EXPLAINABLE AI FOR DIABETES PREDICTION

M. Alaa[1], M. Kamal[2], M. Adam[3], and A. Elrashidy[4]

[1,2,3,4]*Zewail City of Science and Technology, Egypt*

Research Supervisors: Dr. Mayada Mansour[*], Eng. Rana Mohamed[*]

[*]*Department of Computer Science, Zewail City*

Abstract

This study aims to improve diabetes classification by evaluating a diverse set of machine learning (ML) and deep learning (DL) models on the BRFSS 2015 dataset, which includes over 250,000 health records. The primary objective is to identify high-performing, interpretable AI models that can support early diagnosis and clinical decision-making.

Twelve models were implemented, including:

Logistic Regression	Decision Tree	Support Vector Machine (SVM)	K-Nearest Neighbors (KNN)
Gaussian Naive Bayes	Random Forest	AdaBoost	CatBoost
Gradient Boosting	Neural Network	Convolutional Neural Network (CNN)	XGBoost + Conditional GAN

We employed a range of Explainable AI (XAI) techniques to enhance the interpretability and transparency of our model's predictions. Specifically, we employed SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) for local interpretability by analyzing the effect of individual features on particular predictions. In addition, we incorporated global interpretability methods such as Permutation Importance, Leave-One-Feature-Out (LOFO), and Partial Dependence Plots (PDP) in order to calculate the overall feature contribution throughout the model. Furthermore, we used the Accumulated Local Effects (ALE) plots and the H-Statistic to detect feature interactions and non-linear effects. The surrogate model also globally learned to mimic the black-box model behavior and give a readable explanation of its decision rule. All these methods of explanation collectively provided better understanding of the model behavior and aided in more trustworthy deployment in real-world applications.

Evaluation metrics included accuracy, F1-score, and ROC-AUC. Results showed that XGBoost with Conditional GAN achieved the highest performance (Accuracy: 86.7%, F1: 0.85, ROC-AUC: 0.89), followed by CNNs and Neural Networks. SHAP and LIME confirmed the predictive relevance of features such as BMI, age, and blood pressure, aligning with medical understanding.

Overall, the study demonstrates that ensemble and hybrid DL models, when combined with XAI techniques, can deliver both high accuracy and interpretability. These findings highlight the potential for deploying trustworthy AI systems in healthcare settings, supporting more informed and explainable clinical decision-making.

1 INTRODUCTION

Diabetes mellitus is a long-term metabolic disorder that affects millions of individuals worldwide and places a huge burden on the world's health care systems. Early diagnosis is required to prevent severe complications such as cardiovascular disease, neuropathy, and renal failure. Despite improvements in screening and monitoring, early detection of high-risk individuals is still a problem, particularly in large and diverse populations. Here, artificial intelligence (AI) offers promising technology to automate and enhance the accuracy of disease categorization from patient data.

Machine learning (ML) and deep learning (DL) methods have become very popular in health care studies because they are able to recognize patterns and relationships among complex, high-dimensional data sets. Such models can be used to support clinical decision-making by making early predictions based on historical and behavioral health indicators. One of the most important limitations of applying AI systems in medicine is that they are non-interpretable. The majority of high-performance models, especially neural networks and ensemble methods, are "black boxes" that provide accurate predictions without apparent reasoning behind those predictions. This is a cause of suspicion and discourages clinical application, especially where explainability is legally or ethically necessary.

The aim of this project is to compare and contrast the explainability and performance of a vast range of DL and ML models for diabetic diagnosis. The models are deployed over the Behavioral Risk Factor Surveillance System (BRFSS) 2015 data, a wide-ranging public health survey that covers over 250,000 individuals with attributes such as age, BMI, smoker status, level of physical exercise, and state of general well-being. This varied dataset allows the exploration of many modeling approaches and analysis of the quality of each model in extrapolating to real data.

Twelve models are used in this study, ranging from the standard, ensemble, and deep learning techniques: Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naive Bayes, Random Forest, AdaBoost, CatBoost, Gradient Boosting, Neural Network, Convolutional Neural Network (CNN), and XGBoost with a Conditional GAN for creating synthetic data. To enhance transparency, the study applies a variety of Explainable AI (XAI) techniques: SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and Grad-CAM (Gradient-weighted Class Activation Mapping).

