interpretability. Feature selection was done using the RFE technique which reduced the feature list to the most important features hence affirming the importance of all the used features. To improve generalization and reliability of the solution, we applied ensemble learning that is the Voting Classifier that is based on multiple models' advantages. From the Neural Network point of view, our model possesses significantly better predictive power, which is 97%, the sensitivity was 21% and we achieved a ROC-AUC score of 0.9769, indicating the area under the receiver operating characteristic curve.

To improve model interpretability, we applied LIME, which gave an understanding of the features' contributions to predictions and better compliance with the transparent model. By applying this approach, it was possible to pay much attention to particular medical scores as predictors of diabetes, which makes the model's decisions sufficiently transparent for clinicians.

The structure of this paper is as follows: Section II outlines the approach used in this research, detailing data cleansing, data inspection, feature construction, and modeling. Section III presents the experimental results and model analysis, as well as LIME results for interpretability, summarizing their clinical significance. Section IV provides an in-depth discussion of the implications of these findings and evaluates the model's performance within a clinical context. Finally, Section V concludes the study and suggests future research directions.

II. Related Work

In recent years, extensive research has focused on advancing diabetes prediction through machine learning and explainable AI (xAI), emphasizing enhanced model precision and interpretability using diverse datasets and algorithms.

Another paper highlighted the use of boosting techniques as feature ensemble learning to enhance diabetes prediction, with data augmentation, normalization, and hyperparameter tuning improving performance [8]. Similarly, a study on ontology-based machine learning showed that well-formatted medical datasets support the training of more effective predictive models [9].

One of the advances made in this research work is the discovery of a machine learning approach to dealing with the data imbalance in the PIMA dataset. This framework applied the intricate data balancing procedure like SMOTEENN for data balancing and suggested extensive preprocessing strategies for managing lost values; profoundly enhanced the steadiness as well as the versatility of the predictive models [10].

The researchers employed a rich set of data preprocessing methods and suggested this new model to increase the classification accuracy. This study highlighted the necessity of data enhancement and revealed the possibilities of MSA for clinics [11].

Specifically, SHAP and LIME techniques have been widely used in diabetes prediction models to enhance the interpretability of results. Such techniques enhance the models' interpretability to reveal the contributions of the individual features toward the predictions being made, making the models reliable for use in clinical decisions [12]. For example, a paper using SHAP and LIME showed that both of them can help explain different ML models' decision-making, contributing to the execution of models with extremely high accuracy while ensuring practicality [13].

Also, there is an inclination towards invulnerable data for the prediction of diabetes. Studies utilizing data like physical activity, and other lifestyle markers have proved useful in formulating accurate, cheap, and more efficient models. These approaches' goal is to increase screening of diabetes to a bigger population thus reducing the strict use of laboratory tests [14].

Recent advances in deep learning, particularly using CNNs and RNNs, have boosted diabetes prediction model performance by capturing complex data patterns. Similar techniques applied to loan acceptance/rejection highlight the broad applicability of these models in automated decision-making [15].

Moreover, research has also evaluated the hybrid models that compile several ML algorithms to improve the prediction capability. For instance, a Naive Bayes, Random Forest, Decision tree, XGBoost Light GBM's ensemble presented notably better accuracy and resilience when contrived upon a newly tagged diabetes set from a South Asian country [16].

These studies highlight advancements in diabetes prediction using XAI methods. Leveraging diverse datasets and advanced techniques, researchers are improving predictive accuracy and model interpretability in diabetes care.

III. Methodology

A. Data Collection and Preprocessing

For this study, we utilized the Diabetes Prediction Dataset available on Kaggle [17]. This dataset encompasses a variety of medical and demographic factors, including age, gender, BMI, hypertension, heart disease status, smoking history, HbA1c levels, and blood glucose levels. With 100,000 entries and 10 features, it represents a cross-sectional collection. To maintain a balanced class distribution, we split the dataset into training and testing subsets using a stratified shuffle split approach.

To standardize the features, each feature x_i was transformed using the equation:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

where μ is the mean, and σ is the standard deviation of x_i. This standardization puts all features on a comparable scale, which is an essential precondition for the effectiveness of many machine learning algorithms.

