Analytical comparison of machine and deep learning models in diabetes diagnosis: a comparative study of performance and accuracy

Rahma Ebrahim, Jehad Mahmoud, Habiba Arafa, Mariam Hassan
Department of Computer Science - Artificial Intelligence, Zewail City of Science and Technology, Egypt
E-mail: {s-jehad.elhendawy, s-habiba.arafa, s-mariam.hafez}@zewailcity.edu.eg

***Abstract*—Diabetes is one of the biggest health challenges worldwide, and early diagnosis is critical to limiting the development of complications. In this study, we used a diabetes prediction framework using machine learning (ML) and deep learning (DL) models, combining them with explainable AI (XAI) tools to better understand the decisions made by the models.**

During the data preparation phase, we noticed a class imbalance issue. We used techniques such as oversampling and undersampling to improve the data distribution. We also implemented feature selection using 10 different algorithms, followed by voting to select the top 7 features.

After testing several models, the best performance was achieved by the BPNN model, which achieved an accuracy of 89%. This demonstrates that machine learning-based models can provide effective solutions for diabetes prediction, especially when combined with results interpretation and data quality improvement techniques.

I. Introduction

Diabetes is one of the most prevalent chronic diseases worldwide, affecting the lives of millions of people [13]. According to the World Health Organization's 2020 reports, type 2 diabetes was among the leading causes of death [14].

This disease is complex and can lead to serious health complications, such as heart disease, nerve damage, retinal problems, and kidney dysfunction , [15],Therefore, early and accurate diagnosis is crucial to improving patients' quality of life and reducing potential complications .

In recent years, researchers have begun to rely on artificial intelligence techniques, particularly machine learning, to predict the likelihood of developing diabetes from patient data [12], [2], [5]. These models have shown promising results in terms of accuracy [7], [1], but one of the major challenges is the difficulty of interpreting the results of the models, especially if they are complex, which can limit doctors' ability to trust their recommendations.

Hence, a new field known as explainable artificial intelligence (XAI) has emerged, which aims to explain how models make their decisions in a way that is understandable to users, especially physicians who do not have advanced technical expertise in the field of artificial intelligence .

Despite advances in this field, the number of studies that combine machine learning techniques and XAI concepts on real-world medical data, such as diabetes, remains limited. Most previous studies were either small in sample size [2] or relied on only one type of model or explanation tool [9], which reduces the strength and generality of their findings.

Therefore, this study aims to fill this gap by conducting a practical analysis using a variety of machine learning and deep learning models, without using ensemble models, to evaluate their performance in predicting diabetes. The Pima Indian Diabetes Dataset, a real-world dataset available through platforms such as Kaggle, was used.

The primary objective of this study is to evaluate the efficiency and interpretability of different models, thus contributing to the provision of effective tools to support physicians in their decision-making, even if they are not experts in the field of artificial intelligence.

II. Literature Review

Diabetes mellitus poses a major public health challenge. In response, researchers have increasingly relied on machine learning (ML) and deep learning (DL) techniques to improve the accuracy and efficiency of diabetes prediction. While these models have shown impressive performance, their "black-box" nature limits clinical trust. Explainable AI (XAI) techniques were introduced to address this gap by offering insights into how predictions are made [15].

Machine Learning Models for Diabetes Prediction

Various ML models have been applied to diabetes diagnosis, including Decision Trees, Random Forest, Support Vector Machines (SVM), and Logistic Regression. Their success depends on data preprocessing, feature selection, and model tuning. Ensemble methods, such as boosting and bagging, have

demonstrated superior accuracy compared to individual models [8],[9].Kowsher et al. combined SVM with ANN, achieving 94.87% accuracy, while Chou et al. reported an AUC of 0.991 using a boosted tree model. Dharmarathne et al. found XGBoost to be the best-performing model and used SHAP for interpretation [9]. Uysal identified SVM and Random Forest as the top models when tested with XAI [10].A detailed summary of the studies included in this review is presented in **Table 1**."

Use of Explainable AI (XAI)

LIME and SHAP are the most frequently used XAI tools.

- **LIME** provides local explanations for individual predictions, helping identify key features like glucose levels [1][5].
- **SHAP**, based on game theory, offers both global and local interpretations by assigning feature contributions across all combinations. It has been widely used to interpret models in recent studies, including those by Islam et al. [11]

However, many studies used XAI in a limited scope without systematic comparison, reducing their interpretive power and clinical applicability [3].

Ensemble Learning and Interpretability

Ensemble models are gaining attention for their robustness and accuracy. kibria et a.demonstrated that soft voting classifiers outperformed single models and highlighted the importance of dataset balance [5]. Tasin et al. showed that decision trees and XGBoost not only performed well but also supported interpretability [8].Zhang et al. combined Boruta feature selection with ensemble models and achieved 89.4% accuracy on different datasets [7]

Deep Learning Models

Deep learning models, such as DNNs, have outperformed traditional ML models in capturing complex patterns[1]. Zhang et al. used a BPNN with batch normalization, reporting up to . Madan et al. proposed a CNN-BiLSTM hybrid model for real-time prediction, outperforming classic ML [7]

Despite growing interest in XAI, most studies either apply a single explanation technique or treat models as black boxes, without evaluating interpretability in depth. This study aims to fill that gap by applying and comparing multiple XAI techniques (e.g., SHAP, LIME, PDF, Global surrogate model) across different models. The goal is to enhance both interpretability and predictive performance for real-world clinical use.