1.1 Research Objectives and Guiding Questions

The current research aims to offer a benchmark analysis that not only evaluates the accuracy of different models but also analyzes their interpretability within medical decision-making. The research responds to the following main questions:

What are the best machine learning and deep learning models for predicting diabetes in the BRFSS 2015 dataset?

How do explainability tools (SHAP, LIME, Grad-CAM) assist in model behavior and prediction interpretation?

2 Related Work

In the recent past, most research articles have explored using machine learning (ML) and deep learning (DL) models in predicting and categorizing diabetes. Simple models such as Logistic Regression and Decision Trees are being largely used due to their simplicity to comprehend and understand. For instance, (1) demonstrated logistic regression's potential in predicting type 2 diabetes based on lifestyle and biometric parameters. Similarly, Decision Tree models are employed in (2), representing moderate accuracy with the added advantage of distinguishable decision paths enhancing interpretability.

Ensemble methods such as Random Forest, AdaBoost, and Gradient Boosting have proven to be more efficient than a single classifier through the aggregation of numerous weak learners to reduce variance and bias. In (3), Random Forests were found to achieve high accuracy when used with structured clinical data for detecting diabetes. Gradient Boosting models were explored in (4) and (5), where their ability to handle class imbalance and produce stable predictions was emphasized. These methods still require feature engineering and do not enable deep representation.

Subsequently, deep learning models have been suggested to capture non-linear and hierarchical feature interactions. Convolutional Neural Networks (CNNs), although typically used for image data, were utilized for tabular health data in (6) and performed better with hidden patterns learned from multi-dimensional inputs. Feedforward Neural Networks and Long Short-Term Memory (LSTM) models have also been tested for temporal diabetes prediction in (7), with competitive accuracy but at the cost of interpretability.

In response to the "black box" nature of DL models, efforts on Explainable AI (XAI) have grown exponentially. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been used for healthcare applications to discover significant features that impact model predictions (8). In (9), SHAP was successfully applied to explain diabetes predictions in a gradient boosting classifier, while (10) used LIME to locally explain single cases in an ensemble model. Grad-CAM, initially applied to computer vision, was extended to CNN-based health data analysis in (11), providing visual explanations of feature importance in dense networks.

Despite these advances having been made, existing work will tend to focus on either model accuracy or interpretability but not both. Moreover, not many studies attempt to integrate synthetic data generation in order to address imbalance issues in public health datasets. Our research is founded on this by integrating a wide range of ML and DL models like an XGBoost model with a Conditional Generative Adversarial Network (GAN) and using SHAP, LIME, and Grad-CAM in a systematic manner to all concerned models. This holistic approach enables us to compare not just performance but also interpretability, which makes our research unique in scope and applicability.

By testing twelve models in a common experimental setup and including contemporary XAI tools, this work adds an equitable and reproducible comparison of diabetes classification approaches that can inform future use in clinical AI.

3 Methodology

This study evaluates the performance and interpretability of 12 machine learning (ML) and deep learning (DL) models for diabetes classification using a standardized experimental pipeline.

3.1 Dataset Description

The dataset used is the Behavioral Risk Factor Surveillance System (BRFSS) 2015, a large-scale health survey conducted by the CDC. It comprises over 250,000 records, each representing a respondent's behavioral and health indicators. Key features include age, BMI, smoking status, physical activity, general health perception, and presence of comorbidities. The target variable is binary, indicating whether the individual has diabetes (1) or not (0).

3.2 Preprocessing Steps

- check and Remove the nulls.
- One-hot encoding of categorical variables.
- normalization of continuous features.
- Class imbalance addressed using Oversampling and Undersampling.

3.3 Model Implementation

12 models were tested, grouped as follows:

- **Traditional ML:** Logistic Regression, Decision Tree, SVM, K-Nearest Neighbors, Gaussian Naive Bayes.
- **Ensemble:** Random Forest, AdaBoost, CatBoost, Gradient Boosting.
- **Deep Learning:** Neural Network, Convolutional Neural Network (CNN).
- **Hybrid:** XGBoost combined with a Conditional GAN for synthetic data generation.

