Early Prevention of Diabetes Using Explainable Machine Learning

# Abstract

Diabetes, a prominent lifestyle disease, presents a significant challenge to global public health, with a rising number of individuals affected worldwide. The integration of machine learning models into healthcare offers a promising avenue for improving early detection and intervention strategies. This study investigates the effectiveness of three machine learning models Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) in predicting diabetes using a dataset of 520 samples with 16 features. The primary focus is on enhancing model interpretability through post-hoc explainability techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), which provide valuable insights into the factors driving model predictions.

To address class imbalance, the dataset undergoes preprocessing using the Synthetic Minority Over-sampling Technique (SMOTE), ensuring a balanced distribution of classes. The models are evaluated based on metrics such as AUC, F1 score, sensitivity, and specificity. The Random Forest model emerges as the most effective, achieving an AUC score of 0.99 and an F1 score of 0.95. Notably, the analysis identifies polyuria, polydipsia, age, and gender as significant predictors of early diabetes, highlighting the model's ability to discern critical risk factors.

The utilization of explainable AI techniques enhances the transparency and trustworthiness of machine learning models in healthcare. SHAP values offer both global and local explanations of model predictions, underscoring the importance of symptoms like polyuria and polydipsia in diagnosing diabetes. LIME provides localized explanations that further aid clinical decision-making by illustrating individual prediction factors. Despite the computational demands of these techniques, the study demonstrates their potential to support endocrinologists in identifying at-risk individuals and implementing preventive measures, ultimately contributing to more effective diabetes management.

**Keywords**: Explainable AI in diabetes, Early Prediction in diabetes, AI in healthcare

# Introduction

Modern society faces a significant burden from life-threatening diseases that can often be effectively controlled if diagnosed early. According to IDF[1], by 2045, approximately 783 million adults globally, will be living with diabetes.

Building upon predictive ML models, we employed post hoc explainability techniques to enhance interpretability and decision-making in the early detection of diabetes. This approach not only facilitates early detection of diabetes but also sheds light on the key variables driving these conditions, offering a deeper understanding of its causative factors and potentially guiding more effective intervention strategies.

In this study, we have constructed three predictive ML techniques, including RF, LR, and SVM to detect diabetes associated with sixteen predictors. Furthermore, we aim to answer the following three RQs:

1. What are the practical features of a task that predict diabetes?
2. Which ML model – has the highest ROC-AUC scores for predicting diabetes?
3. Can XAI be utilized to assist endocrinologists in predicting diabetes outcomes more effectively?

The rest of the paper is structured as follows. In the following, (A) a literature review of the XAI and ML techniques applied in the healthcare sector, (II) describes the Experimental design of our projects, (III) the experimental results of our XAI analysis, and lastly, (IV) addresses the conclusion, limitations and future work of this study.

## Small-Scale Literature Review

This section highlights the practical potential of XAI and ML in improving decision-making processes within the healthcare sector.

In the modern agile healthcare environment, ML is increasingly utilized to enhance patient care through predictive analytics and personalized treatment approaches. ML models are demonstrating considerable potential in several healthcare fields, including prostate health, oncology, and diabetes management[2, 3].

Type 2 diabetes, often referred to as the “silent killer,” can develop over many years with minimal warning signs, leaving individuals asymptomatic for extended periods.

Among the most widely implemented supervised learning methods for disease prediction and hospital outcome identification are (SVM), (DT), (RF), and (KNN).

According to[4], RF achieved the highest accuracy of 0.99 in predicting type 2 diabetes. Additionally, SHAP bar plots revealed that the primary predictors of this condition include 1) polydipsia, 2) polyuria, and 3) gender.

A comprehensive overview of various explainable machine learning (XAI) techniques, focusing on their types of explainability, strengths, and limitations. SHAP, which is based on cooperative game theory, provides both global explanations (the overall feature importance throughout the model) and local explanations (specific insights for individual predictions) to guarantee a fair attribution of feature importance [5]. Globally, it highlights overall feature importance across the model, while locally, it explains individual predictions. Its strengths lie in its theoretical rigor and fairness, making it widely applicable in high-stakes industries like finance and healthcare. However, SHAP can be computationally expensive, particularly for large models, and its results can be challenging to interpret in cases with complex feature interactions.

Although counterfactuals are useful, their effectiveness depends on the computational work necessary to produce viable, workable alternatives [6]. Counterfactual explanations, which focus on local explainability, allow users to explore "what-if" scenarios by demonstrating how small changes to input data can alter predictions. These explanations provide actionable insights, particularly valuable in healthcare, by offering clear suggestions for improving outcomes. However, generating realistic and practical counterfactuals can be challenging, especially in complex models, and requires significant computational resources.