III. Materials and Methods

The methodology applied (summarized in Fig. 1) uses the Pima dataset, which is available on Kaggle. These steps contain EDA , feature selection, model training, model evaluation, and explainability methods for each model. The goal is to ensure results of models by using XAI to understand the decision-making process. This framework provides a clear direction for evaluating machine learning and explainable AI methods to improve their effectiveness in predicting diabetes..

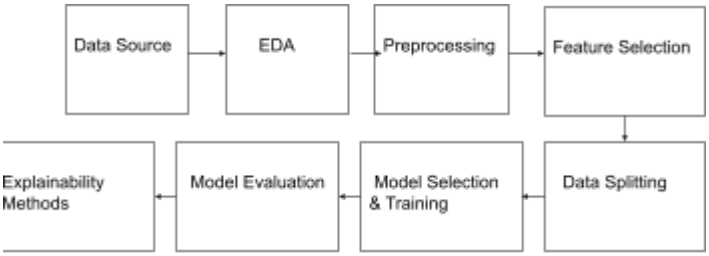


Fig. 1 Methodology Diagram for Explainable Diabetes Prediction Using Machine Learning and Deep Learning

TABLE I

SUMMARY OF THE RELATED WORK.

Reference	Objective	Dataset	Prediction Techniques	XAI Technique
Tanim et al. (2024) [1]	Develop explainable deep learning model for diabetes diagnosis	Three benchmark datasets (specifics not stated)	DeepNetX2 (custom DNN architecture)	in paper:LIME, SHAP our work:LIME, SHAP,PDP,Global surrogate model
Ali et al. (2020) (ResearchGate) [2]	Diabetes classification using KNN	PIMA Indian Diabetes Dataset	K-Nearest Neighbors (KNN)	in paper: N/A our work:Nearest Neighbors.PFI,LIME, SHAP

Kowsher et al. (2023) [3]	Predict Type 2 diabetes treatment response	Unspecified dataset (likely clinical records)	Artificial Neural Network (ANN)	in paper :N/A our work:LIME,SHAP, PFI,Global surrogate model
Brisimi et al. (2018) (ScienceDirect) [4]	Analyze chronic stress and impaired glucose metabolism using ML	Electronic health records (EHRs)	convLSTM	in paper :N/A our work:Saliency Map,Feature Ablation,LIME,Global surrogate model
Kibria et al. (2022) [PMC9571784] [5]	Develop an interpretable diabetes prediction method using ensemble learning	PIMA Indian Diabetic Dataset (PIDD)	Soft Voting Classifier (ensemble of multiple models)	In paper: SHAP, LIME our work: SHAP, LIME,PDP.ICE ,LOFO
Ganguly et al. (2023) [IJACSA] [6]	Implement XAI for diabetes prediction using ensemble methods	PIMA Indian Diabetic Dataset (PIDD)	Ensemble models (XGBOOST)	in paper:LIME, SHAP our work:PDP,ICE, LOFO,PFI,H_statisirics
Zhang et al. (2024) (Springer) [7]	Propose a non-invasive diabetes diagnosis method using deep learning	Pima dataset, CDC BRFSS2015, Mesra Diabetes dataset	Back Propagation Neural Network (BPNN) with batch normalization	in paper :N/A our work:LIME, SHAP, LOFO,PDP
Tasin et al. (2023) [8]	Develop diabetes prediction system using ML and XAI on merged datasets	PIMA Indian + private RTML dataset (Bangladesh)	XGBoost, SVM, RF, KNN, AdaBoost, Voting Classifier	in paper:LIME, SHAP our work :LIME, SHAP, PDP.,LOFO
Dharmarathne et al. (2024) [9]	Develop self-explainable diabetes diagnostic interface	Clinical dataset (unspecified)	Multiple ML models(Select SVM)	in paper:Built-in explainability our work :SHAP, LIME PFI,,Global surrogate model
Uysal (2023)[10]	Create clinically interpretable diabetes prediction model	PIMA Indian Dataset	Ensemble methods(select Random Forest)	in paper: SHAP, LIME our work :SHAP, LIME,PFI,,Global surrogate model
Islam et al. (2024) [11]	Optimize diabetes prediction with comprehensive XAI	PIMA Indian Dataset	Hyperparameter-tuned ML models (Exreme Tree Classifier)	in paper: SHAP, LIME, Partial Dependence Plots our work : SHAP, LIME,PFI,ICE
anda & Mahanta (2023) [arXiv:2311.05665] [12]	Enhance interpretability of diabetes prediction using Random Forest	Public diabetes symptoms dataset	Random Forest Classifier	in paper:LIME, SHAP our work :LIME, ,PDP,ICE, LOFO

