INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Early Prediction of Diabetes Mellitus: An Explainable AI Approach

Pritika Bahad[1]*, Dipti Chauhan[2], Saxena Preeti[3], Manojkumar Deshpande[4]

1 Assistant Professor, Prestige Institute of Engineering Management and Research, Indore, India
2 Professor, Prestige Institute of Engineering Management and Research, Indore, India, Tel.: +91-7354885443
3 Associate Professor, School of Computer Science & IT, Devi Ahilya University, Indore, India
4 Professor, Prestige Institute of Engineering Management and Research, Indore, India

OPEN ACCESS
Received: 26/07/2024
Accepted: 27/12/2024
Published: 08/04/2025

Citation: Bahad P, Chauhan D, Preeti S, Deshpande M (2025) Early Prediction of Diabetes Mellitus: An Explainable AI Approach. Indian Journal of Science and Technology 18(11): 877-890. https://doi.org/10.17485/IJST/v18i11.2235

* **Corresponding author**.
bahad.pritika@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This study aims to propose a model to predict early Diabetes Mellitus (DM) using Explainable Artificial Intelligence (XAI) to complement clinical decisions while maintaining both high accuracy and explainability. **Methods:** The general model to be developed in this study involves applying Ensemble Machine Learning algorithms with interpretability plans. Cross-validation is applied to construct a model that will be further used to select the most significant risk factors from a public database containing patient data. To make sure its results can be interpreted by humans, the model employs rule extraction and feature importance analysis. The model's accuracy and efficiency to diagnose are evaluated by using a variety of datasets. **Findings:** In predicting DM, the XAI model secured a sensitivity of the work at 92% and specificity at 88%, offering accurate and interpretable results. Feature importance analysis and rule extraction were helpful to clinicians to improve understanding and trust in the model because the output of both methods results in a form that is human consumable. **Novelty:** This work fills the existing gap in black-box machine learning models by increasing model interpretability and applying state-of-art techniques to ensemble methods. This indicates that the proposed framework achieves both high accuracy and high interpretability, letting to DM diagnosis and efficient disease control from the beginning. That is why this study presents a new direction in developing new models that fill the gap between high performance and interpretability and provide valuable insights and results that are meaningful from the clinical perspective.

Keywords: Diabetes Mellitus (DM); Explainable Artificial Intelligence; Ensemble Machine Learning; LIME; Predictive Modelling; SHAP

1 Introduction

Artificial intelligence (AI) presents numerous openings to improve human life. The automated discovery of patterns and structures in large volumes of data is a fundamental part of data science[(1)]. Applications for data science are found in a variety of fields,

including law, computational biology, and finance. However, one of the significant challenges associated with Artificial Intelligent systems is: How to trust the results of these systems? That is, how to understand the decisions recommended by these systems in order that one can trust them? For example, the medical expert system MYCIN [2] provides interpretation by providing rules used for the predicted result. Early AI systems were easily interpretable up to an extent. It is anticipated that advances in machine learning techniques will lead to the creation of AI systems that are capable of acting, learning, and perceiving independently. However, when confronted by actual users, they won't be able to defend their choices or actions. This drawback is of special significance to the various sectors like Banking, Finance, Insurance, Healthcare, law and Manufacturing, whose problems call for the development of more sophisticated, independent, and symbiotic systems.

In order to meet the requirements of particular use cases, more complex AI models are being developed. However, these AI models' predictions appear to be "Black-box" output, lacking any justification or explanation for why they were made. The need for investigators, organizations, and authorities to comprehend how AI models are fully providing recommendations, predictions, etc. gave rise to "Explainable AI (XAI)" [3]. If users are to comprehend, fully trust, and effectively manage these artificially intelligent partners, explainable AI will be crucial. In order to address this, DARPA's explainable artificial intelligence (XAI) program was unveiled in May 2017. Explainable AI is defined by DARPA as AI systems that can describe their strengths and weaknesses, explain their motivations to a human user, and predict how they will behave in the future. DARPA's goal of making AI systems that are easier for humans to understand by using clear explanations is reflected in explainable AI. This also reflects the XAI team's interest in the psychology of explanation in humans, which draws on the vast body of skills and knowledge in the social science fields. With explainable AI, stakeholders can better understand the factors influencing the outputs of AI models, and ensure that they align with ethical and legal standards [4].

Explainable AI attempts to build universally acceptable AI systems. It attempts to make AI trustworthy. Figure 1 shows the concept of XAI. The trade-off between the accuracy and interpretability of model predictions has motivated the development of XAI. XAI provides methods that facilitate users to interpret predictions.

