ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.1 / No.1 / 2024 / pp.34-44
https://doi.org/10.69882/adba.cem.2024075

Machine Learning Interpretability in Diabetes Risk Assessment: A SHAP Analysis

Mustafa Kutlu *,[1], **Turker Berk Donmez** *,[2] **and Chris Freeman** β,[3]
*Sakarya University of Applied Sciences, Mechatronics Engineering Department, 54050, Sakarya, Turkiye, βElectronics and Computer Science, University of Southampton, UK.

ABSTRACT Diabetes continues to be a complicated and prevalent metabolic illness, providing a serious burden to public health. While machine learning approaches like extreme gradient boosting (XGBoost) provide intriguing options for diabetes prediction, their 'black-box' nature typically limits clinical interpretability. To overcome this gap, our work applied SHapley Additive exPlanations (SHAP) to give insights into the XGBoost model's predictions. The dataset utilized in this research comprised of 253,680 patients and contained 21 parameters, such as General Health Status, High Blood Pressure Status, Age, and Body Mass Index. After feature selection using Recursive Feature Elimination (RFE), 15 important characteristics were discovered. In the test set, the XGBoost model obtained an accuracy of 86.6%, precision of 54.1%, recall of 17.0%, and an F1-score of 25.9% for the Original dataset. For the RFE dataset, the model displayed an accuracy of 86.6%, precision of 54.9%, recall of 16.5%, and an F1-score of 25.3%. SHAP analysis found that General Health Status, High Blood Pressure Status, Age, and Body Mass Index were the most important characteristics in both the Original and RFE datasets. This work provides as a platform for transparent and clinically applicable predictive modeling, assisting in early diabetes identification and preventive healthcare.

KEYWORDS
Explainable AI
Diabetes
Recursive feature elimination

INTRODUCTION

Diabetes, marked by chronic high blood sugar levels, is a pressing global health concern. Its prevalence continues to rise, leading to increased risks of cardiovascular diseases and other severe health issues. The World Health Organization (WHO) states that around 1.28 billion adults aged 30–79 years suffer from diabetes, predominantly in low- and middle-income nations (Diabetes - World Health Organization (Organization *et al.* 2019). Yet, less than half receive proper diagnosis and treatment, with a mere one in five effectively managing their condition (Mihai *et al.* 2022). Such disparities highlight the pressing need for better diabetes prevention, diagnosis, and management.

In the last decade, the allure of machine learning in refining predictive models for diabetes has captivated researchers (Lin *et al.* 2022; Gómez-Peralta and Abreu 2022; Qin *et al.* 2022; Shankaracharya 2017; Afsaneh *et al.* 2022). This has given birth to innovative methods for predicting various outcomes related to diabetes, from individualized risk evaluations to disease behaviour. These methods provide crucial insights into risk factors, pushing the frontier of tailored healthcare. However, many machine learning tools are criticized for their opaque "blackbox" methodologies (Carreras *et al.* 2021). For diabetes prediction, this opacity means healthcare professionals might struggle to grasp the biomarkers and elements that drive predictions. Considering the intricate relationships between diabetes risk factors, this lack of clarity can limit the models' clinical and public health value.

Several biomarkers critical to diabetes' onset and progression are recognized in medical literature (Gómez-Peralta and Abreu 2022). Dyslipidemia, notable for its impact on cholesterol, is a recognized cardiovascular risk, and its ties with diabetes are well-documented (Gutch *et al.* 2017). Other factors, like electrolyte imbalances (high sodium, low potassium), iron metabolism, vitamin B12, and HbA1c levels, are also significantly linked with diabetes (Hasan *et al.* 2021).

Manuscript received: 24 April 2024,
Revised: 10 May 2024,
Accepted: 15 May 2024.

[1] mkutlu@subu.edu.tr (**Corresponding author**)
[2] turkerberkdonmez@yahoo.com
[3] cf@ecs.soton.ac.uk

To enhance model transparency, the research community has pivoted towards explainable artificial intelligence (XAI) methodologies (Hasan *et al.* 2021). Notably, SHapley Additive exPlanations (SHAP) have risen in prominence which demystifies both global and local feature significance, offering a clearer understanding of diabetes risk factors, and aiding more informed medical and public health decisions (Gong *et al.* 2022).

This study delves deep into the existing literature on diabetes prediction, emphasizing transparent machine learning methods. SHAP is employed to elucidate an XGBoost model's predictions, utilizing a commonly available diabetes dataset from The Behavioral Risk Factor Surveillance System. Study involves 253.680 participants with their diabetes status and relevant clinical data. This includes chronic illnesses, medications, and other markers. Firstly, one of the tree-based machine learning models called XGBoost is selected due to its better performance. This is followed by employing SHAP explanations to uncover novel links between diabetes risk and biomedical parameters. This work enriches the understanding of machine learning models in diabetes prediction and sets the stage for transparent, clinically pertinent predictive modeling in diabetes research and care. The paper is structured as follows:

- The Methodology section delves into our research design, discussing data gathering, analytical methods, the XGBoost model, RFE(Recursive Feature Elimination) feature selection method and the use of SHAP for enhancing interpretability.
- The Results section shows our results, concentrating on the performance measures of the XGBoost model in both the Original and RFE datasets. It also digs into the SHAP analysis, emphasizing the most relevant factors like General Health Status, High Blood Pressure Status, Age, and Body Mass Index in both datasets. The section further covers the modest differences in accuracy, precision, recall, and F1-score between the Original and RFE sets, proving the efficiency of our feature selection approaches like RFE in diabetes prediction.
- A comprehensive discussion interprets our results, underlining the significance of transparent machine learning in diabetes research and healthcare choices in Discussion section.
- Finally, Conclusion section highlights the benefits of transparent machine learning in diabetes prediction, addresses constraints, and suggests future research directions to elevate diabetes risk evaluation, ultimately aiming for superior diabetes care and preventive medicine for better public health.