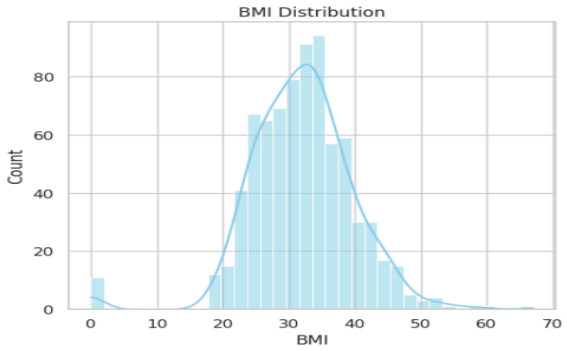
A. Dataset Description

The dataset used in this study is the well-known Pima Indians Diabetes dataset, which consists 768 records of female patients aged 21 and older. Each record includes eight input features and one binary output variable, "Outcome," which indicates whether the patient has diabetes (1) or not (0). The features include the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function (a measure of genetic influence), and age. The dataset was provided in CSV format and went through preprocessing to ensure it was ready for effective machine learning analysis and prediction tasks.

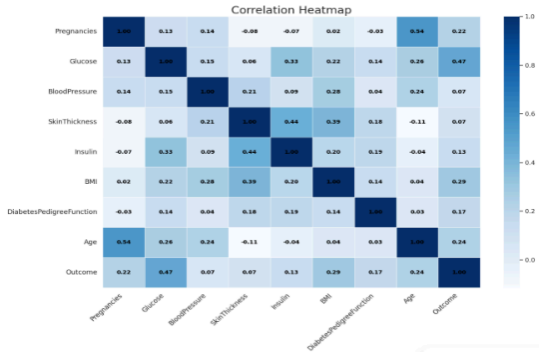
1. Exploratory Data Analysis (EDA)

Exploratory data analysis was did to identify important patterns and relationships in the dataset. A histogram of BMI values showed a right-skewed distribution, suggesting that most patients are classified

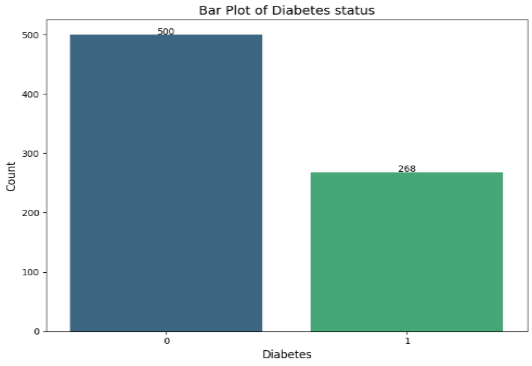
as overweight or obese. This highlights the significance of BMI as a key factor in predicting diabetes.



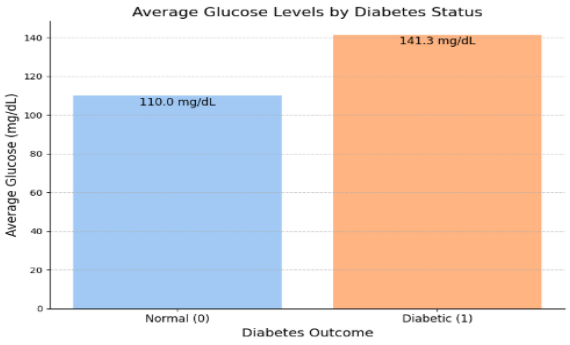
A correlation heatmap was created to examine the statistical relationships between the features and the target outcome. The heatmap revealed that glucose showed the strongest positive correlation with the diabetes outcome (around 0.47), followed by BMI and age. These shows played a key role in the feature selection process and reinforced the inclusion of these variables in the model.



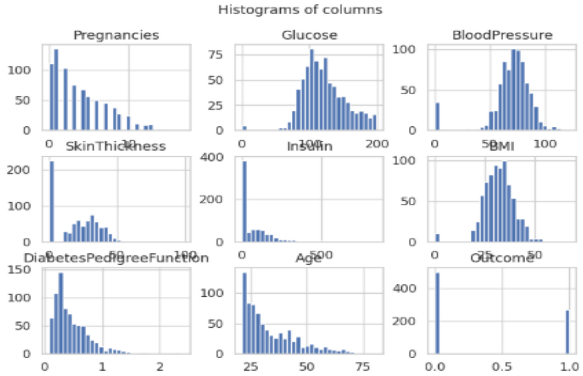
A bar plot depicting the class distribution of diabetic and non-diabetic patients revealed a slight imbalance, with non-diabetic cases being prevalent. To control this balancing, this techniques are applied during preprocessing, ensuring the model would not biased toward the true class.



The chart was used to judge the role of glucose as a predictive feature by comparing the average glucose levels between diabetic and non-diabetic groups. The findings revealed a difference, with diabetic individuals having notably higher average glucose levels (~141 mg/dL) compared to non-diabetics (~110 mg/dL), highlighting the clinical significance of glucose in diagnosing diabetes.