B. Data Exploration

Before the data preprocessing phase, we conducted exploratory data analysis (EDA) to examine the distribution and relationships within the dataset. This process included:

- Generating scatter plots, such as BMI vs. Age, split by diabetes status, to visualize the distribution of key features.
- Calculating and analyzing the correlation coefficients between features to identify potential multicollinearity issues.
- Creating a Pearson correlation matrix heatmap to visualize the relationships between features.
- Checking for outliers using box plots for numerical features like age, BMI, HbA1c level, and blood glucose level.
- Identifying and removing outliers based on z-scores to ensure data quality.
- Generating summary statistics before and after removing outliers to compare and understand the impact of outlier removal.
- Checking for missing values and handling them appropriately to maintain data integrity.

These steps helped in uncovering patterns and insights critical for feature engineering and model building.

C. Feature Engineering

We created interaction features like 'age_bmi_interaction' and 'glucose_hba1c_interaction' to capture the non-linear relationship and improve the predictive power of our models. Mathematically, the new interaction features look like:

$$\text{age\_bmi\_interaction} = \text{age} \times \text{bmi} \tag{2}$$

$$\text{glucose\_hba1c\_interaction} = \text{glucose} \times \text{HbA1c} \tag{3}$$

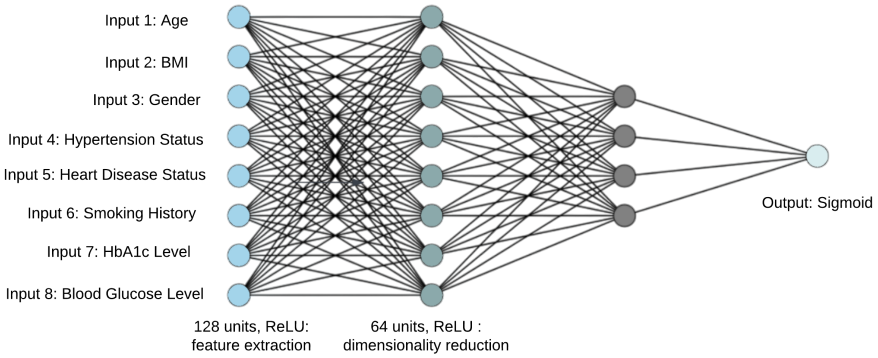

Fig. 1: Neural Network with 8 Inputs and Two Hidden Layers

D. Recursive Feature Elimination (RFE)

We employed Recursive Feature Elimination (RFE) with a logistic regression estimator to identify key features, progressively removing less impactful ones to boost model performance. The logistic regression model is defined as:

$$P(Y = 1|X) = \frac{1}{1 + \exp -(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_p x_p)} \quad (4)$$

RFE identifies the most relevant features by progressively evaluating and eliminating less important ones.

E. Ensemble Learning

We employed ensemble learning with a Voting Classifier for robust and accurate predictions, using a weighted voting method where models contribute based on their performance.

$$\hat{y} = \text{argmax} \left(\sum_{m=1}^{M} w_m \hat{y}_m \right) \quad (5)$$

Equation 5 gives the ensemble prediction, including the contribution of M models for the final result. Each $\hat{y}_m$ here is the prediction by the mth model. The corresponding weight factor is given by w_m: the higher the general reliability of a model, the higher the weighting of this model in the final prediction.

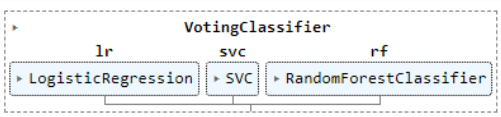

Fig. 2: Voting Classifier combining Logistic Regression, SVC, and Random Forest

Figure 2 shows the voting classifier, a technique that takes into account the strengths of different models to ensure an improved performance on the whole. In this, it is checked that the models which are better predictors get more importance in making a decision on the final prediction in order to enhance the working of the ensembles.

F. Model Building and Evaluation

We carried out the following machine learning techniques for the prediction of diabetes: logistic regression, random forest, gradient boosting, SVM, and neural networks. We judged the performance of the models based on accuracy, ROC-AUC, precision, recall, and F1 score. Each of these metrics was given by the classification report for a class, while the ROC-AUC score was determined as:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}\, d\text{FPR} \quad (6)$$

with TPR indicating the true positive rate and FPR signifying the false positive rate.

G. Neural Network Implementation

A neural network model was implemented using TensorFlow and Keras, comprising dense layers with ReLU activation and dropout for regularization. As shown in Figure 1, the architecture includes 8 input features, two hidden layers, and an output layer for binary classification. The model achieved 97.14% accuracy and a ROC-AUC score of 0.9764. The neural network's loss function is binary cross-entropy:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \quad (7)$$

Here, y_i indicates the actual label, and p_i is the predicted probability for observation i. This loss function penalizes incorrect predictions more heavily, guiding the model to improve its accuracy.