METHODOLOGY

In this work, two variations of the Behavioral Risk Factor Surveillance System (BRFSS) dataset, a yearly telephone survey conducted by the Centers for Disease Control and Prevention (CDC), were subjected to a thorough statistical analysis. The original dataset has two categorical outcome classes: '0,' which denotes the lack of diabetes or the presence of gestational diabetes alone, and '1,' which denotes either prediabetes or diabetes. There were originally 21 feature variables in this dataset, which included a variety of behavioral and health characteristics. Recursive Feature Elimination (RFE) was used as a feature selection strategy to narrow the feature space and improve the model's ability to forecast. A second dataset was created as a result, however only 15 of the most useful features were kept. On both the original and the RFE-selected datasets, subsequent SHapley Additive exPlanations (SHAP) studies were performed to clarify the impact of each variable on the predictive model. This divided method provided a sophisticated assessment of feature relevance and offered insightful information on the characteristics most responsible for the likelihood of developing prediabetes or diabetes. These findings provide healthcare professionals with a strong analytical framework that enables more focused interventions based on the recognized important variables.

70% of the dataset is allocated for training, while the remaining 30% is used for testing, leading to a 70/30 data split. The 'Discussions' section also presents findings from a tenfold cross-validation, providing insights into the model's stability. Model interpretation is further enhanced using SHAP (SHapley Additive exPlanations).

Model Training

In this section, a detailed description of the classifier that has been implemented for the purpose of categorizing the diabetes dataset has been presented. A classifier, a fundamental machine learning algorithm, is utilized to partition the input data into predefined categories. In our specific context, the classifier employs the patient's features as input to discern the presence or absence of diabetes.

High levels of accuracy in predicting diabetes based on a wide range of attributes, including high blood pressure status, high cholesterol levels, cholesterol check status, Body Mass Index, stroke status, heart disease or attack status, physical activity status, heavy alcohol consumption, general health status, mental health status, difficulty walking or climbing stairs, gender, age category, education level, and income level are delineated in Table 1. Models like XGBoost, AdaBoost, Random Forest, and other tree-based standards provided excellent and comparable outcomes in the studies. XGBoost has been chosen as the main model despite alternative models having equivalent performance that has been demonstrated. XGBoost is regarded as the outstanding model due to its usability and practical efficiency, which are especially well-suited to the study objectives. It is the best option for addressing the complexity faced in our work in diabetes forecasting because of its exceptional computational speed mixed with strong predictive skills.

eXtreme Gradient Boosting (XGBoost)

The XGBoost classifier builds upon the Gradient Boosting classifier, which also emphasizes speed and performance. One of the notable features of XGBoost is its regularized learning, which helps to smooth out the final learned weights and prevent overfitting. Overfitting occurs when the model performs well on the training data but poorly on the testing data due to learning both the information and noise from the training data (Fitriyani *et al.* 2020).

The objective function $L(\theta)$ of the XGBoost can be written as:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

where n is the number of training instances, y_i is the actual value for the i^{th} training instance, $\hat{y}_i$ is the predicted value for the i^{th} training instance, $l(y_i, \hat{y}_i)$ is the training loss, which measures the difference between the predicted and actual values for each training instance, K is the number of trees, f_k is the k^{th} tree, and $\Omega(f_k)$ is the regularization term for the k^{th} tree, which penalizes the complexity of the model to avoid overfitting.

The regularization term $\Omega(f_k)$ can be further defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2}$$

Where T is the number of leaf nodes in the k^{th} tree, w is the vector of scores on the leaf nodes in the k^{th} tree, and γ and λ

Table 1 Model Metrics and Confusion Matrixes

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
XGBoost	0.8656	0.5413	0.1701	0.2588	64092 1513 8713 1786
Adaboost	0.8650	0.5402	0.1955	0.2871	63858 1747 8446 2053
Random Forest	0.8599	0.4788	0.4789	0.1748	63608 1997 8663 1836
Decision Tree	0.7979	0.2930	0.3289	0.3099	57274 8331 7045 3454
E. Boosting Machine	0.8654	0.5662	0.1664	0.2573	64266 1339 8751 1748
Naive Bayes	0.7718	0.3166	0.5647	0.4057	52809 12796 4570 5929
Logistic Regression	0.8652	0.5395	0.5395	0.1585	64184 1421 8834 1665
KNN	0.8468	0.3948	0.2066	0.2713	62279 3326 8329 2170
SVM	0.7798	0.2088	0.2184	0.2106	57117 8488 8264 2235

Table 2 Variable names, their indications, and value scales along with patient data.

Variable Name	What It Indicates	Value Scales	Pt6	Pt156	Pt34	Pt265
Diabetes_binary	Diabetes Status	0=no diabetes, 1=prediabetes/diabetes	0.0	1.0	1.0	0.0
HighBP	High blood pressure status	0 = no high BP, 1 = high BP	1.0	1.0	1.0	1.0
HighChol	High cholesterol levels	0 = no high cholesterol, 1 = high cholesterol	0.0	1.0	1.0	0.0
CholCheck	Cholesterol check status	0 = no check in 5 years, 1 = yes	1.0	1.0	1.0	1.0
BMI	Body Mass Index	Continuous Scale	30.0	47.0	24.0	36.0
Smoker	Smoking status	0 = no, 1 = yes	1.0	1.0	1.0	1.0
Stroke	Stroke status	0 = no, 1 = yes	0.0	0.0	0.0	0.0
HeartDiseaseorAtc	Heart disease or attack status	0 = no, 1 = yes	0.0	0.0	0.0	1.0
PhysActivity	Physical activity status	0 = no, 1 = yes	0.0	0.0	0.0	0.0
Fruits	Fruit consumption	0 = no, 1 = yes	0.0	1.0	0.0	1.0
Veggies	Vegetable consumption	0 = no, 1 = yes	0.0	0.0	0.0	1.0
HvyAlcoholCons	Heavy alcohol consumption	0 = no, 1 = yes	0.0	0.0	0.0	0.0
AnyHealthcare	Healthcare coverage	0 = no, 1 = yes	1.0	1.0	1.0	1.0
NoDocbcCost	Avoided doctor due to cost	0 = no, 1 = yes	0.0	0.0	0.0	0.0
GenHlth	General health status	1=excel, 2=very good, 3=good, 4=fair, 5=poor	3.0	3.0	2.0	3.0
MentHlth	Mental health status	1-30 days	0.0	0.0	0.0	0.0
PhysHlth	Physical health status	1-30 days	14.0	0.0	0.0	2.0
DiffWalk	Difficulty in walking	0 = no, 1 = yes	0.0	1.0	0.0	1.0
Sex	Gender	0 = female, 1 = male	0.0	0.0	0.0	0.0
Age	Age category	1 = 18-24, 9 = 60-64, 13 = 80 or older	9.0	11.0	12.0	11.0
Education	Education level	1 = Never attended, 6 = College graduate	6.0	6.0	3.0	5.0
Income	Income level	1 = less than $10,000, 8 = $75,000 or more	7.0	5.0	3.0	4.0