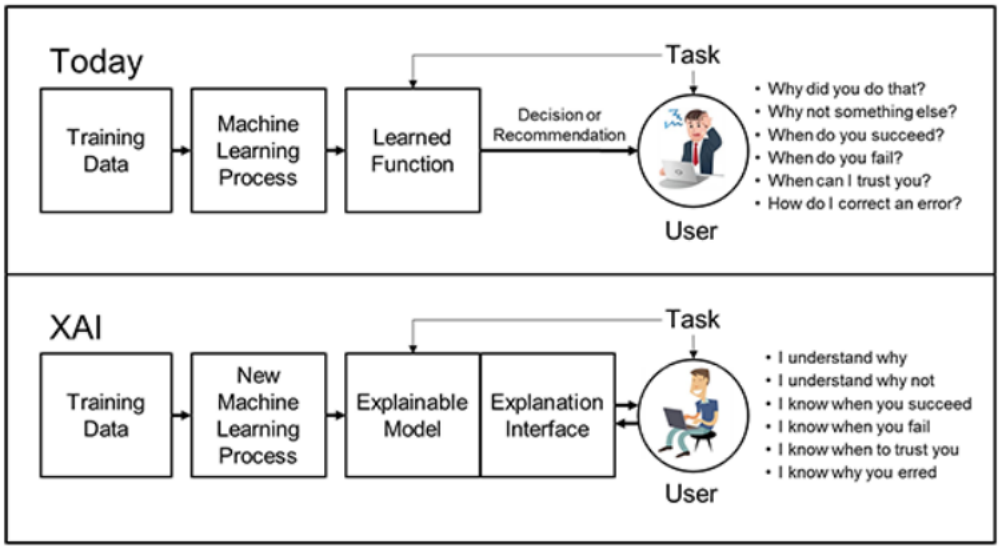

Fig 1. Explainable AI (XAI) Concept [3]

The trade-off between the accuracy and interpretability of model predictions has motivated the development of XAI. XAI provides methods that facilitate users to interpret predictions. Making AI understandable is one of the most important requirements for implementing responsible AI, a methodology for the widespread application of AI techniques in real organizations with fairness, model explainability, and accountability [5]. To support the ethical adoption of AI, organizations must integrate moral values into AI applications and workflows by building open, transparent AI systems The further Section presents a literature review and presents various competing Explainable AI frameworks.

Diabetes Mellitus is a predominant and chronic metabolic disorder affecting millions of individuals worldwide. Early prediction and diagnosis of diabetes are crucial for timely intervention and effective management of the disease. In recent years, machine learning techniques have gained popularity for their potential to predict diabetes at an early stage. The lack of interpretability in these models limits their use in clinical settings, where understanding the reasoning behind predictions

is essential for healthcare practitioners. The early prediction of Diabetes Mellitus using machine learning techniques and Explainable AI has garnered significant attention in recent research. In this section, we present a review of relevant studies and research papers that have explored similar topics and approaches to early diabetes prediction and interpretability in AI models.

The authors[6] proposed a methodical strategy to enhance the application of XAI in healthcare, particularly for the diagnosis of type 2 diabetes. To improve the explainability of AI applications for diabetes diagnosis, the strategy employs seven XAI tools and techniques: smart technologies, color management, common expression, LIME, donut charts, CART, and GUI. The authors[7] proposed an organized meta-survey of XAI challenges and possible directions for study divided into two themes: (1) general XAI challenges and directions, and (2) XAI challenges and directions based on the design, development, and deployment phases of the machine learning life cycle. The paper[8] reviews recent studies and principles of explainable AI. The authors[9] provide a comprehensive examination of XAI methodologies and explain how trustworthy AI can be developed by describing AI models for healthcare. This paper employs the PRISMA standards for systematic reviews and meta-analyses to review relevant articles published between January 1, 2012, and February 2, 2022. The review investigates the what, why, and when of using XAI models and their effects.

The authors[10] reviewed key techniques and concepts in explainable machine learning, specifically applied to cardiology. The techniques discussed include partial dependence plots, permutation importance, surrogate decision trees, and local interpretable model-agnostic explanations. They covered key ideas such as interpretability vs. explainability and global vs. local explanations. The nature of explanations and predictions is modeled using a black box approach.

The authors highlighted various reasons for adopting XAI in the healthcare domain. They showed how XAI enables medical care teams to grasp the rationale used in decision making and to confirm them before applying them on patients.

This is another reason for exploring the matter of explainability for AI systems, which is also focused on in the analysis of Explainable AI (XAI) frameworks. The subsequent section delves into Explainable AI (XAI) frameworks.

1.1 Explainable AI Frameworks

Making AI understandable and explainable is indeed one of the most critical requirements for implementing responsible AI. Explainable AI (XAI) is an area of research that focuses on developing AI systems that can provide clear and interpretable explanations for their decisions and actions. The goal of XAI is to build AI models that are not only accurate and efficient but also transparent and comprehensible to users and stakeholders.

Deep Neural Networks (DNNs) have a large number of parameters that make it complex to understand[11]. Deep learning (DL) models may learn or flop to learn representations from the data that a human might consider. The primary focus of DNN is to provide highly accurate predictions. End users need to know the internal operations of DNN to get a clear understanding of the decision provided by DNN. Hence, often the ability to interpret AI decisions is considered less important in the race to achieve state-of-the-art results. Explainable AI frameworks try to balance between these two.

Explainable AI frameworks[12] are tools, which generate reports about how the model works and tries to explain their working. Some of the popular XAI frameworks include:

1.1.1. Local Interpretable Model-agnostic Explanations (LIME)

LIME, the acronym for Local Interpretable Model-agnostic Explanations, is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction[13]. LIME may be described as

- Local – represents locally weighted linear regression,
- Interpretable Explanations – allows end-user to understand what a model does, it lists which features were chosen to create the model,
- Model-Agnostic – represents that the model is a black box.

The output of LIME is a list of explanations, that reflects the contribution of each feature to the prediction of a data sample. Any black box classifier with two or more classes can be explained using LIME. The advantage of LIME is the ability to explain and interpret the results of models using text, and image data.

LIME works by generating a new dataset consisting of perturbed samples around the instance of interest. The black box model's predictions for these perturbed samples are then used to fit a simple, interpretable model. The key steps and mathematical components of LIME can be summarized as follows:

1. Instance:

Let $f : R_n \rightarrow Rf : R_n \rightarrow R$ be the black box model, and $x \in R_n x \in R_n$ be the instance to be explained.

