Journal of Computer Science

Research Article

A Metaheuristic-Optimized Feature Selection for Early-Stage Diabetes Prediction With SHAP-Guided Insight into Influential Attributes

Esmay Azam Rochy[1], Jannatul Ferdaus[1], Uzzal Biswas[1], Jun-Jiat Tiang[2] and Abdullah-Al Nahid[1]

[1]Electronics and Communication Engineering Department, Khulna University, Khulna, Bangladesh
[2]Centre for Wireless Technology, CoE for Intelligent Network, Faculty of Artificial Intelligence & Engineering, Multimedia University, Persiaran Multimedia, Cyberjaya, Selangor, Malaysia

Article history
Received: 16-05-2025
Revised: 04-09-2025
Accepted: 23-09-2025

Corresponding Authors:
Uzzal Biswas
Email: uzzal.biswas@ece.ku.ac.bd;
Jun-Jiat Tiang
Email: jjtiang@mmu.edu.my;
Abdullah-Al Nahid
Email: nahid@ece.ku.ac.bd;
nahid.ece.ku@gmail,com;

Abstract: Diabetes is a metabolic disorder that causes elevated blood glucose. This long-term health condition can lead to cardiovascular diseases, stroke, kidney failure, visual impairment, neuropathy, and even death in critical cases. So, a Computer-Aided Diagnostic (CAD) system is necessary to diagnose diabetes automatically. A clinician can utilize a machine learning-based CAD system that automatically diagnoses many people. This paper will use a Random Forest (RF) classifier for Machine Learning (ML) classification to identify if any individual is diabetic or non-diabetic. In order to increase the accuracy and robustness of the model, the Zebra Optimization Algorithm (ZOA) and the proposed Nomad Zebra Optimization Algorithm (NZOA) are used to identify the most optimal feature sets based on RF subset selection and RFE (Recursive Feature Elimination) technique. Smoking and Age have been identified as the most influential features with a prediction accuracy of 79.86% with a precision of 75.51%, recall of 88.33%, and F1-score of 81.42% using the proposed NZOA. Finally, to further increase the model interpretability and assist physicians in making decisions without any irrationality, SHAP (Shapley Additive Explanations) is used to explain the outputs of the models based on game theory and optimal credit allocation techniques. It also identifies that smoking has the highest impact on our model.

Keywords: Machine Learning, Diabetes Prediction, Diagnosis, Features, Explainable AI, Optimization, Public Health

Introduction

Diabetes mellitus is a common metabolic disease characterized by dysfunctional insulin activity. It leads to elevated amounts of glucose in the bloodstream and affects a large number of people across the globe. The chronic production of diabetes poses significant dangers to cardiovascular health, kidney function, and neurological well-being (Alamro et al., 2023). The importance of taking proactive measures is evident through various key risk factors, including genetic susceptibility, obesity, and sedentary lifestyles (Sathi et al., 2022). This disorder manifests in Type 1, which involves autoimmune insulin insufficiency, and Type 2, characterized by insulin resistance and reduced production (Sathi et al., 2022; Uddin et al., 2023). Lifestyle interventions such as getting regular exercise, maintaining a healthy diet, and smoking have a significant effect on conditions like Type 2 diabetes and its associated problems (Tomic et al., 2022). For this reason, diabetes has been identified as a non-communicable disease like cancer and cardiovascular diseases because of its high mortality (Budreviciute et al., 2020).

The traditional method for diagnosing diabetes involves blood tests to measure fasting glucose levels, oral glucose tolerance, and glycated hemoglobin (HbA1c). These methods are efficient but have limitations, including intermittent spike oversight and potential inaccuracies in representing short-term glucose fluctuation (Mauricio et al., 2020). Machine Learning uses algorithms such as logistic regression, decision trees, and neural networks to analyze data and predict outcomes beyond conventional methods, ML includes feature selection and ensemble learning, which enhances accuracy (Khalid et al., 2014). Now a days It's become crucial to design an automatic disease classification system to increase efficiency and make the testing process more affordable.