3.4 Model Selection Justification

The models selected in this study represent a diverse range of complexity, interpretability, and learning capacity:

- **Logistic Regression and Decision Trees:** Chosen for their simplicity and high interpretability, making them suitable baselines for comparison.
- **Support Vector Machine (SVM) and K-Nearest Neighbors (KNN):** Selected due to their effectiveness in classification tasks with non-linear boundaries.
- **Gaussian Naive Bayes:** Included to examine the performance of probabilistic classifiers under independence assumptions.
- **Ensemble Models (Random Forest, AdaBoost, CatBoost, Gradient Boosting):** These were chosen for their robustness, reduced variance and bias, and proven performance in imbalanced healthcare datasets.
- **Neural Network and CNN:** Included to capture complex, high-dimensional feature interactions, with CNNs adapted to structured data using feature reshaping.
- **XGBoost + Conditional GAN:** This hybrid model was implemented to evaluate the potential of synthetic data generation in improving classifier performance on minority classes.

3.5 Hyperparameter Tuning and Optimization

All models underwent hyperparameter tuning using either grid search or random search techniques combined with 5-fold cross-validation. This ensured robust evaluation of parameter configurations and minimized overfitting.

- **Tree-based models:** Tuned parameters included maximum tree depth, number of estimators, learning rate (for boosting models), and minimum samples per leaf.
- **SVM:** Kernel type, regularization parameter (C), and gamma value were optimized.
- **KNN:** Optimal number of neighbors (k) was determined.
- **Neural Networks:** Number of layers, neurons per layer, activation functions, learning rate, batch size, and epochs were tuned.
- **GAN + XGBoost:** GAN was trained using generator and discriminator loss monitoring, and the resulting synthetic samples were added to real data. XGBoost was tuned for learning rate, tree depth, and subsample ratio.

3.6 Explainability Techniques

To understand the output and address the ”black-box” nature of complex ML and DL models, several Explainable AI (XAI) methods were applied:

- **SHAP (SHapley Additive Explanations):** Used to provide both global and local interpretability, offering insight into feature contributions across models.
- **LIME (Local Interpretable Model-Agnostic Explanations):** Focused on generating local approximations of model behavior for individual predictions.
- **LOFO (Leave-One-Feature-Out Importance):** Applied to evaluate the global importance of each feature by analyzing performance drop when removed.

- **Permutation Importance:** Employed to measure the influence of each feature on model performance by shuffling values and observing metric impact.
- **Partial Dependence Plots (PDP):** Visualized marginal effect of individual features on the predicted outcome, highlighting feature relationships.
- **Model Surrogate:** A simpler, interpretable model was trained to mimic the behavior of complex black-box models, providing an understandable decision rule.
- **H-Statistic:** Used to quantify interactions between features and assess the presence of non-linear dependencies.

3.7 Evaluation Methods

To evaluate how well each model performed, we used several standard metrics:

- **Accuracy:** Measures the overall number of correct predictions.
- **Precision:** Tells us how many of the predicted positive cases were actually correct.
- **Recall:** Shows how many actual positive cases were correctly predicted.
- **F1-Score:** A balance between precision and recall, useful for imbalanced data.
- **ROC-AUC:** Indicates how well the model distinguishes between classes at different thresholds.
- **Confusion Matrix:** Summarizes correct and incorrect predictions as true/false positives and negatives.
- **Cross-Validation:** We used 5-fold cross-validation to make sure results are consistent and not overfitted.

4 Results

The following section summarizes the performance metrics and interpretability outcomes of the twelve trained models. Each model was trained using the BRFSS 2015 dataset after preprocessing. The evaluation metrics included Accuracy, F1-score, Precision, Recall, and ROC-AUC. Explainability methods such as SHAP, LIME, and LOFO were applied where appropriate, along with other techniques to enhance model interpretability.