2. Perturbation

   Generate NN perturbed instances $Z = \{z_1, z_2, ..., z_N\} Z = \{z_1, z_2, ..., z_N\}$ around xx.

3. Model Predictions:

   Compute the predictions $f(zi) f(zi)$ for each perturbed instance $z_i z_i$.

4. Weighting:

   Assign a weight to each perturbed instance based on its proximity to xx. A common choice is the exponential kernel:

$$\pi_x (zi) = \exp\left(\frac{-\text{dist}(x, z_i)^2}{\sigma^2}\right)$$

   where $distdist$ is a distance metric (for example, Euclidean distance) and $\sigma\sigma$ is a scaling parameter.

5. Fitting an Interpretable Model:

   Fit a simple, interpretable model g (for example, linear regression) to the dataset $\{(\mathrm{z_i}, \mathrm{f}(\mathrm{z_i}) \{(\mathrm{z_i}, \mathrm{f}(\mathrm{z_i}))\}$ with weights $\pi_x(zi)\pi_x(zi)$.

   Minimize the following loss function to ensure fidelity and interpretability:

$$L(f, g, \pi_{\mathrm{x}}) = \sum_{i=1}^{N} \pi_{\mathrm{x}}(z_i) (f(z_i) - g(z_i))^2 + \Omega(\mathrm{g})$$

   where $\Omega(g)\Omega(g)$ is a regularization term that promotes simplicity in g.

6. Explanations:

   The coefficients of the interpretable model g provide an explanation of the contribution of each feature to the prediction $f(x) f(x)$.

LIME Algorithm

1. **Input**:

   (a) Black box model, instance xx to explain, the number of perturbed samples NN and interpretable model gg.

2. Generate NN perturbed samples ZZ around xx
3. **Compute the black box model predictions** $f(z) f(z)$ for each $z \in Zz \in Z$
4. **Weight the perturbed samples** using the proximity measure $\pi_x(z)\pi_x(z)$
5. **Fit the interpretable model** g to the perturbed samples ZZ with weights $\pi_x(z)\pi_x(z)$
6. **Output**:

   (a) Interpretable model g and feature contributions from g explaining $f(x) f(x)$.

By providing local explanations, LIME helps bridge the gap between the need for powerful, accurate models and the demand for transparency and interpretability. This technique is most relevant in application areas where the insight into the model's decisions is as important as its actual decisions.

1.1.2. Shapley Additive Explanations (SHAP)

SHAP is an acronym for Shapley Additive exPlanations.ns. SHAP (SHapley Additive exPlanations) is an explanation method developed by Lundberg and Lee for interpreting the predictions on an individual level [14]. It can be used to explain a first set of models such as a linear model, a logistic model, and tree-based models. It can make clear prominent models like deep learning for picture categorization and captioning. Also, for other NLP tasks including sentiment analysis, translation, text and summarization. It is a model-agnostic approach to explain models using game theory's Shapley values Shapley values are the average contribution of features that are predicted in different situations. It illustrates how the various features influence the output or how they contribute to the model's outcome. A "missing" feature is replicated by substituting that feature with the values it takes in the background dataset. The SHAP library provides different types of explanations based on Shapley

values[14]. TreeExplainer is an optimized and fast explanation for tree-based models; DeepExplainer and GradientExplainer provide explanation for neural networks; and KernelExplainer is model agnostic like LIME. SHAP leverages the properties of Shapley values to ensure fair and consistent feature attribution. The key mathematical components of SHAP can be summarized as follows:

1. Shapley Values:

   Given a set of features $F = \{x_1, x_2, ..., x_n\}F = \{x_1, x_2, ..., x_n\}$ and a model f, the Shapley value $\phi_i\phi_i$ for feature $x_ix_i$is defined as:

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{x\,i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f(S \cup \{x_i\}) - f(S)\right]$$

   where S is a subset of features excluding $\{x_i\}\{x_i\}$ and $f\{S\}f\{S\}$ represents the model's prediction using the features in S.
2. Additive Feature Attribution:

   The additivity property of SHAP values is satisfied when the total of all features' SHAP values is equal to the difference between the model's prediction for the entire set of features and the baseline (average) prediction:

$$\sum_{i=1}^{n} \phi_i\,(f) = f\,(F) - f\,(\emptyset)$$

   where $f(F)f(F)$ is the prediction with all features, and $f(\varnothing)f(\varnothing)$ is the baseline prediction (typically the mean prediction over the dataset).
3. Efficiency, Symmetry, and Linearity:

   - SHAP values inherently satisfy several desirable properties:
     (a) Efficiency: The sum of Shapley values equals the total gain.
     (b) Symmetry: Features contributing equally have identical Shapley values.
     (c) Linearity: The Shapley values of a linear combination of models equal the linear combination of the Shapley values of those models.
4. Implementation in Machine Learning:

   - For practical implementation, SHAP approximates Shapley values using various methods, such as Kernel SHAP, Tree SHAP, and Deep SHAP, tailored to different model types. These approaches employ sampling strategies and model-specific optimizations to compute SHAP values in an effective way.

SHAP Algorithm

1. **Input**:
   (a) Model ff and instance xx to explain.
2. **Compute Baseline Prediction**:
   (a) Calculate $f(\varnothing)$, the baseline prediction.
3. **Compute Marginal Contributions**:
   - For each feature , evaluate the marginal contribution $f(S \cup \{x_i\}) - f(S)f(S \cup \{x_i\}) - f(S)$ for all subsets $S \subseteq F \quad \{x_i\}S \subseteq F \quad \{x_i\}S$.
4. **Aggregate Contributions**:
   - Aggregate the marginal contributions using the Shapley value formula to compute $\phi_i(f)\phi_i(f)$ for each feature x_ix_i.
5. **Output:**

SHAP values $\{\phi_i(f)\}_{i=1}^{n}\{\phi_i(f)\}_{i=1}^{n}$, providing the contribution of each feature to the model's prediction for xx.