In recent years, a handful of research has been conducted using ML to predict diabetes in early stages. Kakoly et al. (2023) employed five machine learning algorithms, obtaining an accuracy of 82.2% and an AUC of 87.2%. This study highlighted the clinical aspect of predicting diabetes. In another study, Cheng et al. (2023) showed that the random forest classifier performed better than other classifiers in predicting diabetes. It has achieved an accuracy of 84%, an AUC of 95%, a sensitivity of 77%, and a specificity of 91% (Cheng et al., 2023). Syed and Khan (2020) highlighted the efficiency of the Decision Forest algorithm by demonstrating its outstanding accuracy, precision, recall, AUC, and F1 score on various datasets. Shrestha et al. (2023) introduced a hybrid model that combines Support Vector Machines (SVM), Radial Basis Function (RBF), and Long Short-Term Memory (LSTM); utilizing this, they have achieved an accuracy of 86.31% and AUC of 82.70% as well as processing time. Sonia et al. (2023) demonstrated a 97% accuracy in diabetes detection using a multi-layer neural network. The study also reported a specificity of 0.95 and a sensitivity of 0.97. Alqushaibi et al. (2023) proposed a Bayesian-based Convolution Neural Network (CNN) with a robust accuracy of 89.36%, using the Synthetic Minority Oversampling Technique (SMOTE) to address the imbalanced classes. All the work mentioned above significantly impacts the scientific community because of its applications to a particular dataset and classification algorithm. However, none have addressed the issue of decreasing the number of features and their importance to predicting diabetes.

Feature Selection is a technique used in ML to remove redundant and irrelevant features from the dataset to more accurately predict the desired outcomes, which is sometimes not possible if the dataset has a noisy feature with many outliers in those feature (Chandrashekar and Sahin, 2014; Miao and Niu, 2016). Metaheuristic algorithms are often used to address optimization problems to increase or decrease fitness function value, and they are also used in feature selection problems to select a subset of features for the highest accuracy (Agrawal et al., 2021; Dash and Liu, 1997). Popular algorithms include the Genetic Algorithm (GA), Particle Swarm Optimization algorithm (PSO), Ant Colony Optimization Algorithm (ACO), Whale Optimization Algorithm (WOA) and Zebra Optimization Algorithm (ZOA)(Agrawal et al., 2021; Trojovska et al., 2022; Mirjalili and Lewis, 2016; Dorigo et al., 2006; Holland, 1984; Wong and Ming, 2019). Sakri et al. (2018) have implemented PSO-based feature selection for breast cancer recurrence. Samee et al. (2022) used PSO in chest X-ray images in order to select important features. In the chronic kidney disease classification dataset Raihan et al. (2023) have implemented biogeography-based optimization (BBO) feature selection to get a hand on the optimum feature set.

Building on the insights from the aforementioned studies, we utilize both the standard ZOA and our proposed variant, the Nomad Zebra Optimization Algorithm (NZOA), as metaheuristic optimizers to select features for the RF classifier. The novel contributions of this work are threefold: firstly, we propose the Nomad Zebra Optimization Algorithm (NZOA) to select an optimal subset of features for diabetes prediction; secondly, our approach significantly reduces the model training time compared to the existing ZOA algorithm; and thirdly, we integrate SHAP to develop an interpretable machine learning model, providing transparent insights into the influence of individual features on prediction outcomes. We have used the RFE feature selection technique and the findings to create different feature sets to train our model.

Methodology

In this section, we discuss our ML-based diabetes prediction system, which consists of six different stages.