Integrated gradients are especially helpful for comprehending complex neural networks because they provide local explainability by indicating the relative contribution of each feature to a model's prediction [7]. One of many benefits is in medical imaging, where it highlights the regions of an image that contributed most to the prediction. Integrated Gradients is computationally efficient but dependent on the choice of baseline and is limited to differentiable models, making it less versatile than SHAP.

For time-series data, various XAI techniques like temporal saliency maps and attention mechanisms provide both global and local insights. Despite their benefits, time-series data interpretation can be challenging, particularly for novices, and many methods have limited generalizability due to their model-specificity [8]. These methods are designed to handle the complexities of temporal dependencies, making them useful in fields like healthcare and finance. However, interpreting time-series models can be challenging, especially for novices, and many of these techniques are model-specific, limiting their generalizability.

Comparing these techniques, SHAP and time-series methods provide both global and local explainability, while counterfactual explanations and Integrated Gradients focus more on local insights. SHAP is model-agnostic, whereas Integrated Gradients and some time-series techniques are model-specific. While counterfactuals and SHAP both offer actionable insights, the former is more intuitive and user-friendly, making it accessible to non-experts. However, both counterfactuals and time-series methods require significant computational resources, especially for complex models. In essence, each method has its strengths and is suited to different contexts, balancing explainability, actionability, and computational feasibility depending on the application.

# Experimental Design

Our study utilizes the UCI dataset “Early prediction of diabetes”. It includes 520 samples with 16 features to predict early-stage diabetes. Our response variable, ‘Class’, is a binary variable that takes a value of 1 (Diabetes) and 0 (No Diabetes).

Figure 1 illustrates the proposed pipeline, which includes data pre-processing, feature selection, and model interpretability. The process begins with data pre-processing, followed by EDA and a Chi-square test. Feature importance is assessed using RF, and feature selection is performed with PCA.

To address the class imbalance, SMOTE is applied before implementing ML models. The dataset is split into an 80:20 ratio for training and testing, using standard performance metrics for classifier evaluation.

Additionally, we employ SHAP, an XAI tool, to enhance the reliability by elucidating decision-making processes.

A diagram of a data processing process

Description automatically generated

*Figure 1: Proposed Experimental Design*

## Data pre-processing

In this section, we will explore EDA and feature importance before constructing our predictive models.

### Exploratory Data Analysis (EDA)

#### EDA: Oversampling in Data

Figure 2 shows out of 520 patients, 320 samples have positive diabetes (62% of the sample), creating class inequality that will result in biased predictions. To address this, we employ SMOTE, which creates a balanced distribution of class.

#### EDA: Gender Distribution on Positive Class

In Figure 3, out of the total sample of 520, Females are 54.06% more likely to test positive for diabetes than Males, 45.94%. The gender column is one of the most informative features.

#### EDA: Chi-square Test

In Figure 4, the strongest positive correlation with the target variable is observed for 'polyuria' (0.67). The most substantial negative correlation is with 'gender' (-0.45). Therefore, the presence of polyuria increases the likelihood of a diabetes diagnosis, while being female is associated with a higher probability of diagnosis. To determine if there is a relationship between ‘class’ and other categorial, the Chi-square test of independence is applied. The hypothesis for this test is as follows:

*H*0: *Variables are independent; there is no association between the target variable and the features*  
*H*1: *Variables are not independent; there are associations between the target variable and the features*

**Inferences based on p-values**

1. p-value between “gender”,“polyuria”,” polydipsia”,’ sudden weight loss’,’ Weakness’,’polyphagia’,’Genital thrush’,’visual blurring’,’ irritability,’ partial paresis’,’muscle stiffness’,’ Alopecia’, ‘Age’ less than significance level (α = 0.05), we can reject the null hypothesis and conclude that there is a statistically significant association between these features and the target variables.
2. The p-value between “itching”, “delayed healing,” obesity” is greater than the significance level (α = 0.05); we cannot reject the null hypothesis and conclude that there is insufficient evidence to suggest an association between features and the target variables.

#### EDA: Outlier

In Figure 5, The age distribution for the target variable 'positive' shows a slight asymmetric pattern including some noise. Furthermore, most patients are in the age group range of 45-50.

### Feature Engineering

#### RF Feature importance

In Figure 6, the Random Forest (RF) model offers a clear measure of feature importance ranked according to their impact. The top four features—Polyuria, Polydipsia, Age, and Gender significantly influence the likelihood of early diabetes diagnosis.