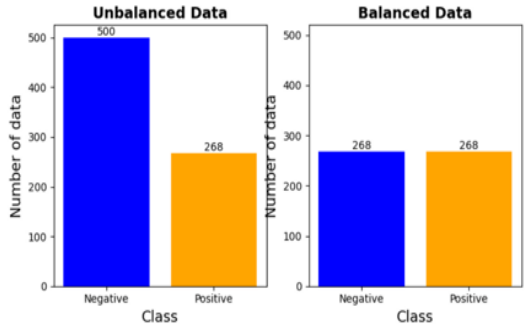


To gain an initial insight into distribution of each feature, a grid of histograms created for all columns in the dataset. This visualization revealed the skewed distributions of features like BMI, insulin, and diabetes pedigree function, while others, such as glucose and blood pressure, exhibited symmetric distributions. This analysis detected outliers and guided the decisions for subsequent preprocessing and feature transformation.

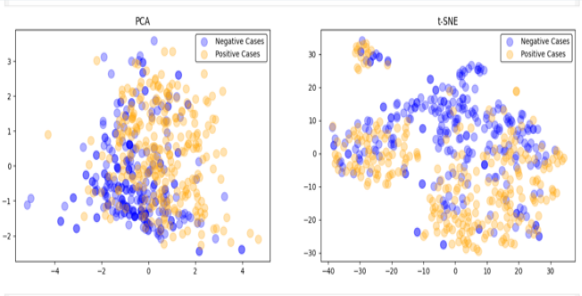


2. *Data Preprocessing*

An examination of class balance revealed a slight imbalance, with more non-diabetic cases than diabetic ones. To address this, undersampling was used to reduce the majority class while maintaining the integrity of the original data distribution. Oversampling was deliberately avoided to prevent the introduction of synthetic bias into the dataset.

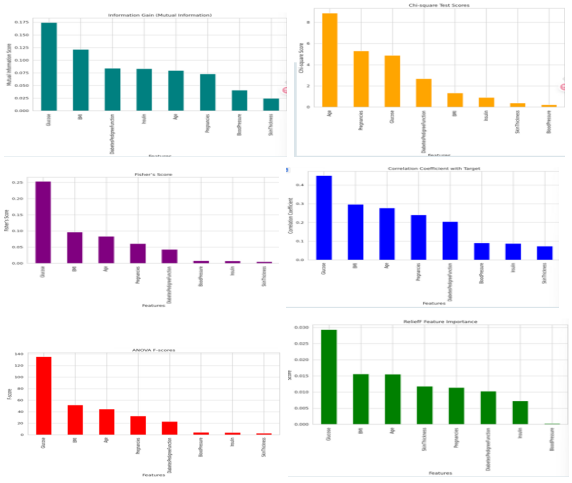


Moreover, dimensionality reduction methods PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) were apply to visualize between diabetic and non-diabetic instances within the feature space.



3. *Feature Selection*

To enhance model performance and interpretability, a thorough feature selection approach was implemented. Several filter methods were used, such as Information Gain, Chi-square tests, Fisher Score, Correlation Coefficient, ANOVA (F-test), and Relief, with each method ranking the features according to their relevance to the target variable.



In addition, wrapper methods like Forward Selection and Backward Elimination were utilized to iteratively test feature subsets. Finally, the Recursive Feature Elimination (RFE) method was used as an embedded approach to identify the optimal set of features.

To consolidate the results from all methods, a voting mechanism was adopted to determine the most consistently selected features across all techniques. The final list of top seven features included: Glucose, BMI, DiabetesPedigreeFunction, Age, Pregnancies, Insulin, and BloodPressure.

B. Data Splitting for Model Evaluation

To ensure a reliable model performance, the dataset was divided into two separate subsets a training set and a testing set, 80% of the data was assigned to the training like model training and feature selection, and 20% was set aside as a testing to evaluate the model's ability to generalize to unseen data. This stratified division maintained the original class distribution, ensuring that both diabetic and non-diabetic cases were evenly represented in both subsets, and now we reduced the risk of overfitting and obtained an accurate estimate of the model's real-world performance.

C. Model Selection, Training and Hyperparameter Tuning

DL and ML models are used for diabetes prediction. Most models have the same preprocessing steps, but some models have different specific preprocessing steps and hyperparameter tuning to enhance model performance.

The models used are traditional machine learning classifiers such as SVM, RF, Extreme Tree Classifier, KNN, XGBoost, and a Voting Classifier to predict diabetes with good accuracy. Decision trees have an interpretable nature and hierarchical structure, which is important for critical decision paths in the data. K-Nearest Neighbor leverages the proximity of data points , providing a robust approach to pattern recognition. Support Vector Classifier excels at separating complex, high-dimensional datasets by finding optimal hyperplanes. Extreme Gradient Boosting has the power of ensemble learning to improve predictive accuracy by sequentially optimizing weak learners.In addition to advanced neural architectures including Artificial Neural Networks (ANN), ConvLSTM networks, Backpropagation Neural Networks (BPNN), and DeepNetX2. ANN is selected because of its ability to model complex nonlinear relationships common in biomedical data. ConvLSTM improved this by capturing temporal and spatial patterns in structured sequences. BPNN provides a flexible multilayer design to refine health-related features.These models were selected based on their effectiveness to analyze health records.