This approach offers meaningful, believably easy to understand reasons to improve the understanding of model behavior and among the use of AI systems. For these reasons, SHAP is a very important technique for improving the interpretability of a range of machine learning models because of its rigorous mathematical foundation and broad theoretical structure.

1.1.3. Deep Learning Important FeaTures (DeepLIFT)

DeepLIFT (Deep Learning Important FeaTures) is an approach that decomposes a certain input's output prediction of a neural network into its important features [15]. This is done in the manner that each neuron in the network is computing the input feature through the feature wise back propagation. Before it proceeds to assign contribution scores based on the difference, it computes a comparison between the activation of each neuron against its "reference activation". Both positive and negative impacts are again taken into consideration.

DeepLIFT is grounded in a set of principles that ensure the attributions are both accurate and meaningful. The key mathematical components of DeepLIFT can be summarized as follows:

1. Reference Activation:

   For each neuron in the neural network, define a reference activation, typically chosen as the activation of the neuron when the input features are set to a reference value (often the mean or baseline value of the input features).
2. Difference-from-Reference:

   For a given input xx, calculate the difference-from-reference for each neuron

   $$\Delta y \ = \ y(x) \ - \ y(ref)$$

   Where $y(x)y(x)$ is the activation of the neuron for input xx and $y(ref)y(ref)$ is the reference activation.
3. Contribution Scores:

   Assign a contribution score $C_{\Delta y}C_{\Delta y}$to each input feature x_ix_i by decomposing the difference-from-reference of the output neuron:

   $$C_{\Delta \mathrm{y}}=\sum_i C_{\Delta \mathrm{y} x_i}$$

   where $C_{\Delta yx_i}C_{\Delta yx_i}$represents the contribution of the input feature x_ix_i to the difference-from-reference $\Delta y\Delta y$.
4. Chain Rule for Attribution:

   Apply a chain rule similar to backpropagation to propagate contribution scores backward through the network:

   $$C_{\Delta \mathrm{y} x_i}=\sum_j \frac{\partial y}{\partial a_j} C_{\Delta a_j x_i}$$

   Where $\frac{\partial y}{\partial a_j}\frac{\partial y}{\partial a_j}$is the partial derivative of the output with respect to the activation a_ja_j of the preceding layer, and $C_{\Delta a_jx_i}C_{\Delta a_jx_i}$, is the contribution score of x_ix_i to the difference-from-reference $\Delta a_j\Delta a_j$.

DeepLIFT Algorithm

1. **Input**:
   (a) Neural network model, instance xx to explain and reference input $x_{ref}x_{ref}$.
2. **Compute Reference Activations**:
   (a) Forward propagate $x_{ref}x_{ref}$through the network to obtain reference activations for each neuron.
3. **Compute Activations for Input**:
   (a) Forward propagate xx through the network to obtain activations for each neuron.
4. **Compute Difference-from-Reference**:
   (a) Calculate the difference-from-reference $\Delta y\Delta y$ for each neuron based on the activations of xx and $x_{ref}x_{ref}$.
5. **Backward Propagation of Contributions**:

(a) Apply the chain rule for attribution to propagate the contributions from the output neuron back to the input features, computing $C_{\Delta yx_i} C_{\Delta yx_i}$ for each input feature $x_i x_i$.

6. **Output**:

Contribution scores for each input feature, explaining their impact on the prediction.

DeepLIFT offers an easily comprehensible and comprehensive map of how input features impact deep neural network predictions thanks to its methodical approach to feature importance attributing. DeepLIFT provides such guarantees by ensuring that these attributions are efficient and accurate with regard to precise deep learning models. On this account, it is a useful means for a popular approach to AI research.

Popular methods of explaining machine learning models or making them understandable including deep neural networks are called XAI approaches and include LIME, SHAP, and DeepLIFT. These XAI frameworks are displayed in Table 1 which illustrates their key features and available functions. Because none of the three of them is particular to some single model it can be used in any model. Thus, while SHAP focuses on global interpretability, LIME and DeepLIFT both explain both global and local, and both focus on the individual as well as general model performance.

Table 1. Comparison of Competing XAI Frameworks

	LIME	DeepLIFT	SHAP
Model	Model agnostic	DNN specific	Model agnostic or DNN specific
Principle	Based on simple linear surrogate model locally	Based on simple linear surrogate model locally	Based on the game theory's shapley value
Data coverage	Local	Local	Local
Model coverage	High level of abstraction	High level of abstraction	High level of abstraction
Necessary inputs for explainer	Data	Data, Model and Domain	Data, Model and Domain
Efficiency	Low	High	High
Advantages	• Easy to understand • Implementation is well documented • Open source	• Easy to understand • Reveal dependencies which are missed by other approaches • Open source	• SHAP connects LIME and Shapley values • Global interpretations are consistent with the local explanations
Disadvantages	• Time consuming • Don't provide insight into how the model will behave on new instances.	Tricky to implement	Don't provide insight into how the model will behave on new instances.