are regularization parameters that control the complexity of the model. Large weights are penalized by the regularization term, which also encourages the model to have more streamlined and comprehensible structural elements. The goal of the model is to minimize this loss over the entire training set. Given that our dataset is also noisy, XGBoost is one of the appropriate classifiers for it Shao and Hu (2022). The values of the hyperparameters are given in Table 3.

Recursive Feature Elimination (RFE)

Finding the ideal subset of features for a particular machine learning study is the goal of the feature selection method known as recursive feature elimination (RFE)Zhang *et al.* (2022). When using RFE, a model is fitted to the data, and the features are then ranked according to how important or relevant they are to the prediction. The least significant characteristics are then gradually removed by RFE until the target number of features is obtained or a stopping requirement is satisfied. By lowering the dimensionality and complexity of the data, as well as by removing noise and multicollinearity among features, RFE can enhance the effectiveness and performance of machine learning models.Chen and Jeong (2007)

Recursive Feature Elimination (RFE) is a feature selection algorithm commonly employed in machine learning for identifying a subset of most predictive features. Given a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with n samples, where each $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector and y_i is the corresponding label, RFE aims to minimize a loss function $\mathcal{L}(f(x;\theta), y)$ by selecting the most relevant features. The optimization objective is $\theta^* = \arg\min_\theta \sum_{i=1}^{n} \mathcal{L}(f(x_i;\theta), y_i)$. The algorithm starts by fitting a predictive model $f : \mathbb{R}^d \to \mathbb{R}$ parameterized by θ to the entire feature set. It then ranks the features based on their importance, often calculated as $\text{Importance}(j) = \left|\frac{\partial \mathcal{L}}{\partial x_j}\right|$ for each feature j, and recursively eliminates the least important ones. This process continues iteratively until the desired number of features is retained. The computational complexity of RFE is generally $O(d \times p)$, where p is the complexity of the base estimator used for ranking features. RFE is particularly useful when model interpretability and simplicity are as crucial as predictive performance, although it can be computationally expensive for high-dimensional data.

Several machine learning models that offer some kind of feature importance metric can be used with recursive feature elimination (RFE). This comprises linear models with coefficients, tree-based techniques with feature significance scores, and support vector machines (SVMs) with their support vectors. It's important to keep in mind, though, that RFE functions as a wrapper technique and requires the fitting of a new model for each subset of characteristics. This might require a lot of work and increase the risk of model overfitting. Table 2 provides more information on the 15 relevant parameters that were chosen for our investigation using RFE.

Model Interpretation with SHAP

To augment the transparency of our XGBoost model and provide clinically meaningful interpretations, we employed SHapley Additive exPlanations (SHAP) as an interpretability tool. SHAP is rooted in cooperative game theory, aiming to fairly allocate "payouts" or contributions among features for a given prediction. Mathematically, the Shapley value ϕ_i of a feature i is computed as:

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

where $f(S)$ is the prediction of the model for feature subset S, and N is the set of all features. These Shapley values represent the marginal contribution of each feature to a specific prediction, enabling us to decipher both global and local feature importance.

The application of SHAP to our XGBoost model adhered to a systematic methodology. Initially, specific data preprocessing steps tailored for SHAP analysis were executed. Utilizing the SHAP library, Shapley values for each feature across all data points were computed. These values were aggregated to form global feature importance metrics, which can be mathematically represented as:

$$\text{Global Importance}(j) = \frac{1}{n} \sum_{i=1}^{n} |\phi_{ij}|$$

where ϕ_{ij} is the Shapley value of feature j for data point i, and n is the total number of data points. This provided an overarching view of the most and least impactful features in diabetes prediction. Interestingly, our analysis revealed that certain biomarkers, traditionally not emphasized in diabetes literature, played a significant role in model predictions.

SHAP also excels in local interpretability, enabling us to dissect individual predictions. For example, in case studies of specific patients, the SHAP values highlighted the biomarkers that were instrumental in classifying them as high-risk or low-risk for diabetes. This local interpretability is invaluable for clinicians, as it provides personalized insights into each patient's risk factors. Mathematically, local interpretability can be examined through the individual Shapley values ϕ_{ij}, providing a nuanced understanding of how each feature contributes to a specific prediction.