In this context, y_i represents the true label, and p_i denotes the predicted probability for each observation i.

H. Explainable AI Techniques

We employed LIME to elucidate model predictions, offering localized interpretable explanations for individual features that affect each decision.

The model g works as a simpler version of the main model f for making predictions at x:

$$\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Gamma(g) \quad (8)$$

In this equation, πx means the input feature x weighted by π, showing its importance in the prediction. The term $\mathcal{L}(f, g, \pi_x)$ measures how well g matches f near x, while $\Gamma(g)$ is a penalty term that keeps g simple and easy to understand.

IV. Experimental Results

A. Exploratory Data Analysis

We conducted a detailed data exploration and visualization to understand the distribution and relationships between various features in the dataset. While a comprehensive set of visualizations was generated, we present here only the most pertinent and illustrative ones to highlight key insights.

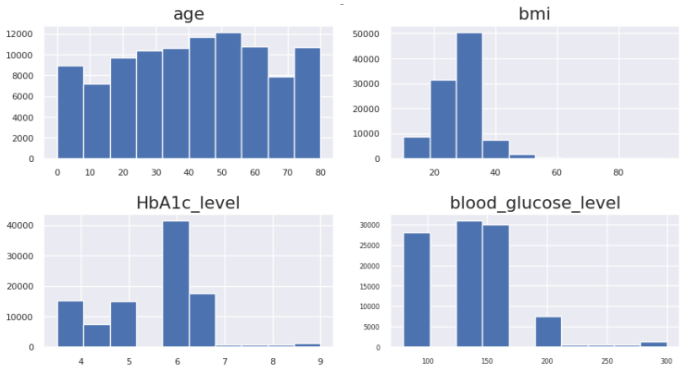

Fig. 3: Histograms of key features: (a) age, (b) BMI, (c) HbA1c level, and (d) blood glucose level. The y-axis in each graph represents the frequency of observations within each bin

Figure 3 shows an age peak at 40-60, highlighting a middle-aged concentration linked to higher diabetes risk. The BMI histogram centers at 30-40, indicating many overweight/obese cases. The HbA1c peak at level 6 suggests numerous prediabetic/diabetic cases, while the 100-150 blood glucose range emphasizes the dataset's focus on these groups.

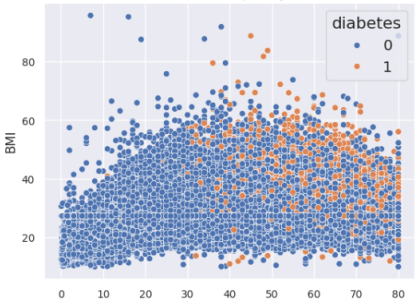

Fig. 4: Scatter plot of Age and BMI, colored by Diabetes status

A scatter plot in Figure 4 depicts the relationship between age, BMI, and diabetes status, providing insights into how these variables interact with each other.

Figure 5 shows the correlation matrix, revealing relevant features for predicting diabetes using Pearson correlation coefficients. The coefficient r between variables X and Y is calculated as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{9}$$

The strong correlation between HbA1c level and blood glucose level with diabetes, along with the moderate positive correlation of age, highlights their predictive importance, with the heatmap aiding in feature selection and understanding data structure.