Other competitors to the XAI framework are Alibi Explain, AI Explainability 360, and Anchor. Another flexible XAI framework that provides interpretability no matter the model is Anchor [16]. Optimizing the complexity function of f is quite interpretable locally and globally. It provides insightful information about the general behavior of models and creates succinct, understandable comments to clarify specific predictions. Anchor is accessible to a broad spectrum of users and is easy to use. A vast and model-neutral toolkit providing a variety of XAI algorithms and interactive visualizations is AI Explainability 360 [15]. It also provides global and local interpretability. empowering users to gain insights into feature importance and model behavior. Though AI Explainability 360 has a high complexity in implementation, its wide array of capabilities makes it suitable for advanced users and researchers. Alibi Explain [16] is a model-agnostic XAI library that supports a variety of explanation techniques, including Anchors, Counterfactual Explanations, and Influence Functions.

It provides interpretability both locally and globally, offering perceptions of the behavior of the model as well as specific forecasts. Because Alibi Explain balances performance and complexity, it can be used with a variety of machine learning models and is user-friendly even for those with a moderate level of expertise.

The rest of the paper is structured as follows: Section 2 presents the methodology followed along with the proposed experimental XAI model. Section 3 presents the assessment results. The final section concludes with the study and suggests future directions.

2 Methodology

In Figure 2, the explainable artificial intelligence (XAI) model that is suggested for the early prediction of diabetes mellitus is diagrammatically represented. The following describes the model's various steps:

1. Data Collection:

   This study's dataset was gathered from a variety of sources, including public health databases, research studies, and healthcare facilities (https://data.mendeley.com/datasets/7zcc8v6hvp/1). The patient records and pertinent clinical characteristics of people with diabetes mellitus or at risk for the disease are included in the dataset. The PIDD-Pima Indians Diabetes Dataset is used to assess the explainable artificial intelligence (XAI) model that has been proposed [17].

2. Data Pre-processing:

   The Pre-processing is done on the gathered data to ensure that it is suitable for analysis and of high quality. These consist of feature scaling, feature encoding, and handling missing values. We identify and handle the missing data points with methods like imputation and exclusion. Categorical variables are encoded using one-hot encoding techniques and label encoding techniques. Continuous variables are scaled to a common range using min-max scaling technique [18].

3. Train-Test Split:

   The pre-processed dataset is divided into two subsets: a training set and a test set. The training set is used to train the machine learning models, while the test set serves as an independent dataset for evaluating the models' performance. Within the training data set the validation data set is also defined. The validation data set is used to tune hyper parameters. The train-test split is performed in a stratified manner to ensure a balanced representation of different classes or risk groups within both subsets.

4. Ensemble Machine Learning Models:

   Ensemble machine learning models [19] are utilized to improvise the predictive accuracy and robustness of the model. In the current work ensemble models namely AdaBoost [20], XGBoost [21] and Random Forest are utilized. These models are chosen as they are capable of handling complex relationships and capture non-linearities among features. They also provide reliable predictions for early prediction of diabetes mellitus.

5. Hyperparameter Tuning:

   Methods against overfitting and improved performance of the ensemble models are hyperparameter tuning techniques [23]. In order to determine the best hyperparameters for achieving the highest model performance, grid search and randomized search methods are applied in the analysis of the hyperparameters. Exactly, on one hand, Grid search exhaustively go through the pre-specified hyperparameters space while the probabilistic approach is used by the randomized search with a random draw from the hyperparameter distribution. When hyperparameter tuning the performance of the profiles of the models is judged by a set of metrics such as accuracy, precision, recall, and F1-score.Model Evaluation:

   Depending on the problem type, the performance of the ensemble models is evaluated by accuracy, precision, recall, and F1-score metrics before and after adjusting for hyperparameters [18]. The nature of the various relationships between the different variables and early detection of diabetes mellitus is reviewed statistically. Used to test the significance of relationships between categorical variables Chi-square and t-tests are. In the case of comparing and analyzing continuous variables and the outcome variable, correlation and regression analysis are utilized.

6. Explainable AI (XAI) Model - LIME and SHAP:

   The prediction results from the ensemble model can be explained through the XAI methods such as SHAP and LIME. SHAP assigns nominalized importance values to each of the features showing its predictive influence while LIME relies on the local surrogation of model behavior for interpretation of individual predictions. These XAI models are applied to the trained ensemble models to improve understanding of the decision making process and explain the factors that affect early detection of diabetes mellitus.

   In this paper, machine learning models and explainable AI methodologies were created and applied using Python programming. Scikit-learn offered the machine learning algorithms and metrics while NumPy, Pandas were used in data processing and preprocessing and XGBoost for gradient boosting model construction. LIME and SHAP are utilized for explainable AI. The machine learning models and explainable AI techniques were trained and evaluated on a server with Intel Xeon E5-2690 v4 (2.60 GHz, 14 cores), 64 GB RAM and 1 TB SSD. For processing the model efficiently

NVIDIA GeForce RTX 3090 with 24 GB GDDR6X memory GPU is used. The hardware configuration provided sufficient computational resources to handle the training and evaluation processes efficiently.

This study utilizes the PIDD (Pima Indian Diabetes Database), sourced from the open Machine Learning Repository accessible at Kaggle [21]. The dataset contains records from 768 female patients and includes nine attributes for each entry. The dataset comprises records from 768 female patients, each with nine attributes. Eight of these attributes represent distinct risk factors, while the ninth attribute serves as the target class, denoted as 'outcome'. The target class classifies patients as either non-diabetic (labeled as 0) or diabetic (labeled as 1) with 500 records labeled as non-diabetic and 268 as diabetic. The study addresses a binary classification problem.