1. Dataset collection and pre-processing
2. RF classifier
3. ZOA base feature selection on RF classifier
4. NZOA base feature selection on RF classifier
5. Performance evaluation and comparison
6. SHAP base model interpretability

The dataset is collected from online, preprocessed, and encoded in ready-to-use standard. After that, we train them using ML base classifier which is RF, as the target feature is labeled as 0 and 1. Then NZOA and ZOA were used on the same dataset to find optimal feature sets that influence the output targets most. Finally, each optimization model output is fed through SHAP interpretable system to get an inside view of the black box. So that researchers and physicians can identify which features or attributes have a lager contribution to becoming diabetic or non-diabetic patients. Our work process, depicted in Figure 1, provides a clear overview of the sequential steps of the methodology, from feature selection to classification and subsequent analysis using SHAP-based interpretability.

Dataset and Preprocessing

The dataset we used in this study was composed of a cross-sectional survey by Syed and Khan (2020) with 4896 individuals, of whom 990 had diabetes and 3906 did not. Demographic data like area, gender, and age were included when considering ten risk factors. The description of the feature sets and their respective values is shown in Table 1. The data collector performed preprocessing through transformations and categorizations of the explanatory variables. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) has applied to guarantee the target class balance. It used minority class to create new synthetic observations from the present values (Fernandez et al., 2018). After SMOTE analysis, the dataset now contains 3906 non-diabetes patients and 3900 diabetes patients.