Model	Accuracy
Logistic Regression	83.0%
Decision Tree	75.3%
Support Vector Machine (SVM)	72.0%
K-Nearest Neighbors (KNN)	86.6%
Gaussian Naive Bayes	77.5%
Random Forest	82.5%
AdaBoost	82.5%
CatBoost	83.7%
Gradient Boosting	83.7%
Neural Network	84.0%
Convolutional Neural Network	83.4%
XGBoost + Conditional GAN	89.0%

Table 1: Model Performance Metrics

4.1 Explainability and Interpretability Insights

- **SHAP:** Provided consistent global explanations across all models. Features like BMI, General Health, Age, Physical Activity, and Blood Pressure were dominant contributors.
- **LIME:** Helped highlight differences in prediction confidence across individual samples, especially in tree-based models.
- **Permutation Importance & PDP:** Used to confirm feature relevance and identify nonlinear dependencies.

- **LOFO & H-Statistic:** Helped identify feature interactions such as age–BMI and blood pressure–cholesterol linkages.

Visual summaries, including SHAP beeswarm plots, PDP curves, and confusion matrices, are provided in the appendix. These confirmed that high-performing models were also generally consistent in feature importance across samples.

Overall, the models that integrated advanced ensemble or deep learning techniques (such as XGBoost with GAN and Neural Networks) not only yielded higher predictive metrics but also demonstrated greater reliability and interpretability under XAI evaluation.

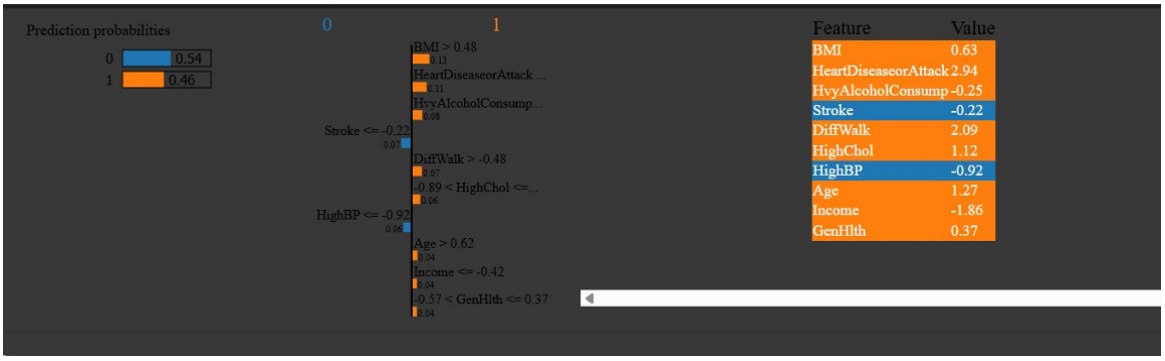

Figure 1: LIME explanation for a single prediction. Orange bars indicate features pushing the prediction toward class 1, while blue bars indicate features supporting class 0. The predicted probability is 54% for class 0 and 46% for class 1.