A graph with blue and white bars

Description automatically generated with medium confidence

Figure 6: RF Feature importance

#### PCA Feature Selection

To further understand the significant patterns in the data, we employed PCA. Figure 7 presents a scree plot showing the proportion of variance explained by each principal component. Using the elbow method, we determined that two principal components should be retained.

Figure 8 displays the biplot of the first (PC1), which accounts for 24.4% of the variance, and (PC2), which explains 38.3% of the variance, a total of (62.7%). Features with high loading on PC1 include itching, age, delayed healing, muscle stiffness, visual blurring, and polyphagia, while high loadings on PC2 include gender, alopecia, and genital thrush.

In the following section, we will build our predictive models. To avoid introducing bias and to prevent potential underfitting, we decided not to exclude any features.

A graph with a red line and a red line

Description automatically generated

*Figure 7 : Scree plot*

A screen shot of a graph

Description automatically generated

*Figure 8 : Bi plot*

## Predictive Model Construction

In our study, to predict diabetes, the following three ML techniques are utilized. A brief introduction of the ML algorithm for model selection is given below:

### Random Forest

(RF) is a fast and efficient ensemble learning method that accommodates both categorical and continuous features, including those with missing values. It offers insights into feature importance and benefits from its ensemble approach, which reduces sensitivity to variations in the training data.

**Pros**: Resilient to overfitting because predictions are based on averaging. It also handles high-dimensional data effectively.

**Cons**: High substantial computational time for training due to its complexity.

### Logistic Regression

(LR) employs a sigmoid function to predict categorical outcomes and continuous variables, modelling the probability that a given input belongs to a specific category through the logistic function applied to a linear combination of input features.

**Pros**: It is straightforward to implement and interpret. It performs effectively on low-dimensional datasets and is efficient when features are linearly separable. It is less likely to overfit compared to more complex models.

**Cons**: Assumes linear relationships between target and predictor variables and struggles to capture complex interactions. It may require significant feature engineering and preprocessing to achieve good performance, especially if the data is not linearly separable.

### Support Vector Machine

(SVM) are effective for both linear and non-linear classification and regression tasks. They identify an optimal hyperplane in a high-dimensional space to separate data points from different classes. The objective is to maximize the margin between classes, which enhances classification performance, even for complex, non-linear problems.

**Pros**: Robust against outliers and, with the use of kernel tricks, can address complex issues.

**Cons**: Selecting an appropriate kernel and fine-tuning hyperparameters, such as cost (C) and gamma, can be challenging.

## Performance Evaluation

Table 1 and Figure 9 show that all three models achieved high AUC scores for predicting positive TPR and FPR. RF model had the highest AUC score of 0.99, F1 score (0.95), sensitivity score of 0.95 for identifying diabetes, and high specificity score (0.90) in identifying non-diabetes.

LR demonstrated relatively well in achieving high AUC score of 0.97, F1 score (0.92), sensitivity (0.92), however, it has the lowest specificity score (0.88) in identifying non-diabetes.

SVM achieved an AUC score of (0.98), a sensitivity score of 0.92, an F1 score of 0.95, and a superior specificity score (0.98) for predicting non-diabetes.

Considering computational cost and interpretability, we recommend selecting the RF model over the other two machine learning models.

Table 1: Three ML models metric comparison in predicting positive diabetes

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Model 1: RF** | **Model 2: LR** | **Model 3: SVM** |
| Accuracy | 0.93 | 0.90 | 0.94 |
| AUC | 0.992 | 0.965 | 0.975 |
| Precision | 0.94 | 0.92 | 0.98 |
| Sensitivity /Recall | 0.95 | 0.92 | 0.92 |
| F1 Score | 0.95 | 0.92 | 0.95 |
| Specificity | 0.90 | 0.88 | 0.98 |

A graph of a function

Description automatically generated

Figure 9: ROC Curve of three ML models with AUC Values

# XAI Experimental Results

## XAI - SHAP

In this section, we present SHAP values from Kernel Explainer to highlight the importance of both global and local features for diagnosing diabetes risk. Figures 10 and 11 show that patients with **polyuria** and **polydipsia** are at higher risk of diabetes. Figure 12 depicts various characteristics and their impact on diabetes prediction. It is evident that at a base value of 1.00, **female** patients have a higher risk of developing diabetes with the symptoms of itchiness, polyphagia, polyuria and polydipsia.