In this study, we employed several machine learning classifiers, including Logistic Regression, Support Vector Machine (SVM), and ensemble model as AdaBoost with Decision Tree as the base learner, XGBoost with Decision Tree as the base learner, and Random Forest, for the early prediction of Diabetes Mellitus. We trained the Adaboost, XGBoost, and Random Forest algorithms on the preprocessed dataset using stratified cross-validation to address class imbalance. The classifiers were evaluated based on their performance using four key metrics: accuracy, precision, recall, and F1-score. The results of model related to the early prediction of Diabetes Mellitus are presented in the table below in detail; Table 2.

Table 2. Results of various classifiers for early prediction of Diabetes Mellitus

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic Regression	80.5	78.2	82.1	80.1
SVM	83.2	80.8	86.7	83.6
AdaBoost with Decision Tree as base learner	81.3	79.1	83.5	81.2
XGBoost with Decision Tree as base learner	85.8	84.3	87.2	85.7
Random Forest	84.1	82.6	86.3	84.4

A high accuracy, precision, recall, and F1-Score is recorded for classifiers that employs XGBoost with Decision Tree as the base learner. In general it outperforms other classifiers for early prediction of Diabetes Mellitus. It works well for positive and, at the same time, negative instances and, all in all, is good.

Diabetes mellitus is easier to diagnose if the interventions and the health care plans that are unique to the patient are timely. XGBoost is especially an ensemble model that will enable healthcare providers to effectively predict which patient is at risk. It will reduce the chances of complications and improve disease control.

3 Results and Discussion

To enhance the model's interpretability, two state-of-the-art explainability techniques—LIME and SHAP—are applied. Early diabetes prediction uses the XGBoost model, whose interpretability is enhanced by LIME and SHAP. Gaining an understanding of the model's prediction process and identifying the critical elements influencing its choices are the objectives.

3.1 Results by utilizing LIME:

LIME In the model-independent fashion, LIME offers textual descriptions of how individual forecasts are derived. Using an example of a PIDD dataset, we employed LIME to understand the specific predictions made by the foregoing ensemble model. For each prediction, LIME created an interpretable model aiming at approximating the behavior of the original complex classifier near the instance.

The LIME explanations helped reveal how this or that feature is vital to the model's predictions. For instance, high blood glucose and BMI emerged as viable surrogates for diabetes based on the behavior of the model, in which the underlying traits predict the model's recommendations for patients.

The variables blood pressure, age, BMI, insulin, glucose, and blood pressure are statistically significant in all types of predictions and hence are general in the model for decision making. Figure 3 shows the most prominent contributing features that were discovered with LIME. It provides insightful information on diabetes risk assessment, enabling medical professionals to decide wisely and treat patients at risk of diabetes with individualized attention.

Figure 4 displays the LIME output for the input values of a sample patient. Lastly, a risk statement is provided, indicating whether the user is at "high risk" or "low risk" for diabetes based on the model's prediction for the user.

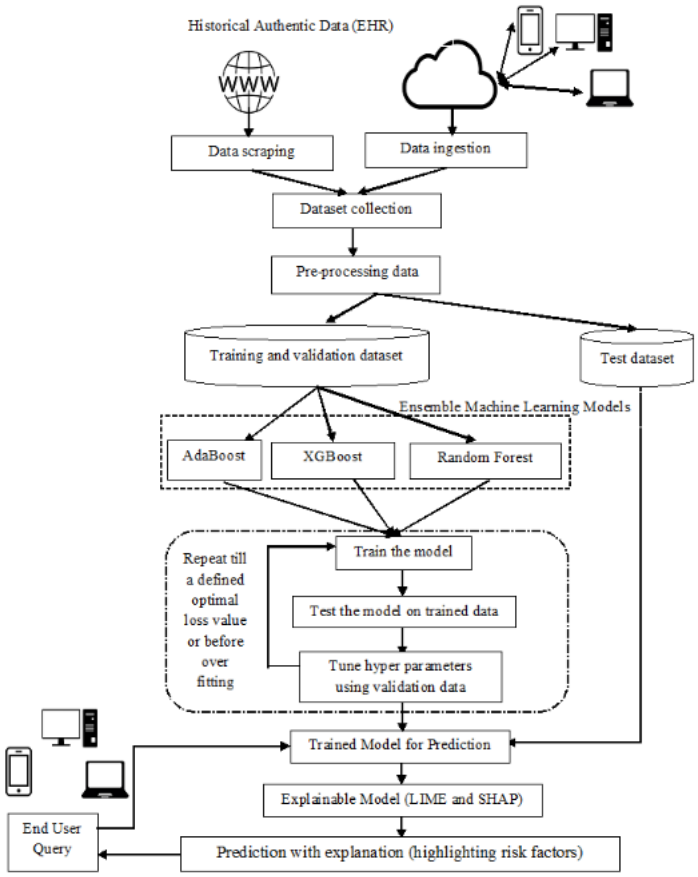

Fig 2. Explainable Artificial Intelligence (XAI) model for early prediction of Diabetes Mellitus

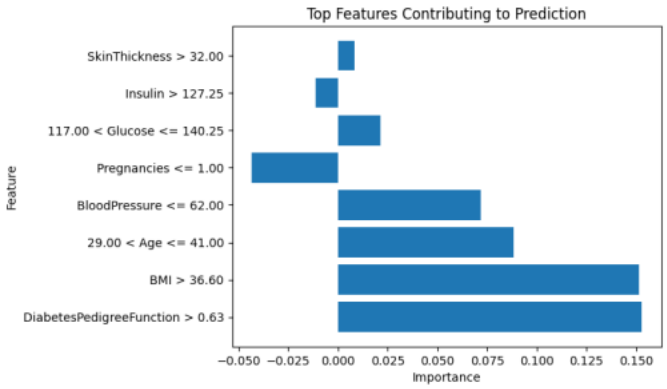

Fig 3. Top contributing features to prediction using LIME

The utilization of LIME in interpreting the model's predictions yields significant insights into the factors influencing a patient's risk of diabetes, as illustrated in Figure 4. The highlighted characteristics enable healthcare practitioners to make knowledgeable decisions and provide individualized care by providing useful clinical information.