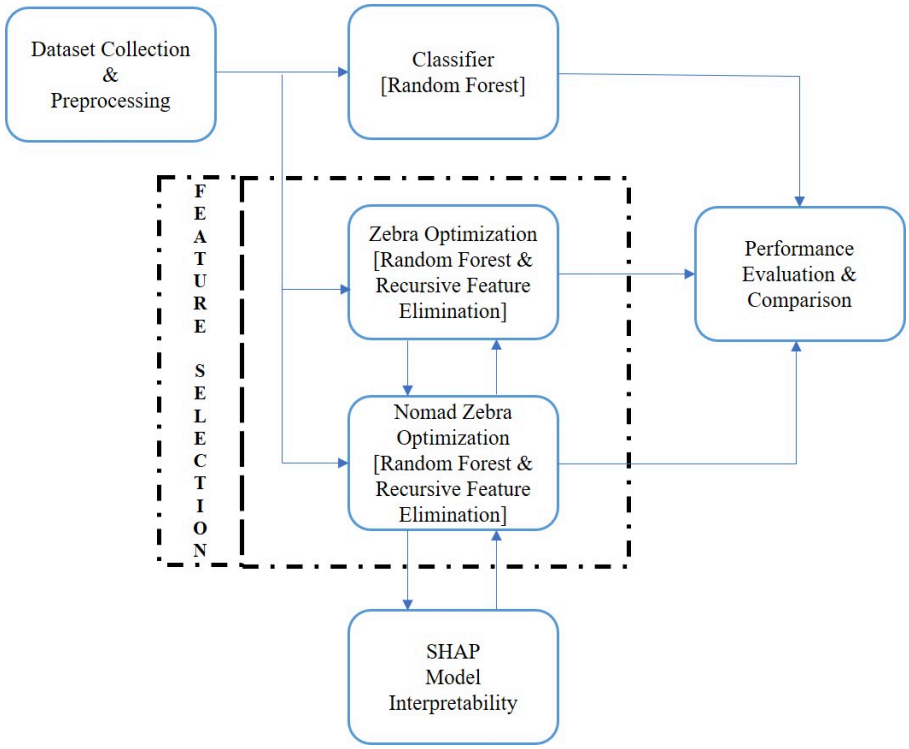

Fig. 1: Flowchart of the working procedure

Table 1: Description of Study Variable and Their Categories

Variable	Description	Categories/Values
Diabetes Status	Presence of Diabetes in individuals	0 = Non-diabetic (3906), 1 = Diabetic (990)
Gender	Binary gender value	0 = Female, 1 = Male
Age	Age categories	0 = ≤ 40 Years, 1 = 40-49 Years, 2 = 50-59 Years, 3 = > 60 Years
Body Mass Index (BMI)	Weight divided by the square of height (kg/m^2)	1 = ≤ 25 Kg/m^2, 2 = 25-30 Kg/m^2, 3 = > 30 Kg/m^2
Waist Circumference	Waist circumference levels	Categorized into three levels (details not specified)
Physical Activity	Physical activity status	0 = Yes, 1 = No
Healthy Eating	Healthy eating habits	0 = Yes, 1 = No
Blood Pressure	Blood pressure medication status	0 = Medication not taken, 1 = Medication taken
Family History of Diabetes	Family history categorized by relation	0 = No family history, 1 = Grandparents affected, 2 = Parents affected
Smoking Habits	Smoking status	0 = Non-smoker, 1 = Smoker
Heart Disease	Presence or absence of heart disease (based on fasting plasma glucose levels)	Binary value indicating presence or absence (details not specified)

ML Classifier

Classifier is an ML-based algorithm that is used to predict predetermined classes. It used complex mathematical and statistical methods to identify the likelihood of the data being recognized in the label form. There are different classifiers such as Naïve Bayes, Random Forest (RF), Logistic Regression, k-Nearest Neighbors and Support Vector Machine (Amancio et al., 2014; Alnuaimi and Albaldawi, 2024). We have used an RF classifier for our dataset as the dataset is balanced and has a numerical value. There is substantial evidence proving that RF classifier have outperformed other classification algorithms in disease classification, including cardiovascular diseases, cancer, heart diseases, skin diseases, and arrhythmia (Kumar and Sahoo, 2017; Palimkar et al., 2022; Mao et al., 2020; Bulbul et al., 2023).

Random Forest Classifier

Random forest is an efficient ensemble learning method in machine learning that combines decision trees to provide accurate outcomes in classification and regression tasks. The classifier uses a random selection of data and features to mitigate the possibility of overfitting. Classification involves majority voting as the final decision-making process, while regression depends on average forecast. When faced with a complex dataset with various data types of features, practitioners often

select Random Forest as their base classification model. The successful application of random forest for diabetes leverages the strength of performance as its attributed nature as an ensemble method (Octavially et al., 2022).

Metaheuristic Algorithm

This subsection discusses the metaheuristic algorithm, a computationally intelligence paradigm used to solve intricate problems. Most problems in engineering and daily life are in non-linear time space (Agrawal et al., 2021; Abdel-Basset et al., 2018). So, using the heuristic approach to solve those problems is relatively inefficient as most of them is designed to tackle a particular problem. However, metaheuristic algorithm can address all the liner and non-linear problems classified as NP- hard (Mauricio et al., 2020). Based on metaphor and non-metaphor characteristics, this algorithm has different searching schemas like Genetic Algorithm (GA), Harmony Search (HS), Sine Cosine Algorithm (SCA), Simulated Annealing (SA) and Teaching-Learning-Based Optimization (TLBO). Feature selection deals with irrelevant and noisy features. There are three approaches in feature selection: Wrapper Methods, Embedded Methods, and Filter Methods. The Wrapper method relies on subset generation based on a modeling algorithm (Agrawal et al., 2021). This modeling algorithm used different search strategies. In this stage, a metaheuristic algorithm is employed to find optimal value for each solution, which leads to the most optimal feature set selection (Raihan et al., 2023; Agrawal et al., 2021).

Zebra Optimization

This study uses a bio-inspired metaheuristic algorithm called the Zebra Optimization algorithm (ZOA) (Trojovska et al., 2022). The Zebra optimization starts by initializing the population size of zebras and the number of iterations. First, the position of the zebra is randomized into the population matrix, and the objective function is evaluated based on given conditions for each position. The positions of the zebra is updated through a specific number of iterations. There are two main phases to updating zebra positions: foraging behavior and defense strategies against predators. In the foraging phase, the position gets updated based on pioneer zebra position and random factor, as shown in Equation (3). Then, compare the values if they show an improved solution, as per Equation (4). Next, for defense strategy based on random probability, zebras are involved in one of two strategies of Equation (5): Strategy 1 is employed when the probability is less than 0.5 (exploitation phase), while Strategy 2 is employed when the probability is greater than 0.5 (exploration phase). The updated new positions are evaluated according to Equation (6). This process continues until the best candidate solution for the optimization problem is found. The optimizer follows the following mathematical equations model (Trojovska et al., 2022).