Furthermore, we applied hyperparameter tuning using GridSearchCV or manual selection, depending on model complexity. For SVM, a linear kernel was applied and validated using 5 fold cross validation. Random Forest and Extreme Tree Classifier models were tuned for parameters

such as max depth, n estimators, and samples split. KNN models were tested with different k values.
This approach allowed a stable tuning technique by avoiding overfitting of the models and guaranteeing that the frequency within all data folds was well-tuned.

In the XGBoost model , it trained with maximum depth = 4 and to mitigate class imbalance, ADASYN (Adaptive Synthetic Sampling) was applied, generating synthetic instances for the minority class before training led to enhance the classifier's ability to detect diabetes and improved generalization.

The voting classifier combined the predictions of Random Forest and XGBoost through soft voting, which averages the predicted probabilities instead of depending on the majority class, using their strengths to generate robust, balanced predictionsIn NN models, the ANN architecture contained sequential dense layers with ReLU activations, batch normalization, and dropout for regularization and was trained using the Adam optimizer, binary cross entropy loss, and mini-batch training over 50 epochs. ConvLSTM was utilized with convolutional LSTM layers followed by dense classifiers. BPNN was trained using three hidden layers and an output layer, with sigmoid activation and batch normalization.

D. Model Evaluation

The performance of both ML models and the DL models was accurately evaluated using a comprehensive set of metrics: accuracy, ROC-AUC, precision, F1-score, sensitivity (recall): $TP/TP+FN$ (true positives / (true positives + false negatives)), and specificity: $TN/TN+FP$(True negatives / (true negatives + false positives)) with learning and validation curves.

E. Explainability of Models

We utilized several well-established explainability techniques to gain a more thorough understanding of our predictive model's behavior. These methods provide clarity on how the model arrives at its decisions and enable us to assess the influence of key features, such as glucose, BMI, and age, on the model's predictions. The techniques we examined include SHAP, LIME, Partial Dependence Plots (PDP), Tree Surrogate Models, Permutation Feature Importance (PFI), and Leave-One-Feature-Out (LOFO). Each of these methods provides a distinct perspective on the model's predictions, and together, they enable us to verify that the model is both reliable and understandable.

SHAP (Shapley Additive Explanations) is a powerful technique grounded in cooperative game theory. It assigns a value to each feature based on its contribution to a specific prediction, allowing us to quantify the impact of each feature in isolation. By calculating the Shapley value, SHAP provides a precise measure of how much each feature (e.g., glucose, BMI, or age) influences the model's prediction relative to the average outcome. In our analysis, SHAP was invaluable for identifying which features played the most significant role in the model's decisions. For instance, SHAP revealed that glucose levels had the most substantial impact on the model's predictions, which was consistent across various instances. This helped highlight the importance of glucose as a predictor in the health-related outcomes we were studying, offering a clear understanding of how individual predictions were made.

Partial Dependence Plots (PDP) were used to visualize the relationship between individual features and the predicted outcome while keeping all other features constant.Partial Dependence Plots (PDP) were used to visualize how individual features relate to the predicted outcome while holding other features constant. This technique was particularly useful in understanding how each feature affects the prediction across a range of values. For example, the PDP for glucose demonstrated that as glucose levels increased, the likelihood of a positive prediction (such as a diagnosis of diabetes) also rose, which was a crucial insight for interpreting the model's behavior. PDPs helped us understand not only the direction of the relationship between a feature and the outcome but also the strength of this relationship across the entire dataset.

Tree Surrogate Models were employed to approximate the predictions of a complex model using a simpler, interpretable decision tree. The tree surrogate acts as an approximation of the black-box model, allowing us to understand how the complex model combines different features to make predictions. By fitting a decision tree to mimic the behavior of the original model, we could observe how features interacted and influenced the final outcome. This approach provided a clear, intuitive understanding of the model's decision rules, showing, for example, how a specific combination of glucose and BMI levels led to a higher probability of a certain prediction. The tree surrogate technique helped in simplifying the model's decision-making process and presented an easily understandable logic behind the predictions.

For instance-level interpretations, LIME was used in the study because it provides clear heatmaps of the features' contributions to the achieved prediction, understanding what features helped to make the decision to be diabetic and other contributing percentages for non-diabetic. The heatmap was a noteworthy feature because it allowed for comparison and explanation of predictions irrespective of the models used in the application.

LOFO is used to test model performance without a specific feature every time, so it shows the feature's importance and its effect on model performance. Glucose is one of the most important features as LOFO shows the drop in model performance when it was eliminated.

PFI is used to measure feature importance by measuring how the performance of a trained model will change by shuffling its values randomly, and with the conclusion taken from all models, PFI shows BMI, glucose, and age are the most important features .