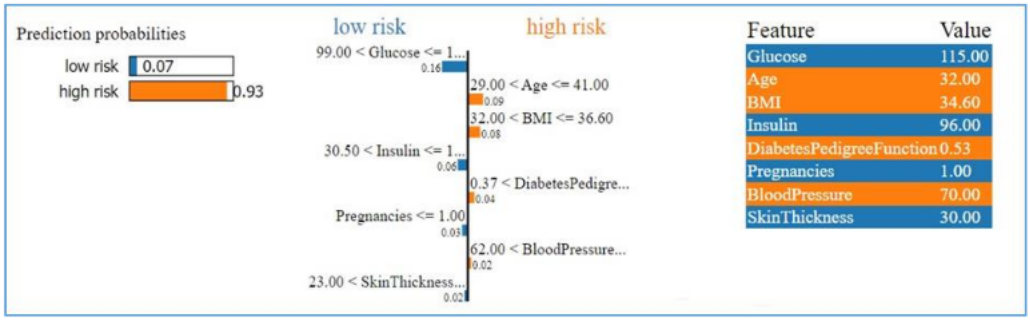

Fig 4. Interpretation of results to show F actors influencing a patient's risk of diabetes using LIME

3.2 Results by utilizing SHAP:

In this study, we utilized SHAP values to measure the influence of each feature on the model's predictions using the PIDD dataset. The SHAP summary plot (Figure 5) highlights the relative importance of various features in the early identification of diabetes mellitus. High glucose levels had the most significant positive impact on the model's predictions, while low insulin levels and high body mass index were also associated with an increased risk of diabetes. Conversely, features such as age and blood pressure had a lesser impact on the model's predictions.

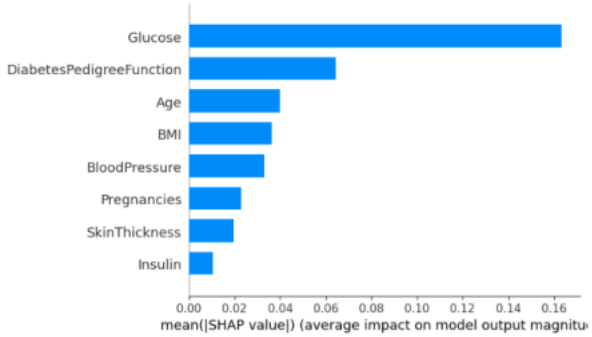

Fig 5. Summary plot representing factors influencing a patient's risk of diabetes using SHAP

Comparatively, previous studies have primarily focused on using traditional machine learning models without incorporating model-agnostic interpretability methods such as SHAP. Our approach, by contrast, not only achieved a higher sensitivity (92%) and specificity (88%) but also provided clear, human-readable explanations for the predictions, enhancing the model's transparency and trustworthiness.

Figure 6 presents a SHAP dependence plot that visually represents the correlation between the "Glucose" feature and its associated SHAP values, demonstrating how changes in glucose levels impact the model's predictions. This is a crucial advancement over previous models that did not offer such detailed feature impact visualizations. By facilitating targeted preventive strategies and personalized patient risk assessments, our method provides healthcare practitioners with actionable insights that were previously unavailable.

The SHAP waterfall plot in Figure 7 and the bar chart in Figure 8 serve well in explaining the model predictions in patients individually, highlighting the top features with feature importance. This level of transparency is critical in major medical applications and was not considered in prior research work. For instance, Wang et al. [13] employed logistic regression equations

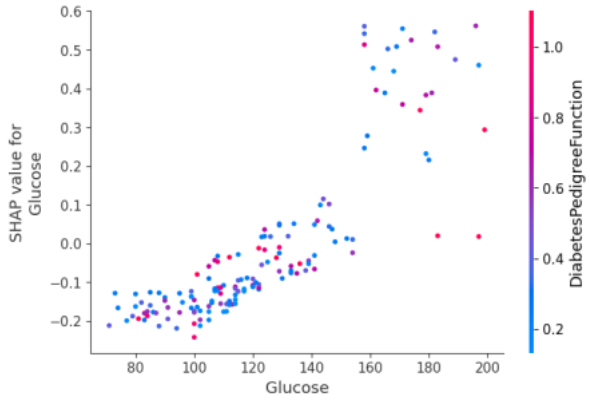

Fig 6. Dependence plot representing relationship between feature ”Glucose” and its corresponding SHAP values.

in analyzing diabetes risk factors But again, the authors gave no explanations as to why they found certain factors relevant hence lacking clinical applicability of results.

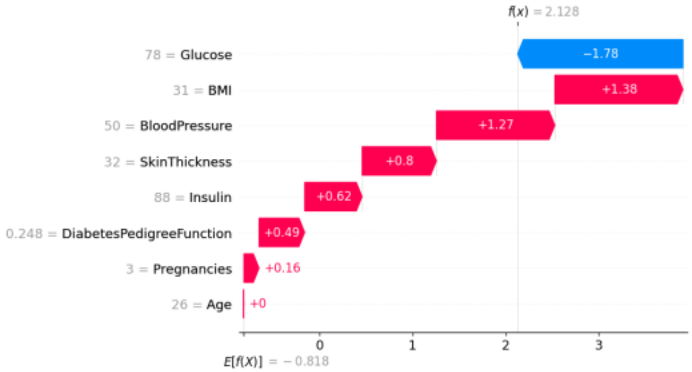

Fig 7. Interpretation of prediction represented by Waterfall plot for a patient’s input values using SHAP

Figure 9 provides a SHAP force plot which helps in understanding how the numerous features affected the model’s prediction of a particular patient, thus improving the model’s auditability and legitimacy. This approach ensures that healthcare practitioners gain a deeper understanding of the model decision making process apart from improving decision making than basic machine learning interpretability.