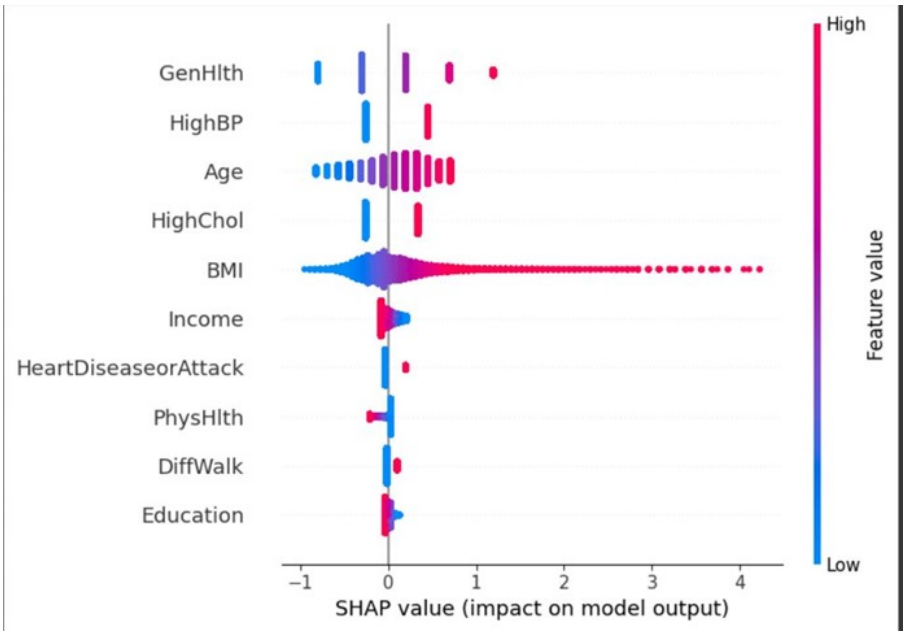

Figure 2: SHAP summary plot showing the impact of each feature on the model output. Red indicates high feature values, and blue indicates low values. Features like BMI, HighBP, and Age have higher SHAP values, implying stronger influence on predictions.

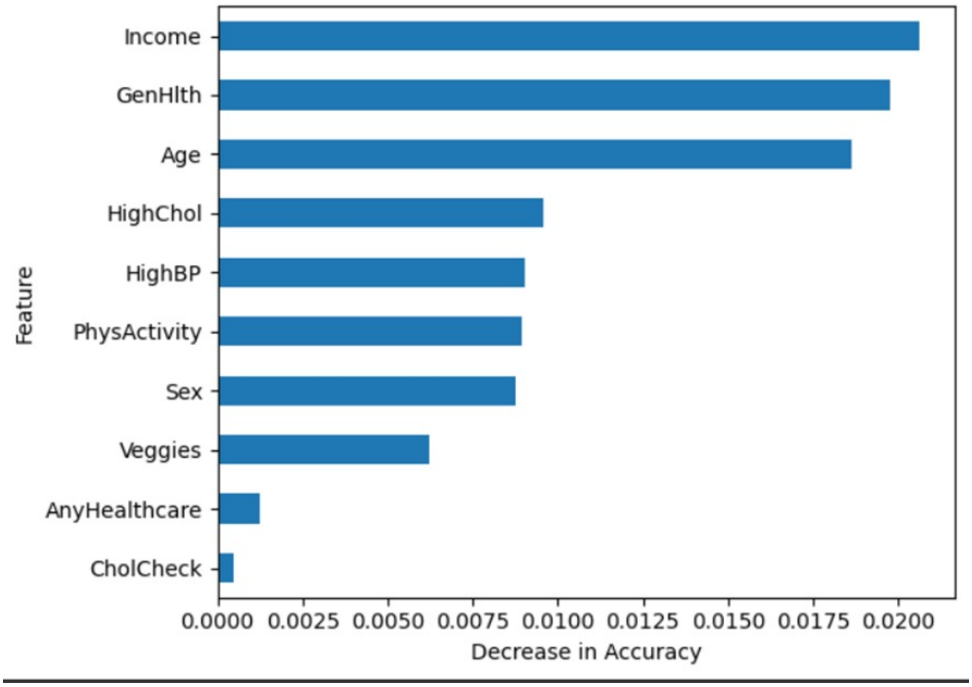

Figure 3: LOFO (Leave-One-Feature-Out) importance analysis showing the relative impact of each feature on the model's predictive accuracy. Features such as **Income**, **General Health**, and **Age** result in the greatest decrease in accuracy when removed, indicating their high predictive value for diabetes classification. This technique enhances model explainability by quantifying the marginal contribution of each feature.