F. Explainability Metrics

Several explanatory metrics were applied to assess the quality and reliability of the model explanations.

- ***Fidelity*** *checks if an explanation is consistent with the model's original predictions, accurately representing the model's behavior.*
- ***Faithfulness*** *checks if explanations reflect the mechanisms of the model accurately , providing real insights into the model's decision-making process. .*
- ***Stability*** supports similar inputs to have similar explanations so generating more reliable model outputs.
- ***Consistency*** *validates that different models agree on feature importance as its main value for validating results in varied situations*

These measures provide a strong foundation to assess the quality of model explanations and aligning their interpretability and accuracy in the context of healthcare. .

IV. Analysis of the Results

In this section, we present and analyze the results of our study for best model performance, explainability methods.

The highest predictive accuracies were achieved by individual models, such as XGBoost, BPNN, and SVM, because these can model nonlinear relationships and complex interactions in healthcare datasets

These models are robust to variations in data structures and are, therefore, suitable for predicting diabetes mellitus. The best result among the three models was XGBoost. Also, it is not the highest accuracy, but its explanation methods results show a successful decision-making process without any problem for overfitting, achieving a test accuracy of 75%. This balance between accuracy and practical utility makes such a combination highly relevant for clinical decision-making.

.

We used various explainability techniques to detect global patterns and explain feature interactions, generate personalized insights to help healthcare professionals understand and trust machine learning predictions, and build a robust framework.

1) Global Interpretability - Key Predictors Driving Diabetes Risk: We used global interpretability tools to identify the most important predictors and their overall contribution to diabetes risk. SHAP was used to explain the importance of global and local features. Important three drivers contributing to diabetes risk included Glucose, BMI and Age (see Fig. 2). The results also indicated that Glucose level has a positive correlation, and general health and physical activity have a moderating impact, as they act as buffer variables. SHAP enhances the depiction of such associations to facilitate the identification of clinical intervention priorities.

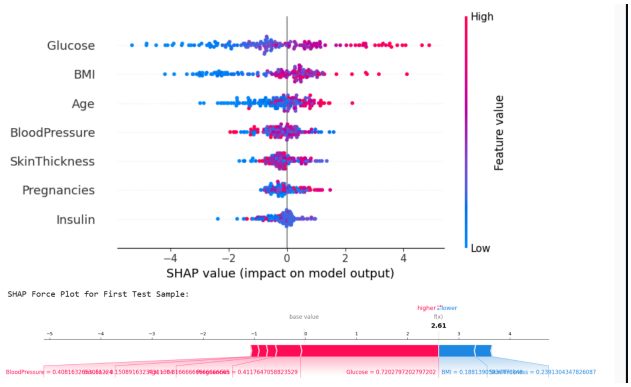


Fig. 2. Key features influence diabetes prediction - SHAP Summary Plot

The results from the XGBoost indicated that it reached accurate predictions while being comprehensible by clinicians and patients. As shown in Fig. 3, XGBoost demonstrated the particularity of ranks, where BMI and Glucose were placed as the highest priority of the most important predictors. XGBoost combines accuracy and interpretability by construction increases its effectiveness in solving specific healthcare problems.

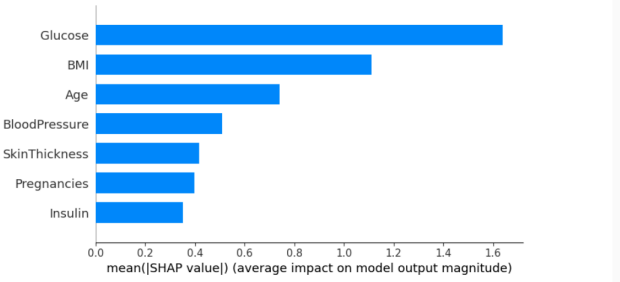


Fig. 3. SHAP for XGBoost – Top Global Predictors

The output for XGBoost model using LOFO in figure 4 represent Glucose has the largest bar, with a decrease in ROC AUC of around 0.04. This indicates that Glucose is the most critical feature for the model's predictive performance. Removing it causes the largest decreasing in the model's performance to predict classes.BMI: The second most important

feature, with a decrease in ROC AUC of about 0.015 plays a significant role in the model's performance.Age: The third most important feature, with a decrease in ROC AUC slightly above 0.01. Age contributes slightly to the model. SkinThickness and BloodPressure have a moderate impact on the mode ROC AUC , while Insulin and Pregnancies are the least influential. This suggests they impact less to the model's performance compared to Glucose, BMI, and Age.