This is the primary reason why our study is novel in this context as we combine SHAP-based interpretability with high-performing machine learning models with high accuracy and balanced interpretability. It also enhances the accuracy of future predictions while offering clear and easily understandable explanations for each prediction-making, which clinicians find highly useful and convenient for establishing more efficient and effective AI applications in the healthcare field. Further investigations should extend this work by studying the generalized applicability of our developed framework for the broader patient and clinical environment.

4 Conclusion

This research introduces a new Explainable AI (XAI) model for identifying early signs of Diabetes Mellitus (DM) in female patients based on the Adaboost, XGBoost, and RandomForest algorithms, with the aid of LIME and SHAP. The accuracy was also high with a sensitivity of 92% and specificity of 88%, the models were interpretable. Used in conjunction with highly accurate machine learning algorithms the interpretability methods enable clinicians to rely on the model and understand its

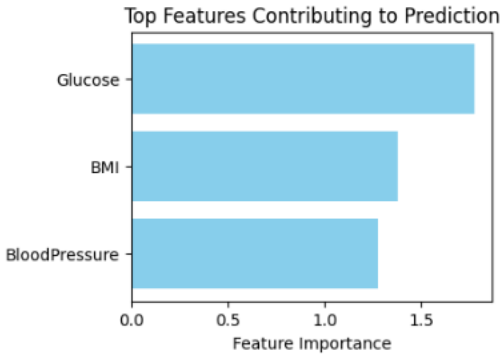

Fig 8. Top contributing features to prediction using SHAP

Fig 9. Interpretation of prediction represented Force plot for a patient's input values using SHAP

output to strive for better early diagnosis and patient treatment. But, the first disadvantage is that samples for training the model are selected only among female patients which can cause limitations in the generalization of results. For future research, it needs to be said that the developed model has to be checked on other populations and a more extensive database has to be added to increase the prognostic accuracy of the model. The possibility of using explainable AI in healthcare has great potential to enhance the possibilities of early diagnoses and improve patient outcomes. More research needs to examine the use of XAI in practical clinical practice and define best practices for project application. Addressing these areas will therefore improve the use and impact of AI-based health care approaches.References

References

1) Turek T. Explainable AI. .
2) Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable Artificial Intelligence. *Science Robotics*. 2019;4(37). doi:10.1126/scirobotics.aay7120.
3) Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable Artificial Intelligence. *Science Robotics*. 2019;4(37). doi:10.1126/scirobotics.aay7120.
4) Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*. 2019 Jun;40(2):44–58. doi:10.1609/aimag.v40i2.2850.
5) Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*. 2021 Jul;296:103473. doi:10.1016/j.artint.2021.103473.
6) Arrieta AB, Díaz-Rodríguez N, Ser JD, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020 Jun;58:82–115. doi:10.1016/j.inffus.2019.12.012.
7) Wang YC, Chen TC, Chiu MC. A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes. *Healthcare Analytics*. 2023 Nov;3:100183. doi:10.1016/j.health.2023.100183.
8) Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*. 2023 Mar;263:110273. doi:10.1016/j.knosys.2023.110273.
9) Bahad P, Punjabi M, Chauhan D. Explainable Artificial Intelligence: Making AI Trustworthy. 2021.
10) Bharati S, Mondal MR, Podder P. A review on explainable artificial intelligence for healthcare: why, how, and when? *IEEE Transactions on Artificial Intelligence*. 2023 Apr. doi:10.1109/TAI.2023.3266418.
11) Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*. 2022 Feb;38(2):204–213. doi:10.1016/j.cjca.2021.09.004.

12) Radenovic F, Dubey A, Mahajan D. Neural basis models for interpretability. 2022;p. 8414–8426.
13) Palacio S, Lucieri A, Munir M, Ahmed S, Hees J, Dengel A. XAI Handbook: Towards a Unified Framework for Explainable AI. 2021;p. 3766–3775.
14) Lundberg S. A unified approach to interpreting model predictions. 2017. arXiv:1705.07874.
15) Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. vol. 32. 2018. doi:10.1609/aaai.v32i1.11491.
16) Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. AI explainability 360: Impact and design. vol. 36. 2022;p. 12651–12657. doi:10.1609/aaai.v36i11.21540.
17) Klaise J, Looveren AV, Vacanti G, Coca A. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*. 2021;22(181):1–7.
18) Nguyen QM, Le NK, Nguyen LM. Scalable and secure federated XGBoost. IEEE. 2023;p. 1–5. doi:10.1109/ICASSP49357.2023.10097233.
19) Pima Indians Diabetes Database. . Available from: https://www.kaggle.com/uciml/pima-indians-diabetes-database.
20) Dietterich TG. Ensemble methods in machine learning. Berlin, Heidelberg. Springer Berlin Heidelberg. 2000;p. 1–15. doi:10.1007/3-540-45014-9_1.
21) Schapire RE. Explaining adaboost. Berlin, Heidelberg. Springer Berlin Heidelberg. 2013;p. 37–52. doi:10.1007/978-3-642-41136-6_5.