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N\times M} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,m} \\ x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,m} \end{bmatrix}_{N\times M} \tag{1}$$

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N\times 1} \tag{2}$$

$$x_{i,j}^{\text{new},P1} = x_{i,j} + r \cdot (PZ_j - I \cdot x_{i,j}) \tag{3}$$

$$X_i = \begin{cases} X_i^{\text{new},P1}, & \text{if } F_i^{\text{new},P1} < F_i \\ X_i, & \text{otherwise} \end{cases} \tag{4}$$

$$x_{i,j}^{\text{new},P2} = \tag{5}$$

$$\begin{cases} x_{i,j} + R \cdot (2r-1) \cdot \left(1 - \frac{t}{T}\right) \cdot x_{i,j}, & P_s \le 0.5 \quad \text{(S1)} \\ x_{i,j} + r \cdot (AZ_j - I \cdot x_{i,j}), & \text{otherwise} \quad \text{(S2)} \end{cases} \tag{5}$$

$$X_i = \begin{cases} X_i^{\text{new},P2}, & \text{if } F_i^{\text{new},P2} < F_i \\ X_i, & \text{otherwise} \end{cases} \tag{6}$$

These equations are the basic algorithm behind Zebra Optimization (Trojovska et al., 2022) depicted in Table 2.

Table 2: The basic algorithm of Zebra Optimization

Zebra Optimization Algorithm	
Step 1	Input: The optimization problem information.
Step 2	Set the number of iterations (T) and the number of zebras population (N).
Step 3	Initialization • Initialize the positions of *N* zebras • Evaluate the objective function for each zebra
Step 4	Initialize the iteration counter: `iter = 1`

Step 5	While `iter ≤ T` do:
	• Repeat until the maximum number of iterations (T) is reached
	• Update the pioneer zebra's position and status
	• For each zebra in the population:
	  • Foraging Behavior: Update zebra's status and position using specific equations
	  • Defense Against Predators: Generate a random probability (Ps)
	  • If: P_s< 0.5: Use Strategy 1 (exploitation) to update status
	  • Else: Use Strategy 2 (exploration) to update status and position
	  • Track the best solution found so far
	  • Increment the iteration counter.
	• End when all iterations are completed

Step 6	The best solution obtained by ZOA for the given optimization problem is identified.
Step 7	End the optimization algorithm.

However, no optimization algorithm is suitable to address all types of problems (Droste et al., 2002). According to No Free Lunch (NFL) theorem with a search heuristic for each objective function that can be optimized efficiently leads to the creation of new functions where the same heuristic can perform very poorly (Droste et al., 2002; Wolpert and Macready, 1997). For this reason, we will propose a new version of the zebra optimization algorithm suitable to tackle our objectives to select an optimal feature set efficiently.

Nomad Zebra Optimization

In this section, we will introduce the proposed modification of the zebra optimization algorithm named nomad zebra optimization. This proposed version works on the same mathematical Equation of zebra optimization algorithm. But to increase new search space and reduce the time complexity, we are adding a track new candidate solution step after the defense strategy using Equation (5), where it will store all the new candidate solutions and compare whether the solutions are unique or not. Because of the use of random variables in Equation (5), the new solution falls into the local minima loop, which had made the algorithm exhaustive and give the candidate solution in a given range of values. So this track new candidate solution help to increase the search space of random value. The new candidate solution cannot go to the next iteration without satisfying these sections. In Figure 2 the dotted line is the new proposed modification from the existing zebra optimization algorithm.