A graph with red and blue dots

Description automatically generated

Figure 10: Determinants of positive diabetes diagnosis and the direction of associations for local feature importance.

A graph with blue squares

Description automatically generated

Figure 11 : Global explanation of features for positive diabetes

A red text on a white background

Description automatically generated

Figure 12: Features significantly contributing to diabetes

## XAI- LIME

In Figure 13, LIME local explanations predicted a probability of 0.97 for diabetes in a 38-year-old female exhibiting mild symptoms of polyuria, polydipsia, alopecia, polyphagia, and obesity.

A screen shot of a computer

Description automatically generated

Figure 13: LIME Local explanations of predicting diabetes

# Conclusion

Polyuria and Polydipsia are well-known early signs of diabetes. In this study, we applied the SMOTE technique to three different machine-learning algorithms to identify these early indicators. Unlike previous research, we also integrated interpretable machine learning methods, specifically SHAP and LIME, to explain how our models arrived at predictions and to determine the relationships between target variables and predictors. Our findings revealed that the Random Forest (RF) model performed best, achieving an impressive F1 score of 95% and an AUC score of 0.99. These insights can aid endocrinologists in identifying individuals at risk of developing diabetes based on early symptoms, offering a preventive approach in diabetes management.

Throughout the project, an extensive analysis of explainable AI (XAI) techniques was conducted, though several limitations emerged. One significant challenge was scalability, particularly with SHAP, which becomes computationally intensive when applied to large datasets or complex models. Additionally, LIME, while popular for its simplicity, exhibited variability in its explanations, raising concerns about consistency and reliability. Addressing these limitations could be a focus of future work, such as exploring hybrid explainability approaches that combine different XAI methods to leverage their strengths while mitigating weaknesses. Furthermore, applying these techniques to alternative domains, like finance or criminal justice, could provide valuable insights into their generalizability and effectiveness in various fields.

Overall, this project highlights the crucial role of explainable AI in making machine learning models transparent, trustworthy, and actionable. By systematically evaluating LIME and SHAP, we have created a clear framework for selecting appropriate XAI techniques based on the needs of specific applications. As machine learning continues to influence high-stakes decision-making environments, explainability becomes increasingly vital. This project emphasizes the need for ongoing research and development to ensure that AI systems not only deliver powerful performance but also offer understandable and justifiable outcomes.

# References

[1] I. D. Federation. "Diabetes Facts and Figures " <https://idf.org/about-diabetes/diabetes-facts-figures/> (accessed May 07, 2024).

[2] H. Habehh and S. Gohel, "Machine Learning in Healthcare," (in eng), *Curr Genomics,* vol. 22, no. 4, pp. 3] J. Singh, S. Rani, and G. Srilakshmi, "Towards Explainable AI: Interpretable Models for Complex Decision-making," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, 18-19 April 2024, vol. 1, pp.

[4] M. S. Islam, M. M. Alam, A. Ahamed, and S. I. A. Meerza, "Prediction of Diabetes at Early Stage using Interpretable Machine Learning," in *SoutheastCon 2023*, 1-16 April 2023, pp.

[5] C. K. Leung, E. W. R. Madill, J. Souza, and C. Y. Zhang, "Towards Trustworthy Artificial Intelligence in Healthcare," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 11-14 June 2022, pp.

[6] Y. Wang and Z. Luo, "Causal Discovery and Causal Inference Based Counterfactual Fairness in Machine Learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4-10 June 2023, pp.

[7] D. Gaspar, P. Silva, and C. Silva, “Explainable AI for intrusion Detection Systems: LIME and SHAP applicability on Multi-Layer Perceptron,” *IEEE Access*, vol. 12, pp. 30164–30175, Jan.

[8] Y. Mei, Q. Chen, A. Lensen, B. Xue, and M. Zhang, "Explainable Artificial Intelligence by Genetic Programming: A Survey," in *IEEE Transactions on Evolutionary Computation*, June 2023, vol. 27, no. 3, pp.

# Appendix

## Imbalanced Distribution of Target variable

A pie chart with numbers and a diagram

Description automatically generated

Figure 2: Distribution of Target Variable

## Gender Distribution

A graph with numbers and a bar

Description automatically generated

Figure 3: Gender Distribution on Target Variable

## Correlation between target variable & Chi-Square Test

A screenshot of a graph

Description automatically generated

A blue and orange line with white text

Description automatically generated

Figure 4: Chi Square Test

## Outlier detection

A chart with a number of boxes

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

Figure 5: Outlier Detection between Age vs Target variable