In XGBoost, SHAP values may be particularly useful for comprehending how various characteristics interact with one another. The authors' modification of SHAP values to take interaction effects into consideration led to the idea of SHAP interaction values. These interaction values guarantee the explanations of each of XGBoost's unique estimations for interaction effects. In order to highlight important interactions that XGBoost captures but that could otherwise go undetected, SHAP interaction values discriminate between main effects and interaction effects for specific model predictions.He *et al.* (2023)

An efficient tool for understanding XGBoost models is the SHAP values. They constitute a significant theoretical breakthrough by overcoming consistency problems and permitting the separate assessment of main and interaction effects for individual model predictions.This might highlight important interactions that XGBoost has recorded but that were otherwise missed. Data scientists are now better prepared to develop models that are more clear and intelligible thanks to the introduction of SHAP values in XGBoost.Gao *et al.* (2023)

RESULTS

Results of Classifications

First, two separate sets were created from the original dataset after a data preparation step. The original collection, referred known as the "Original Set," had 253,680 samples and kept all 21 features. The second set, referred to as the "RFE Set," was created by choosing 15 important features from the initial 253,680 samples using recursive feature elimination (RFE).

RFE was used to improve feature selection and reduce model complexity. This approach chose 15 characteristics, which are listed in Table reftab:attributes, including high blood pressure status, high cholesterol levels, cholesterol check status, body mass index, and others.

Table 3 Hyperparameter values for the model.

Hyperparameter	Value
eta (learning rate)	0.3
n_estimator (number of gradient-boosted trees)	100
Gamma (min split loss)	0
Max depth	6
Min child weight	1
Max delta step	0
subsample	1
Sampling method	Uniform

The two sets—'Original' and 'RFE'—were each put through an XGBoost classifier separately to evaluate the precision and robustness of our feature selections, as shown also in Table reftab:analysis-results. Despite some variations in the measures, the classifier performed well on both sets of data.

The classifier's performance for the 'Original' set was accuracy of 86.6%, precision of 54.1%, recall of 17.0%, and F1-score of 25.9%.

Similar results were obtained for the 'RFE' set, where the classifier's accuracy was 86.6%, precision was 54.9%, recall was 16.5%, and F1-score was 25.3%.

These findings validate the efficacy of our feature selection approaches by showing that both feature sets are useful for classification, although with minor changes in accuracy and recall measures.

Explaining Model with SHAP

Machine learning algorithms have emerged as effective tools for disease prediction in the evolving healthcare landscape. However, the 'black-box' nature of these models often makes it challenging to interpret their predictions and understand the relative importance of different input factors. To address this issue, interpretability tools as SHAP have gained popularity. SHAP values provide a nuanced understanding of the factors influencing illness risk by quantifying the impact of specific attributes on the model's prediction. In this context, we explore the intricate network of health factors affecting the likelihood of developing hypertension, a common and complex cardiovascular disease. Our goal is to analyze the interaction of various health metrics, ranging from medication use and standard risk factors to less obvious influencers, and how they collectively influence the model's predictions using patient-specific SHAP values. This investigation highlights how machine learning interpretability tools can enhance our understanding of disease prediction and potentially guide therapeutic decisions.

study highlights the primary factors impacting hypertension across various dataset modifications through the analysis of global SHAP values. Several common trends and intriguing variances are observed. The average magnitude of the SHAP values for each feature over the entire dataset is evaluated, as depicted in Figure 1a.

To acquire a better understanding of how each feature effects the model's predictions, SHAP (Shapley Additive Explanations) values were generated. Among all the characteristics, General Health Status, High Blood Pressure Status, Age, and Body Mass Index emerged as the most significant in both the Original and RFE datasets. In the Original dataset, the SHAP values were notably 0.674 for General Health Status, 0.523 for High Blood Pressure Status, 0.413 for Age, and 0.401 for Body Mass Index. Similarly, in the RFE dataset, the comparable SHAP values were 0.673 for General Health Status, 0.530 for High Blood Pressure Status, 0.411 for Age, and 0.404 for Body Mass Index as also shown in figure 1. These values provide as a quantifiable indication of the average influence each of these qualities has on the model's output, therefore underlining their relevance in diabetes prediction.

These results validate the success of our feature selection techniques by showing the resilience of the chosen features in predicting diabetes, regardless of the feature set employed.

A more sophisticated knowledge of the model's prediction ability is revealed when the individual SHAP values are thoroughly analyzed and compared to the actual health situations of particular patients. Negative SHAP values for General Health Status were seen starting with Patient 6, who was correctly identified as not having diabetes (True Negative),as shown in the Figure 2, -1.3621 in the "Original" set and -1.4022 in the "RFE" set. These data support the model's accurate prediction of a decreased risk of diabetes, together with a negative SHAP value for high blood pressure status (-0.5726 in "Original" and -0.5969 in "RFE").

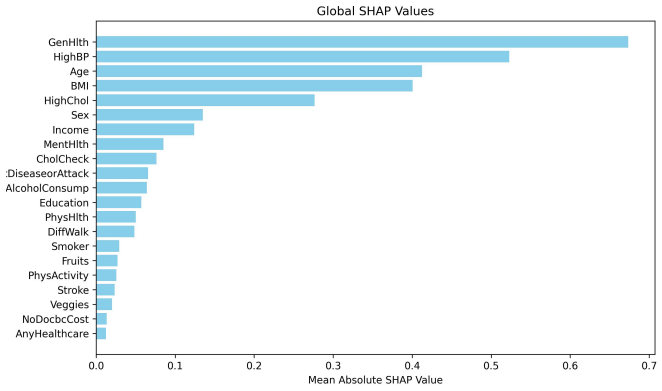
Patient 156 had substantial positive SHAP values for BMI, 0.9521 in the "Original" set and 0.8867 in the "RFE" set, and was appropriately diagnosed as having diabetes (True Positive).As shown in the figure 2, positive SHAP values for High Blood Pressure Status (0.5819 in 'Original' and 0.6005 in 'RFE') are confirmed to have a significant influence in the model's ability to diagnose diabetes accurately.

However, the model's forecasts did not show infallibility. Positive SHAP readings for High Blood Pressure Status were found in both sets for Patient 34, who was incorrectly diagnosed as not having diabetes (False Negative), as shown in the Figure 3, namely 0.5234 in the "Original" and 0.5418 in the "RFE". These were countered by a negative SHAP value for General Health Status (-0.5440 in 'Original' and -0.4934 in 'RFE'), indicating that these competing signs may have caused the inaccurate prediction to be confused.