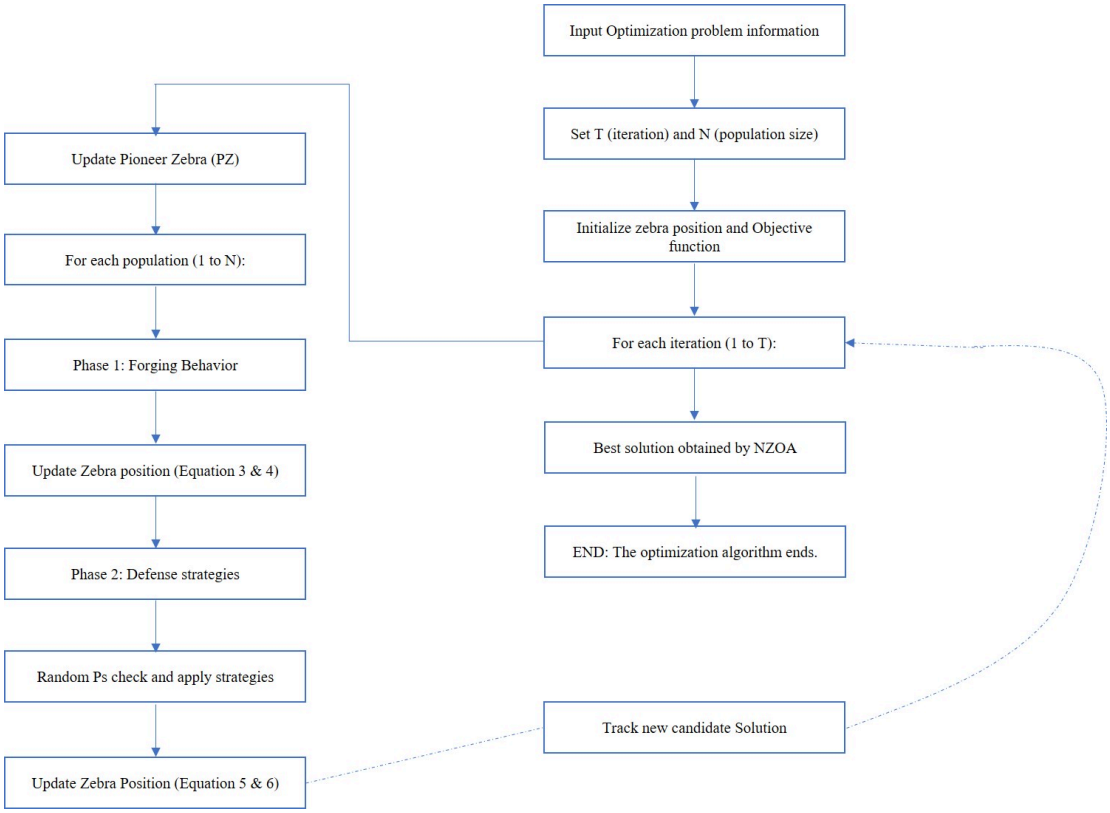

Fig. 2: Flowchart of Nomad Zebra Optimization

Feature Selection

The feature selection approach is essential in machine learning and data science (Chandrashekar and Sahin, 2014; Miao and Niu, 2016; Dash and Liu, 1997). The primary objective is identifying attributes or feature sets from a dataset. It minimizes the problem of overfitting and improves prediction efficiency by applying methods like correlation-based selection, recursive feature elimination, Least Absolute Shrinkage and Selection Operator (LASSO), and Principal Component Analysis (PCA). (Khalid et al., 2014). In the field of feature selection, choosing an algorithm can have a high impact on accuracy, and the method of feature selection varies depending on the dataset characteristic (Oreski et al., 2017). Our research enhances interpretability and performance by employing random forest and Recursive Feature Elimination (RFE) for effective feature selection. Random forest identifies each feature's importance in the decision tree ensemble, and the recursive feature elimination continuously improves models by repeatedly removing less significant features (Khalid et al., 2014; Menze et al., 2009).