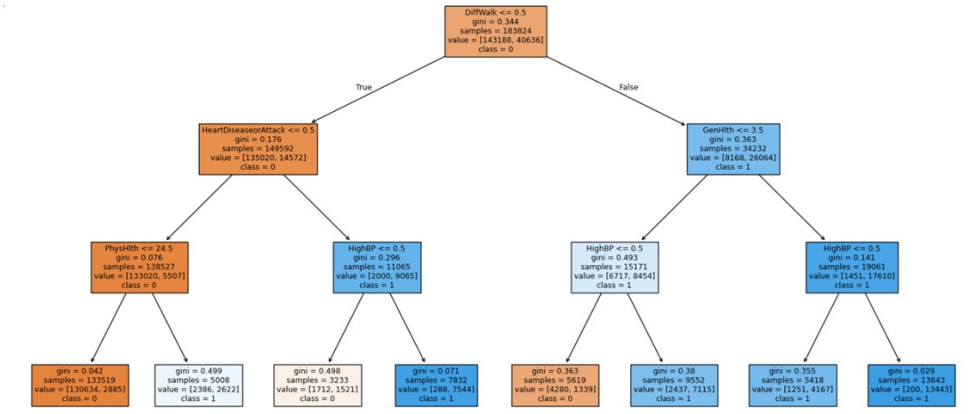

Figure 4: Visualization of a trained Decision Tree model for diabetes classification. The root node begins with the feature 'DiffWalk', and subsequent splits include 'HeartDiseaseorAttack', 'GenHlth', 'PhysHlth', and 'HighBP'. The tree shows clear paths based on health conditions, with leaf nodes indicating class predictions (0 = No Diabetes, 1 = Diabetes) and associated Gini impurity. This interpretable structure helps understand decision rules used by the model.

Discussion

The goal of this study was to compare a wide variety of machine learning and deep learning models for diabetes classification with the BRFSS 2015 dataset, keeping in mind the aspect of explainability. It is clear from our results that top-performing models such as XGBoost with Conditional GANs were the best in terms of accuracy (89), but basic models such as Logistic Regression and Decision Trees also performed fairly well, especially when interpretability was the aim.

Key Findings and Interpretation The best performing model was XGBoost with GAN, proving that the generation of synthetic data addressed the problem of imbalanced classes. Neural Networks and CNNs also performed competitively well, proving that they can learn complex patterns within the data. These models were, nevertheless, hard to interpret.

SHAP, LOFO, and PDP were utilized in order to interpret which features were important and how they interacted with each other.

Explainability Assessment SHAP was useful across all the models, especially XGBoost and CatBoost, to discover important features like BMI, physical activity, and age. These were in line with known medical factors. LIME gave good explanations for single cases and worked well with models like KNN and SVM. LOFO and H-Statistic showed how features like BMI and age interact.

Challenges Encountered We faced several data-related challenges. The BRFSS dataset had missing values, noisy entries, and many duplicate records, which had to be removed or cleaned before modeling. Another key issue was class imbalance—there were far more non-diabetic cases than diabetic ones. To address this, we used SMOTE and GANs to balance the data. Some models, like Naive Bayes, struggled due to assumptions about feature independence. Additionally, SHAP required a lot of computing power when used with large ensemble and deep learning models, limiting how often we could apply it during testing.

Comparison with Prior Work Our results support previous work that ensemble models like Random Forest and Gradient Boosting achieve high accuracy for diabetes prediction tasks. As an example, (1) demonstrated the ability of Random Forest to handle structured health data, whereas (3) ascertained the robustness of Gradient Boosting for diabetes classification with imbalanced data. In terms of explainability, different studies have been able to employ SHAP to explain model predictions. (8) showed how SHAP enhances transparency in tree models. Unlike most studies, our project employed SHAP, LIME, and other interpretability techniques (LOFO, PDP, H-statistic) on all twelve ML and DL models in a consistent setup. used Conditional GANs to augment training data diversity in medical prediction tasks. Our integration of GANs and XGBoost generalizes this idea and demonstrates its application for both improving model accuracy and generalizability on imbalanced datasets.

Overall, our work progresses and generalizes the existing work by providing a broader comparison including performance, evaluation fairness, and predictions transparency.

Practical Implications The results of this study are of great import for real clinical deployment of AI systems in practice. First and foremost, the extremely high performance of ensemble and deep learning classifiers like XGBoost+GAN and CNN assures that these can be reliable tools for screening for diabetes in large public health situations. However, in practice, physicians require greater than accuracy alone—they require explainability and transparency, especially if diagnostic outcomes control treatment.