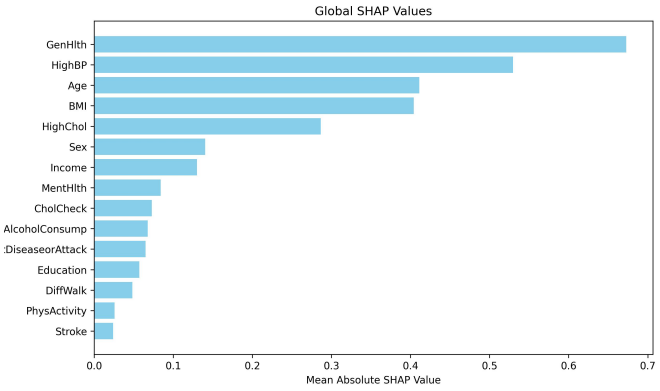
Similar positive SHAP readings as shown in the figure 3, for

Table 4 Attribution of the dataset and patient details with RFE Selection

No.	Attribution	Variable	RFE	Pt6	Pt156	Pt34	Pt265
1	High blood pressure	HighBP	Picked	1.0	1.0	1.0	1.0
2	High cholesterol levels	HighChol	Picked	0.0	1.0	1.0	0.0
3	Cholesterol check status	CholCheck	Picked	1.0	1.0	1.0	1.0
4	Body Mass Index	BMI	Picked	30.0	47.0	24.0	36.0
5	Smoking status	Smoker		1.0	1.0	1.0	1.0
6	Stroke status	Stroke	Picked	0.0	0.0	0.0	0.0
7	Heart disease or attack	HeartDorA	Picked	0.0	0.0	0.0	1.0
8	Physical activity status	PhysActivity	Picked	0.0	0.0	0.0	0.0
9	Fruit consumption	Fruits		0.0	1.0	0.0	1.0
10	Vegetable consumption	Veggies		0.0	0.0	0.0	1.0
11	Heavy alcohol consmp.	HvyAlcoholC	Picked	0.0	0.0	0.0	0.0
12	Healthcare coverage	AnyHealthcr		1.0	1.0	1.0	1.0
13	Avoided doctor due to cost	NoDocbcCos		0.0	0.0	0.0	0.0
14	General health status	GenHlth	Picked	3.0	3.0	2.0	3.0
15	Mental health status	MentHlth	Picked	0.0	0.0	0.0	0.0
16	Physical health status	PhysHlth		14.0	0.0	0.0	2.0
17	Difficulty in walking or climbing stairs	DiffWalk	Picked	0.0	1.0	0.0	1.0
18	Gender	Sex	Picked	0.0	0.0	0.0	0.0
19	Age category	Age	Picked	9.0	11.0	12.0	11.0
20	Education level	Education	Picked	6.0	6.0	3.0	5.0
21	Income level	Income	Picked	7.0	5.0	3.0	4.0

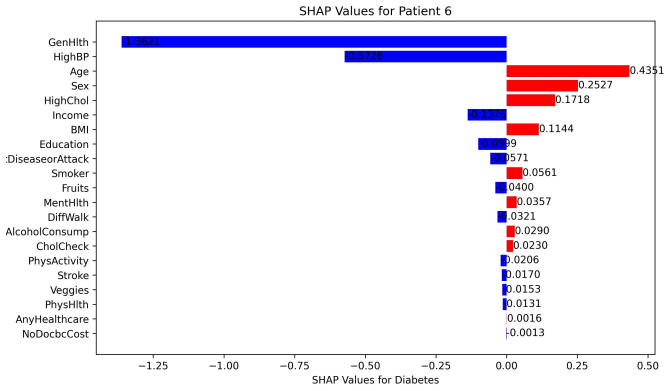


(a) Global SHAP Values

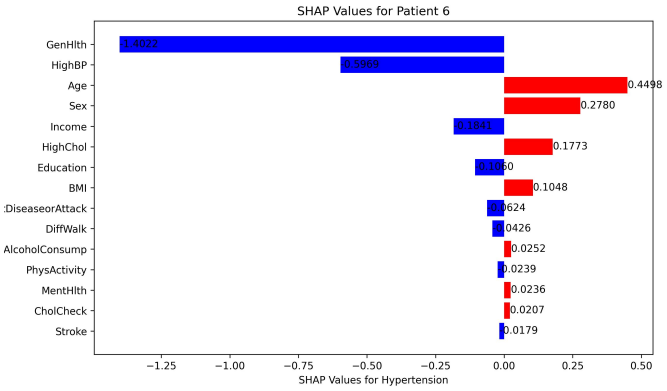


(b) Global SHAP Values (RFE)

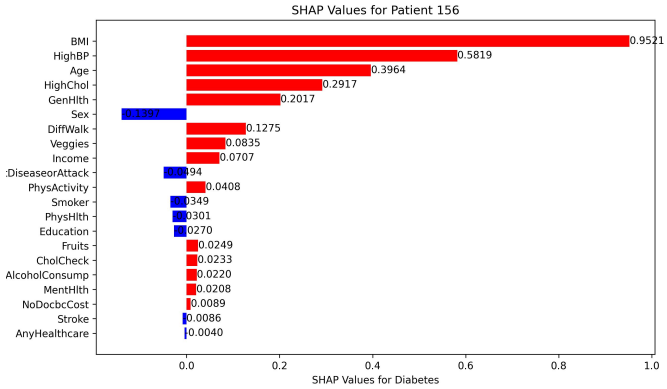
Figure 1 Comparison of Global SHAP Values