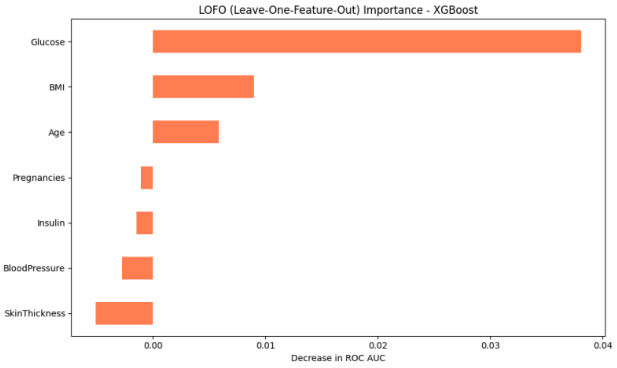


Fig. 4.Leave One Feature Out for XGBoost model

In fig 5 PDP for XGBoost with ADASYN model When the value of feature 0 ranged between approximately 0 and 0.17, there was a decrease in the value of the pdp from 0.38 to 0.36 and from 1.7 to 1.9, it rose again to 0.38, and when the value of the feature was approaching 0.2 to 0.3, there was a clear decrease in the value of the pdp, then from approximately 0.37 to 0.45.The largest increase occurred and the pdp value reached approximately 0.43 and then decreased again, which shows that this The feature has a varying effect on the prediction of this class

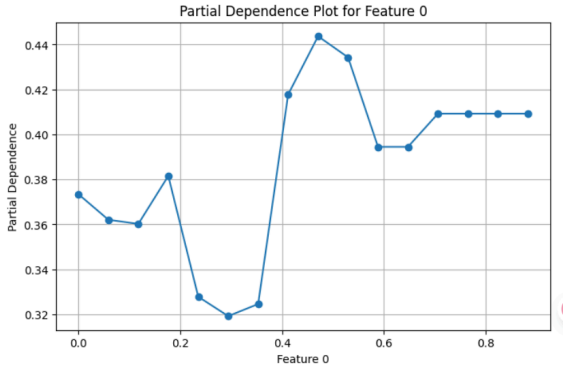


Fig. 5.PDP Plot Out for XGBoos with ADASYN model

2) Local Interpretability - Personalized Insights: Local interpretation helped resolve ambigu- ity at the micro level and gave a precise idea about some local predictions to help clinicians devise fine-tuned patient treatment plans. LIME provides post hoc, understandable, and concrete examples related to each prediction. As shown in Fig. 5, two features that impacted the prediction of the particular patient were Glucose and BMI.

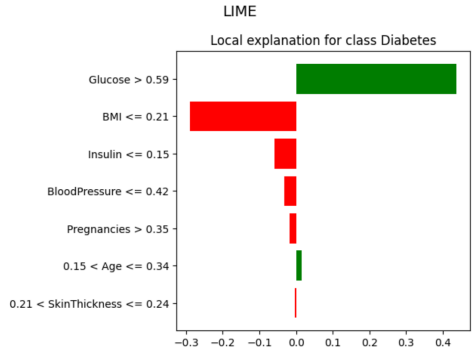


Fig. 5. LIME Explanation for Instance Level Predictions

The SHAP force plots (Fig.6) showed decomposed individual predictions into an additive feature contribution. For instance, BloodPressure was found to have a strong positive relationship with diabetes in a specific patient. Still, features such as BMI were found to decrease diabetes. These visualizations enable clinicians to easily distinguish the nature of various factors at the patient level.

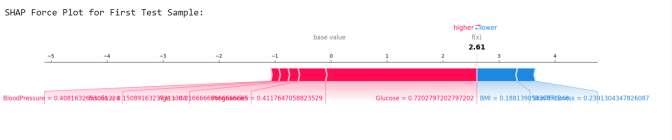


Fig. 6. Positive and Negative Feature Effects for a Single Prediction using SHAP Force Plot

V. Discussion

Key Takeaways and insights:

The Random Forest model achieved 89% accuracy, which is the best accuracy in all models. XGBoost, Voting Classifier (Random Forest, XGBoost), they have also achieved more than 85%, which indicates the good performance of ensemble algorithms. While these models achieved more than 95% accuracy in previous studies [5][6][12].

Feature Selection techniques, e.g.Information gain, Chi-square, Recursive feature elimination(RFE), were effective in improving the models by giving importance to the correct features, and this appears in reducing overfitting, decreasing the model training time, and boosting generalizability

The use of XAI tools (SHAP, LIME, PFI, and Global Surrogate Trees) provided excellent transparency, helping us understand how the model made its predictions. SHAP provides the shap values, which are a good indicator of how each feature impacts the prediction of the model and which of them have a wide range, so their effect is greater than others.LIME provided insights for a specific instance(patient), which led to this classification for this instance.PDP (Partial Dependence Plot) helps visualise the average effect of a feature on the model's prediction. ICE (Individual Conditional Expectation) is similar to PDP, but it plots a separate line for each instance, providing insight into how the model behaves at an individual level rather than showing just the average. LOFO (Leave-One-Feature-Out) measures feature importance by retraining the model multiple times, each time leaving out one feature and observing the performance drop.PFI (Permutation Feature Importance) is similar in purpose to LOFO but avoids retraining by randomly shuffling the values of a feature and measuring how much the model's performance decreases or increases. H-statistic quantifies the interaction between features, helping to understand whether features act independently or interactively in influencing the prediction. Global Surrogate Trees offered a simpler decision-making process from the root to the final decision through a sequence of decisions.