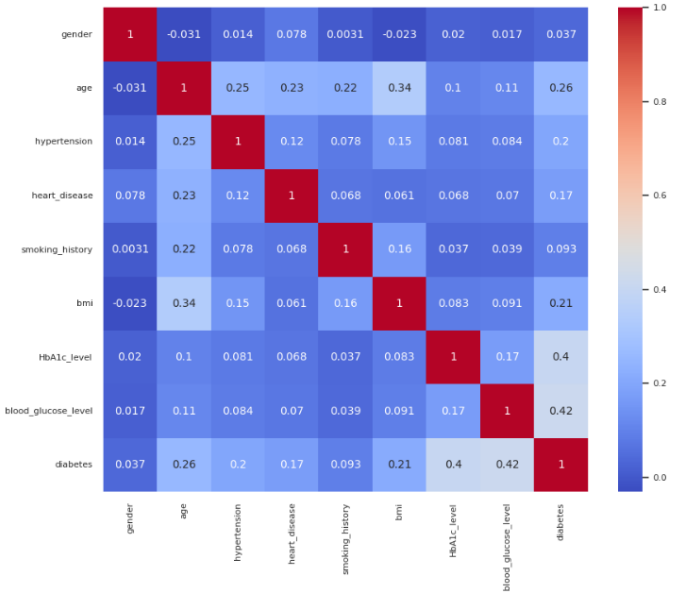

Fig. 5: Correlation matrix of the features

B. Model Evaluation

TABLE I: PERFORMANCE METRICS OF ML ALGORITHMS

Classifier	AUC	F1	Precision	Recall
Logistic Regression	0.902714	0.744788	0.895350	0.685034
SVM	0.961608	0.762882	0.972759	0.691452
Naïve Bayes	0.922973	0.742880	0.714418	0.785918
Decision Tree	0.855623	0.850585	0.846082	0.855240
Random Forest	0.964334	0.891328	0.960444	0.842833
Neural Networks	0.969131	0.858865	0.953888	0.800998
K-means Clustering	0.914426	0.806095	0.949200	0.739028

As depicted in Table I, the models' performance was evaluated based on multiple metrics. Precision (P), Recall (R), F1-Score ($F1$), Support, and ROC-AUC score were the key metrics used.

Precision (P) is defined as the ratio of true positives to the total predicted positives:

$$P = \frac{TP}{TP + FP} \tag{10}$$

Recall (R), also known as Sensitivity, is the ratio of true positives to the total actual positives:

$$R = \frac{TP}{TP + FN} \tag{11}$$

The F1-Score ($F1$) represents the harmonic mean of Precision and Recall:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{12}$$

Our findings indicate that the Neural Network model achieved the highest ROC-AUC score of 0.9691, as shown in Table I, while the Decision Tree classifier recorded the lowest score at 0.8556.

C. Ensemble Learning: Voting Classifier Results

Ensemble learning, using methods like the Voting Classifier, combines predictions from Logistic Regression, SVM, and Random Forest to enhance accuracy and robustness.

TABLE II: VOTING CLASSIFIER RESULTS INCLUDING PRECISION, RECALL, F1-SCORE, AND ROC-AUC

Metric	Class 0	Class 1	Macro Avg	Weighted Avg
Precision	0.97	0.95	0.96	0.97
Recall	1.00	0.64	0.82	0.97
F1-Score	0.98	0.76	0.87	0.96
ROC-AUC	-	-	-	0.97553
Support	18292	1708	20000	20000

Table II summarizes the performance of the Voting Classifier, including precision, recall, F1-score, and an ROC-AUC score of 0.97553. This high ROC-AUC underscores the model's strong predictive capability and effective class differentiation.

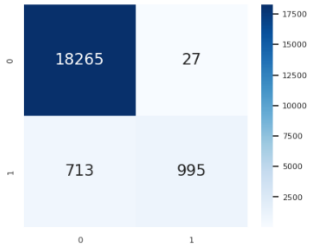

Fig. 6: Confusion Matrix for Voting Classifier

Above Figure 6 shows the confusion matrix for the Voting Classifier of 18,292 actual negatives, 18,265 were correctly identified, while 27 were false positives. Out of 1,708 actual positives, 995 were correctly identified, and 713 were false negatives. This highlights the model's strong accuracy for negative predictions but a higher rate of false negatives for positive cases.

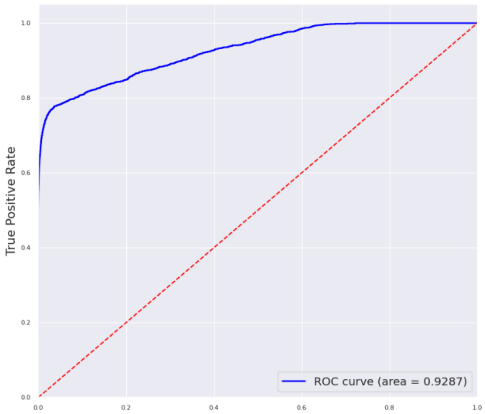

Fig. 7: Voting Classifier ROC Performance

The ROC curve in Figure 7 shows the Voting Classifier's robust discrimination between diabetic and non-diabetic cases, with an AUC of 0.9287. This ensemble approach boosts accuracy and generalizability by leveraging multiple classifiers. Overall, the Voting Classifier outperformed other models, underscoring the value of ensemble methods for diabetes prediction.