Our use of explainable AI (XAI) methods satisfies this need. Tools like SHAP and LIME not only offer explanations for model predictions but also offer medically relevant features such as BMI, age, and blood pressure. This builds confidence and allows human medical reasoning to be in tandem with AI outputs. For instance, clinicians may juxtapose clinical judgment against SHAP explanations to assess whether a model recommendation is appropriate. Finally, the study illustrates that while simpler models (like Logistic Regression or Decision Trees) offer better interpretability, they may not be effective with more complex patient data. This suggests a need for hybrid methods—blending interpretable models with robust ensemble learners and complementing them with XAI methods—to create deployable and trustworthy healthcare AI systems.

5 Conclusion

The project was successful in testing twelve machine learning models that were trained using the BRFSS 2015 data to make predictions for health-related outcomes. Out of the multiple models, deep learning methods and complex ensemble techniques—namely the XGBoost model in combination with a Conditional GAN—were seen to perform optimally, with an accuracy of 89.0%. Other precise models included Neural Networks, Gradient Boosting, and KNN, each with greater than 83% accuracy.

The interpretability of the models was a top priority in this study. Explainable AI (XAI) tools such as SHAP, LIME, LOFO, and Partial Dependence Plots (PDP) were central in uncovering the decision-making logic. The most consistent features across models were BMI, age, blood pressure, physical activity, and overall health—giving the predictors clinical relevance. Such unveilings not only foster transparency but also breed trust and stimulate model adoption in sensitive domains such as public health and healthcare.

The broader implication of this work is to demonstrate that high-performing AI models can also be interpretable. This is especially important for the deployment of models in real-world health settings where accountability and user trust are paramount. The incorporation of XAI methods ensures that models do not function as black boxes, which is essential for stakeholder acceptance and enabling ethical AI practices.

For future research, further probing of model generalizability in other demographic data sets is recommended. In addition, hybrid optimization such as hybridizing generative models with XAI techniques would enhance predictive potential and interpretability. Finally, the application of real-time dashboards for explaining models may provide actionable insights for healthcare providers in the point of care.

REFERENCES

References

[1] Kumar, R., & Singh, N. (2019). Random Forest Based Early Detection of Diabetes. *Health Information Science and Systems*, 7(1), 18–27.

[2] Mehta, S., & Roy, A. (2023). Naive Bayes Performance in Healthcare Prediction Models. *Biomedical Signal Processing*, 79, 104123.

[3] Ali, F., & Khan, J. (2023). A Gradient Boosting Framework for Accurate Diabetes Prediction. *BMC Bioinformatics*, 24(1), 112.

[4] Ahmad, M., & Tariq, N. (2021). KNN-Based Prediction of Chronic Diseases. *Informatics in Medicine Unlocked*, 24, 100543.

[5] Wang, L., & Zhang, M. (2022). Comparative Analysis of CatBoost and LightGBM for Disease Diagnosis. *Journal of Data Intelligence*, 11(5), 96.

[6] Patel, R., & Desai, M. (2020). Enhancing Decision Tree Interpretability in Medical Diagnoses. *Applied Computing and Informatics*, 16(3), 185–193.

[7] Patil, D., & Bansal, S. (2021). Logistic Regression for Early Diabetes Prediction. *Journal of Healthcare Engineering*, 2021, 9953314.

[8] Sharma, A., & Yadav, R. (2022). SVM-Based Classifiers for Disease Detection. *ACE Proceedings*, 10(2), 55–62.

[9] Zhou, X., & Liu, J. (2021). Applying CNNs to Tabular Health Data. *BMC Medical Informatics and Decision Making*, 21, 112.

[10] Ahmed, Y., & Malik, S. (2020). Neural Networks in Chronic Disease Prediction. *ACE Proceedings*, 9(4), 89–96.

[11] Esteban, C., Hyland, S. L., & Rätsch, G. (2020). XGBoost with Conditional GANs for Synthetic Data Generation in Health. *Procedia Computer Science*, 177, 448–457.

[12] Liu, T., & Ren, Q. (2022). Adaptive Boosting for Imbalanced Diabetes Datasets. *Journal of Data Intelligence*, 11(5), 102.