Combining global (SHAP, PFI, ICE, LOFO, H_statistics, Global surrogate tree) and local (LIME) XAI methods solved the interpretability issue by providing interpretability for the model in general and for specific instances of data.

Challenges and Limitations:

Class imbalance: the recorded instances of diabetes and non diabetes patients are not equally which caused a challenge and led to skewed learning

Data issues, like missing insulin values and BMI outliers, required normalisation and scaling during preprocessing. This may lead to inaccurate data, which will affect in the behaviour and performance of the model.

The computational complexity of XAI techniques, especially SHAP, made it impractical for real-time clinical applications, requiring heavy computational resources for large datasets and being time-consuming to use.

Effectiveness of Explainability Techniques

SHAP provided accurate global and local feature importance

LIME's explanations were particularly helpful for borderline cases, offering actionable advice like reducing BMI, which can help in ƈrotecting the patient from diabetes if he adheres to the instructions.

PFI validated feature rankings, which improve the model performance by filtering specific features and which makes its performance worse

PDP, ICE were weaker when features were interdependent, showing that feature independence assumptions don't always hold.

LOFO: Good indicator of feature importance by retraining the model and measuring performance after deleting each feature, not only the important features are shown, but also the features affect the model negatively will be shown with a negative value

Unexpected Observations and Explanations:

There are some models, although of higher accuracy, but their output of explanation techniques is different to other models or the output of the most important features in feature selection techniques this because of the different pre-processing and number of data used in some models or not learned well-learned.

The low accuracy of neural network models (around 70%) because of the more needs to instances in the dataset

VI. Conclusion

This research paper explains how different machine learning and deep learning algorithms can help in the early detection of diabetes before complications occur, based on information such as Glucose level, BMI, Diabetes Pedigree Function,etc.Through this information, the different models were able to build their predictions

By using explainability techniques, it is possible to build transparency in these models by explaining the reason behind the model prediction, whether it is local or global methods, which gives doctors a better insight into the prediction and leads to them making better decisions and enabling them to rely on these models.

Finally, although these results indicate that these models can contribute to the development of treatment methods and medical practices, it is not possible to be satisfied with the data used only; rather, these models must be tested on larger and

more diverse data for real patients and the interpretation of their results improved. Its development must continue in the future to keep pace with changes and different behaviours.

REFERENCES

[1]A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, & S. Abdulwahed. (2020). Diabetes classification based on KNN. IIUM EngineeringJournal,21(1),176–181.
https: //www.researchgate.net/publication/338687609

[2] Gangani Dharmarathne, Jayasinghe T.N., Madhusha Bogahawaththa, D.P.P. Meddage, & Upaka Rathnayake. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. HealthcareAnalytics,5,100301.https://doi.org/10.1016/j.health.2024.1001

[3] Ganguly, R., & Singh, D. (2023). Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach. International Journal of Advanced Computer Science and Applications, 14(7). https://doi.org/10.14569/ ijacsa.2023.0140717

[4] Ilhan Uysal. (2023). Interpretable Diabetes Prediction using XAI in Healthcare Application. Journal of Healthcare Informatics Research, 8(1), 20–38. https://www. researchgate.net/publication/376208551

[5] Islam, M. M., Rahman Rifat, H., Shamim, M., Akhter, A., Uddin, M. A., & Mohi, M. (2024). Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME. Engineering Reports, 6(3). https://doi.org/10.1002/eng2.13080

[6] Kibria, H. B., Nahiduzzaman, M., Goni, Md. O. F., Ahsan, M., & Haider, J. (2022). An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. Sensors, 22(19), 7268. https://doi.org/10.3390/ s22197268

[7] M. Kowsher, M. Y. Turaba, M. M. M. Rahman, & T. Sajed. (2023). Type 2 diabetes treatment prediction using artificial neural network. arXiv preprint arXiv:2301.03093v1. https://arxiv.org/abs/2301.03093

[8] Sharia Arfin Tanim, Aurnob, A. R., Shrestha, T. E., Islam, R., & Ullah, S. (2024). Explainable deep learning for diabetes diagnosis with DeepNetX2. Biomedical Signal Processing and Control, 99(8). https://doi.org/10.1016/j.bspc.2024.106902

[9] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters. https://doi. org/10.1049/htl2.12039

[10] T. S. Brisimi, T. Xu, & I. C. Paschalidis. (2018). Chronic stress and impaired glucose metabolism: A machine learning approach. IEEE Journal of Biomedical and Health Informatics, 22(5), 1632–1640. https://doi.org/10.1016/S1476927120304692

[11] Panda, M., & Ranjan, M. S. (2023). Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP. ArXiv.org. https: //arxiv.org/abs/2311.05665

[12] Zhang, Z., Ahmed, K. A., Hasan, M. R., Gedeon, T., & Hossain, M. Z. (2024). A Deep Learning Approach to Diabetes Diagnosis. Communications in Computer and Information Science, 87–99. https://doi.org/10.1007/978-981-97-5937-8_8