Proposed ML Model

In this section, we will witness how our ML model has been designed to find optimal feature sets that

significantly impact diabetes prediction. Figure 3 shows the systematic workflow of metaheuristic base feature selection technique. The process starts with dataset loading, then defining the optimization problem, determining the goal of maximization, and storing the results of the execution period with the fitness value of each iteration. The ML model used RF classifier coupled with RFE. A preliminary check has been made in the fitness function declaration to ensure at least two features have been selected. It will split the dataset into 75-25 ratio from a valid subset of selected features. In the meantime, RF classifier is initialized, and RFE is employed to eliminate feature which has less significance in the prediction. Then, the accuracy of this prediction is calculated and returned as a fitness value. This workflow guarantees that only significant features have been trained by the model, enhancing its predictive performance.

SHAP

In data science and machine learning, it is crucial to understand the reasoning behind a model prediction, especially when accuracy is lacking due to the complexity of data.The feature selection algorithm has its own weakness, which is that it is unable to interpret the reason behind choosing the features (Marcilio and Eler, 2020). So, it is very important to understand behind the show of the feature selection for a particular disease when it comes to clinical datasets. Model interpretability is essential, leading to an agreement between accuracy and explainability. In dealing with this type of problem, a novel approach has emerged, which is known as SHAP (Le et al., 2022). The SHAP framework tackles this issue by addressing measures for the importance of new features. These indicators provide Shapley values to each feature for specific prediction and use approximation through weighted linear regression (Marcilio and Eler, 2020; Lundberg and Lee, 2017). One of the special features of SHAP is that it has both local and global explanations, which help us understand how a single datapoint contributesthe output and shows average feature contribution in negative or positive (Marcilio and Eler, 2020; Le et al., 2022).