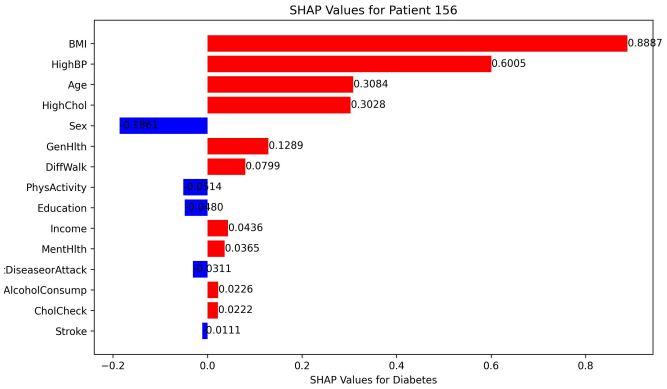
(a) Local SHAP values for original dataset



(b) Local SHAP values for RFE dataset



(c) Local SHAP values for original dataset



(d) Local SHAP values for RFE dataset

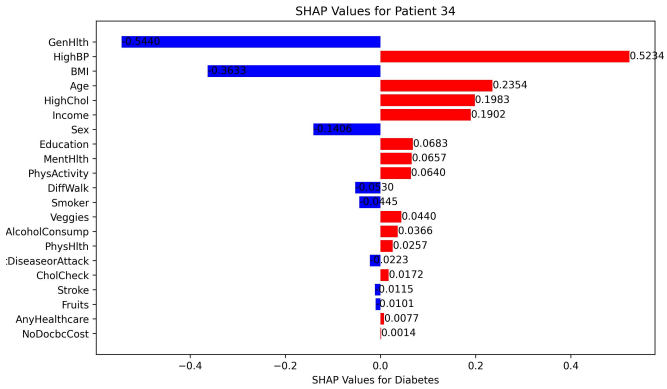
Figure 2 Comparison of local SHAP values for Patient 6 and 156

BMI (0.5051 in 'Original' and 0.5029 in 'RFE') and High Blood Pressure Status (0.5122 in 'Original' and 0.4965 in 'RFE') were reported in the case of Patient 265 who was misdiagnosed as having diabetes (False Positive). Despite these signs, the model overestimated the danger, which led to an inaccurate prognosis.

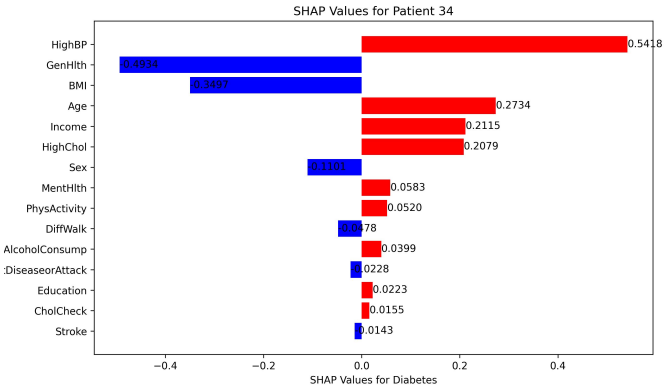
These distinct studies confirm the robustness of the feature selection while also highlighting potential areas where more model fine-tuning may be necessary for more precise and dependable predictions.

SHAP Dependences for Variables

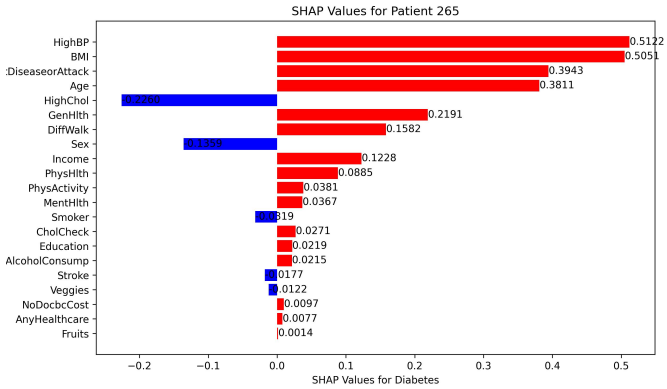
In the study, the clarification of how individual features influence the model's predictions was provided by SHAP dependence plots. Additionally, how these relationships are modulated by other in-



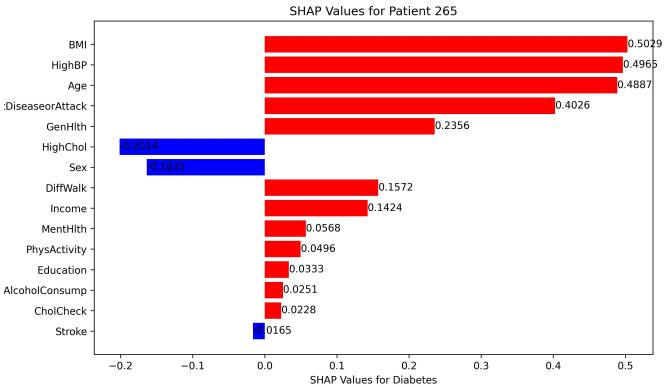
(a) Local SHAP values for original dataset



(b) Local SHAP values for RFE dataset



(c) Local SHAP values for original dataset



(d) Local SHAP values for RFE dataset

Figure 3 Comparison of local SHAP values for Patient 34 and 265

teracting variables was indicated by the color of each point on the plots. A comprehensive understanding of the key variables and their interactions that drive the model's decisions was offered by these plots.

Two distinct sets were utilized for this analysis: the 'Original' set, which encompasses all 21 variables, and the 'RFE' set, a refined subset containing 15 selected features. In both sets, the identification of the most impactful features was made: General Health Status (GenHlth), High Blood Pressure Status (HighBP), Age, and Body Mass Index (BMI), listed in descending order of their importance. A more comprehensive understanding of the implications of these critical variables was enabled by the utilization of both sets, thereby enriching the interpretability of the model.

It must be noted that rigorous statistical analysis has not been applied to the SHAP values, which are intended to provide an initial understanding of the relationships between the variables and diabetes risk. For example, a greater likelihood of diabetes was suggested by higher SHAP values for General Health Status and High Blood Pressure Status in both sets, as illustrated in the dependence plots.