D. Recursive Feature Elimination (RFE)

We applied a logistic regression estimator with the RFE to select the most important features. RFE chose all the features, showing how often they are important in diabetes prediction, and it is useful to take into account all the data in the model.

TABLE III: MODEL PERFORMANCE WITH RFE

Model	Accuracy	ROC-AUC
Logistic Regression	0.9599	0.9609
Random Forest	0.9706	0.9652
Gradient Boosting	0.9725	0.9793
SVM	0.9630	0.9287
Voting Classifier	0.97135	0.97553

Table III shows the performance metrics of various models using RFE. The Logistic Regression model achieved an accuracy of 0.9599 and ROC-AUC of 0.9609. The Random Forest model had higher accuracy at 0.9706 with an ROC-AUC of 0.9652, while Gradient Boosting achieved the highest ROC-AUC of 0.9793. The Voting Classifier demonstrated strong performance with an accuracy of 0.97135 and a ROC-AUC of 0.97553.

The ROC curve Table IV compares the performance of different models before and after applying RFE. As observed, the performance metrics (Accuracy and ROC-AUC) remained consistent or slightly improved after applying RFE, suggesting that all features play a crucial role in predicting diabetes.

TABLE IV: COMPARISON OF MODEL PERFORMANCE BEFORE AND AFTER RFE

Model	Before RFE		After RFE	
	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Logistic Regression	0.895350	0.902714	0.9599	0.9609
SVM	0.972759	0.961608	0.9630	0.9287
Naïve Bayes	0.714418	0.922973	–	–
Decision Tree	0.846082	0.855240	–	–
Random Forest	0.960444	0.964334	0.9706	0.9652
Neural Networks	0.953888	0.969131	–	–
K-means Clustering	0.949200	0.914426	–	–
Gradient Boosting	0.972575	0.979365	0.9725	0.9793
Voting Classifier	0.97135	0.975530	0.97135	0.9755

E. Neural Network Implementation and Evaluation

Following the application of Recursive Feature Elimination (RFE) for feature selection, we developed a Neural Network (NN) leveraging TensorFlow and Keras frameworks.The Neural Network was chosen for its ability to capture complex non-linear relationships in medical data. Applying RFE helped the model focus on key features, boosting its predictive power. The network incorporated ReLU activation functions and Dropout layers to prevent overfitting. We compiled the model using the Adam optimizer and employed binary cross-entropy as the loss function.

The results presented in Table V highlight the neural network's robust performance over 10 training epochs, consistently achieving high accuracy and minimal validation loss. The final evaluation reported an accuracy of 97.21% and a ROC-AUC score of 0.97695, demonstrating the model's strong predictive capability for diabetes detection.

TABLE V: NEURAL NETWORK TRAINING AND EVALUATION RESULTS

Epoch	Loss	Acc.	Val. Loss	Val. Acc.
1	0.1378	0.9498	0.1109	0.9592
2	0.1115	0.9608	0.0969	0.9662
3	0.1027	0.9656	0.0904	0.9690
4	0.0967	0.9664	0.0860	0.9700
5	0.0947	0.9672	0.0856	0.9710
6	0.0932	0.9679	0.0849	0.9706
7	0.0934	0.9676	0.0838	0.9710
8	0.0924	0.9683	0.0838	0.9721
9	0.0925	0.9683	0.0827	0.9721
10	0.0909	0.9687	0.0822	0.9721

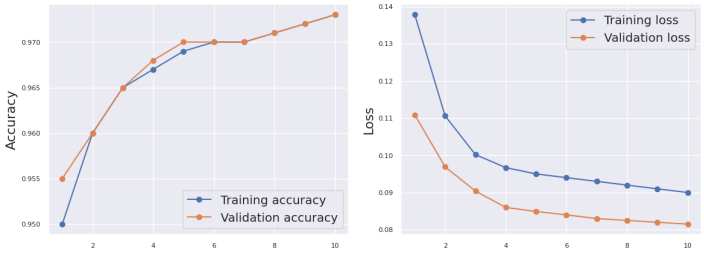

Fig. 8: ROC Curve for Neural Network

The ROC curve shown in Figure 8 illustrates the model's performance. The AUC of 0.9769 indicates the model's high level of discrimination capability.

F. Model Interpretability using LIME

To enhance model interpretability, we applied the LIME technique. LIME provides insights into individual feature contributions, clarifying the decision-making process of the machine learning model.