Regarding Age, an increase in the risk of diabetes as age increases was indicated by the SHAP values. This is consistent with broader medical understanding, in which older age is linked to a higher likelihood of developing chronic conditions, including diabetes.

Similarly, a link between elevated Body Mass Index (BMI) values and a higher risk of diabetes was established, corroborating existing medical literature that associates obesity with diabetes.

In summary, the validation of the efficacy of the feature selection process was achieved through the SHAP dependence plots generated from both the 'Original' and 'RFE' sets. Nuanced insights into how these selected variables interact to influence the model's predictive capabilities were also offered.

DISCUSSION

In the current work, the efficacy of machine learning methods in predicting diabetes risk was examined, with a special emphasis on the XGBoost classifier. Two unique datasets were employed: the 'Original' set, encompassing all 21 variables, and the 'RFE' set, a refined subset of 15 chosen characteristics. The SHAP dependent charts were developed to give subtle insights into the correlations between various variables and the model's prediction capabilities. It was discovered that General Health Status (GenHlth), High Blood Pressure Status (HighBP), Age, and Body Mass Index (BMI) were the most impacting factors in both groups. These results accord with current research, strengthening the assumption that these characteristics are key markers of diabetes risk.

The investigation also demonstrated that the XGBoost classifier performed with a high degree of accuracy on both the 'Original' and 'RFE' sets. However, it should be noted that the model's performance indicators, such as accuracy and recall, differed marginally across the two sets. This implies that although feature selection

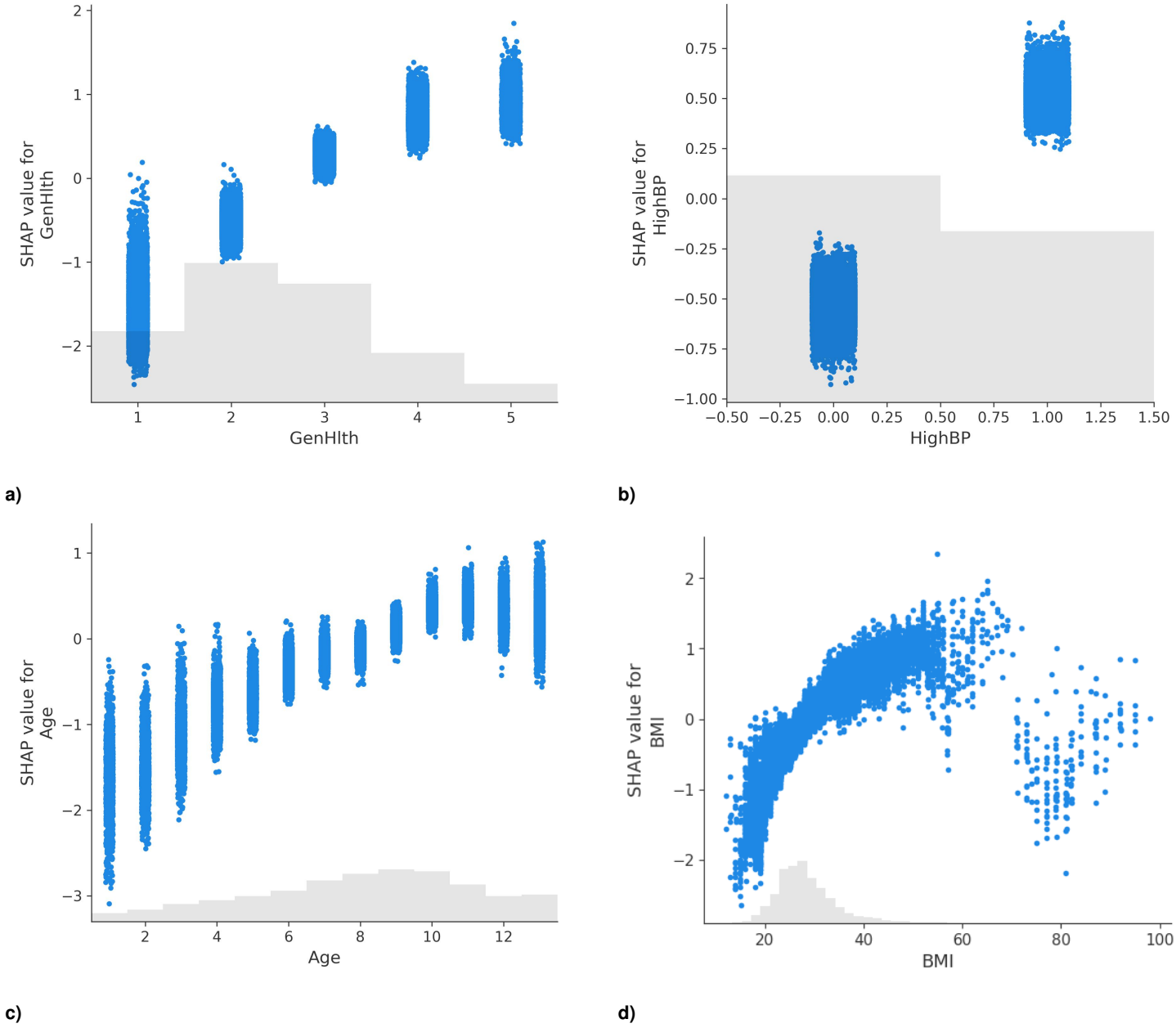


Figure 4 Dependence plots in original dataset for variables **a)** General health status, **b)** High blood pressure, **c)** Age, **d)** Body Mass Index (BMI)

strategies like RFE may be efficient in decreasing model complexity, they may not necessarily lead to a substantial gain in predictive accuracy. Therefore, the option between employing all accessible data or a subset of chosen features should be made carefully, considering the unique aims and restrictions of the study.