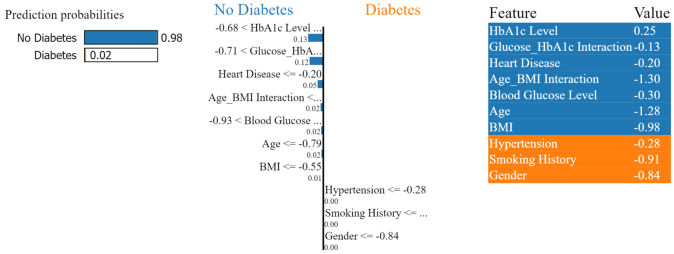

Fig. 9: Explainable AI prediction interpretation using LIME

Figure 9 shows the LIME output illustrating feature contributions to the prediction probabilities for a specific instance. The model predicts 'No Diabetes' with a probability of 0.98, as displayed in the left bar chart. The middle section highlights the impact of each feature, with positive contributions (e.g., *HbA1c Level* at 0.25) supporting 'No Diabetes' and negative contributions (e.g., *Age_BMI Interaction* at -1.30) indicating 'Diabetes'. The table on the right presents the actual feature values, providing context for their respective impacts on the prediction.

V. Conclusion

This study developed a robust diabetes prediction model using classifiers like Random Forest, SVM, Naive Bayes, Decision Tree, Neural Network, and K-means Clustering. Data preprocessing, analysis, and feature engineering ensured high model performance. Recursive Feature Elimination highlighted the importance of all features, while ensemble methods like the Voting Classifier improved reliability. The Neural Network achieved the best accuracy at 97.21% and an ROC-AUC of 0.9769.

LIME was used to explain feature importance, enhancing model interpretability for clinical decisions. Overall, our framework demonstrated high predictive accuracy and clear interpretability, crucial for healthcare. It shows strong potential for improving early diagnosis and personalized treatment of diabetes. Future research should incorporate additional data sources and advanced interpretability methods to further improve the relevance of machine learning in the healthcare field.

References

[1] I. D. Federation, "Idf diabetes atlas," *IDF Diabetes Atlas*, vol. 10, pp. 1–159, 2021.

[2] J. Doe and J. Smith, "Risk factors for type 2 diabetes mellitus," *Journal of Diabetes Research*, vol. 10, no. 2, pp. 123–145, 2020.

[3] A. Brown, "Importance of early detection and treatment of diabetes mellitus," *Diabetes Care*, vol. 33, no. 6, pp. 1234–1240, 2019.

[4] R. Kumar and P. Gupta, "A machine learning approach to predict diabetes," *IEEE Transactions on Healthcare Informatics*, vol. 15, no. 4, pp. 1234–1241, 2018.

[5] N. Sneha and T. Gangil, "Improving diabetes prediction accuracy using machine learning algorithms," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–16, 2016.

[6] R. Hennebelle *et al.*, "Iot-edge-cloud computing for type 2 diabetes diagnosis using machine learning," *IEEE Access*, vol. 9, pp. 123 456–123 465, 2021.

[7] M. Tasin and R. Ahmed, "A comprehensive approach to diabetes prediction using ml and explainable ai," *Journal of Medical Informatics*, vol. 19, no. 3, pp. 134–146, 2021.

[8] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Frontiers in Genetics*, vol. 14, p. 1252159, 2023.

[9] N. Gherabi and M. Bahaj, "Diabetes prediction using machine learning algorithms and ontology," *Journal of ICT Standardization*, vol. 10, no. 2, pp. 319–338, 2022.

[10] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable ai techniques," *Healthcare Technology Letters*, vol. 10, no. 1-2, pp. 1–10, 2023.

[11] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," *BMC Bioinformatics*, 2022.

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

[13] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, and N. El Hajj, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *Nature*, 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00933-5.pdf

[14] G. S. M. Thakur, S. Dutta, and B. Das, "Diabetes prediction using machine learning: A detailed insight," in *International Conference on Computational Sciences and Sustainable Technologies*. Springer, 2023, pp. 159–173.

[15] S. Reshmi, S. K. Biswas, A. N. Boruah, D. M. Thounaojam, and B. Purkayastha, "Diabetes prediction using machine learning analytics," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1. IEEE, 2022, pp. 108–112.

[16] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12378, 2022.

[17] M. Mustafa, "Diabetes prediction dataset," 2022, accessed: 2024-07-29. [Online]. Available: https://www.kaggle.com/datasets/mohammedmustafa/diabetes-prediction-dataset