One of the disadvantages of this research is the absence of comprehensive statistical analysis done to the SHAP values. While the SHAP dependency plots give helpful early insights into the correlations between factors and diabetes risk, they are not a replacement for a complete statistical study. Future study might benefit from a more in-depth investigation, potentially adding additional statistical approaches to confirm the results further.

In conclusion, this work adds to the expanding corpus of research on the application of machine learning algorithms in healthcare. It reveals that XGBoost, when trained on well chosen characteristics, may be a strong tool for predicting diabetes risk. The insights gathered from the SHAP dependent plots give an extra layer of interpretability to the model, making it more transparent and perhaps more trustworthy in a clinical environment. Further study is required to confirm these results across diverse demographics and healthcare systems.

CONCLUSION

In our comprehensive research journey into the domain of machine learning interpretability, an emphasis on diabetes risk assessment was pronounced, leveraging the usability of SHAPley analysis. Drawing from the Behavioral Risk Factor Surveillance System (BRFSS) dataset, the study precisely analyzed two distinct variations (original and feature selected) of the dataset, which is a

renowned annual telephone survey collected by the Centers for Disease Control and Prevention (CDC). The dataset presented a separation in its outcome classes, defining the presence of diabetes, offering a granular understanding of this pervasive disease. Firstly, encompassing 21 varied feature variables, the dataset provided a kaleidoscopic view of a many of behavioural and health characteristics. Identifying the potential for refinement, we harnessed RFE as a strategic tool for feature selection. This not only rationalised our feature space but significantly augmented our model's prediction ability. In conclusion our efforts underscore the enormous benefits of adopting a transparent machine learning paradigm, especially in the diabetes prediction.

The horizon of future endeavours, our study gazed into several areas of research. The profound depth of the BRFSS dataset hints at the potential for uncovering more intricate patterns and correlations, thereby exploring diabetes risk prediction. While our current methodologies have proven the usability, the dynamic nature of healthcare mandates continuous evolution. This could clear in the form of integrating more sophisticated machine learning algorithms, or perhaps investigating the distinctions of SHAP to loosen more complex model behaviors. Collaborative research performed with medical professionals could surface the real-world results for the application of our models, offering concrete benefits to patients. It is aspired by amalgamating feedback from the clinics, refining methodologies, and developing innovation, to redefine the landscape of machine learning in diabetes risk assessment.

Availability of data and material

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Afsaneh, E., A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, 2022 Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetology & Metabolic Syndrome **14**: 1–39.

Carreras, J., S. Hiraiwa, Y. Y. Kikuti, M. Miyaoka, S. Tomita, *et al.*, 2021 Artificial neural networks predicted the overall survival and molecular subtypes of diffuse large b-cell lymphoma using a pancancer immune-oncology panel. Cancers **13**: 6384.

Chen, X.-w. and J. C. Jeong, 2007 Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pp. 429–435, IEEE.

Fitriyani, N. L., M. Syafrudin, G. Alfian, and J. Rhee, 2020 Hdpm: an effective heart disease prediction model for a clinical decision support system. IEEE Access **8**: 133034–133050.

Gao, X. R., M. Chiariglione, K. Qin, K. Nuytemans, D. W. Scharre, *et al.*, 2023 Explainable machine learning aggregates polygenic risk scores and electronic health records for alzheimer's disease prediction. Scientific Reports **13**: 450.

Gómez-Peralta, F. and C. Abreu, 2022 Clinical research on type 2 diabetes: A promising and multifaceted landscape.

Gong, H., M. Wang, H. Zhang, M. F. Elahe, and M. Jin, 2022 An explainable ai approach for the rapid diagnosis of covid-19 using ensemble learning algorithms. Frontiers in Public Health **10**: 874455.

Gutch, M., S. Rungta, S. Kumar, A. Agarwal, A. Bhattacharya, *et al.*, 2017 Thyroid functions and serum lipid profile in metabolic syndrome. Biomedical journal **40**: 147–153.

Hasan, M. A., M. F. K. Chowdhury, S. Yasmin, S. Paul, T. Ahmed, *et al.*, 2021 Association between glycemic control and serum lipid profile in type 2 diabetic patients: Experience in a medical college hospital. Bangabandhu Sheikh Mujib Medical University Journal **14**: 138–143.

He, Z., Y. Yang, R. Fang, S. Zhou, W. Zhao, *et al.*, 2023 Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems. Frontiers in Environmental Science **10**: 1057081.

Lin, Y., Y. Li, X. Huang, L. Liu, H. Wei, *et al.*, 2022 Analysis of diabetes clinical data based on recurrent neural networks. Computational Intelligence and Neuroscience **2022**.

Mihai, D. A., D. S. Stefan, D. Stegaru, G. E. Bernea, I. A. Vacaroiu, *et al.*, 2022 Continuous glucose monitoring devices: A brief presentation. Experimental and Therapeutic Medicine **23**: 1–6.

Organization, W. H. *et al.*, 2019 Classification of diabetes mellitus .

Qin, Y., J. Wu, W. Xiao, K. Wang, A. Huang, *et al.*, 2022 Machine learning models for data-driven prediction of diabetes by lifestyle type. International Journal of Environmental Research and Public Health **19**: 15027.

Shankaracharya, S., 2017 Diabetes risk prediction using machine learning: Prospect and challenges. Bioinformatics, Proteomics and Immaging Analysis **3**: 194– 195.

Shao, N. and H. Hu, 2022 Exploring the path of enhancing ideological and political education in universities in the era of big data. Journal of Environmental and Public Health **2022**.

Zhang, Y., X. Zhang, J. Razbek, D. Li, W. Xia, *et al.*, 2022 Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome. BMC Endocrine Disorders **22**: 1–15.

How to cite this article: Kutlu, M., Donmez, T. B., and Freeman, C. Paper Name. *Machine Learning Interpretability in Diabetes Risk Assessment: A SHAP Analysis*, 1(1), 34-44, 2024.