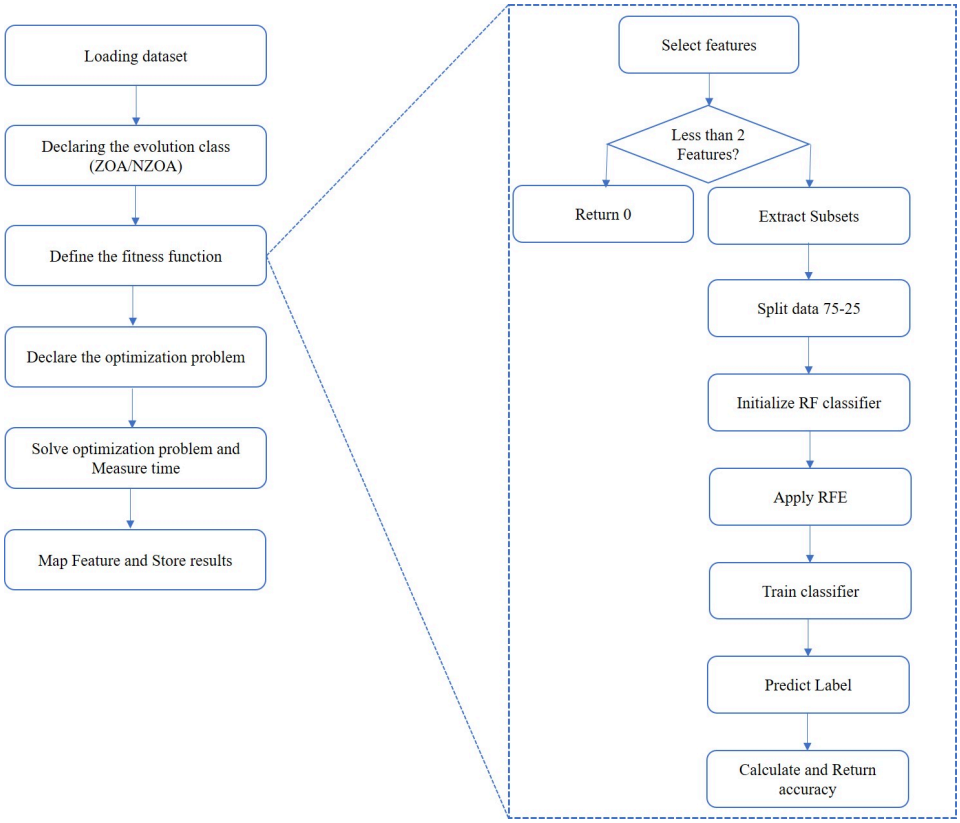

Fig. 3: Metaheuristic Optimization Process for Feature Selection

Performance Metrics

The accuracy and effectiveness of the model are measured using performance metrics. They can help measure a model's ability to precisely categorize data into different classes and also deliver insights regarding its reliability and fitness for various tasks (Hossin and Sulaiman, 2015; Vujović, 2021). The key metrics used in our work are accuracy, precision, recall, F-1 score, Receiver Operating Characteristic (ROC), and confusion matrix. Accuracy measures correctly predicted results given as a percentage. Here, TP = True positive, TN = True negative, FP = False positive and FN = False negative.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Precision assesses the correct prediction of positive instances and handles false alarms.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

Recall shows the ability to correctly identify positives and mitigate false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

The F1-Score metric measures binary classification models by considering accuracy and recall.

$$\text{F1 Score} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}} = \frac{2TP}{2TP+FN+FP} \quad (10)$$

The confusion matrix describes the number of true positives, true negatives, false positives, and false negatives. Additionally, the ROC (Receiver Operating Characteristic) visually represents the model's ability to identify positive and negative classes. The Area Under the Curve (AUC) summarizes the ROC curve into a single value, which shows the ability of the model to differentiate between classes- with a value closer to 1 representing better performance (Vujović, 2021).

Results

Classifier Performance Evaluation

In this subsection we will look at a performance metrics comparison between different ML classifiers for the dataset with 12 features.

Table 3: Classifier performance evaluation without feature selection

Model	Accuracy	Precision	F1-Score	Recall
Decision Tree	83.8	88	83	78
Random Forest (RF)	85.2	91	84	78
AdaBoost	79.5	83	78	73
Gradient Boosting	81.5	87	80	74
XGBClassifier	84.9	90	84	78
LGBMClassifier	82.9	88	81	76
CatBoost	83.94	88	81	76

Based on the comparative analysis presented in Table 3, the RF classifier demonstrated superior performance, with an accuracy of 85.2%, precision of 91%, recall of 78%, and an F1-score of 84%. Therefore, RF was chosen as the base classifier for our proposed ML model.

Feature Selection

As discussed in the methodology, two subsets of features will be created from ZOA and NZOA optimizer using RFE feature selection techniques. We run our model for 10 iterations to get the most relevant and essential features of diabetes. We took each round's highest fitness value of feature as the selected feature, as position value indicates its relevancy with the target feature. The feature with higher position value is more important as shown in Table 4.

The tick indicates it has been the highest relevant selected feature from one round of iteration. Multiple ticks indicate it has appeared several times as the most relevant feature for diabetes prediction. For example "Age" feature has appeared two times out of the ten iteration period of the RFE base ZOA. In contrast, the "Smoking" feature appeared three times out of ten of the RFE base NZOA iteration periods. Table 4 shows that we get a recursive feature elimination base for Zebra optimization (RFE_ZOA), which gives seven features as relevant features for prediction. Finally, the recursive feature elimination base Nomad Zebra optimization (RFE_NZOA) also offers seven of the most relevant features for diabetes prediction.

Table 4: Selected features using different optimization algorithm

Feature	RFE_ZOA	RFE_NZOA
Region		
Age	✓✓	✓✓
Gender	✓	
BMI		✓
Waist_Size	✓	✓
Physical_Activity		✓
Diet	✓	
BP	✓✓✓	✓
Family_History	✓	✓
Smoking	✓	✓✓

Feature Subset

We have shown the selected feature sets of RFE_ZOA, and RFE_NZOA in Table 5. The features that have appeared multiple times in the iteration process were selected as having the highest fitness value for predicted diabetes patients. As these features mitigate overfit in the testing period, they contribute the most to the generalization.

Table 5: Most frequently selected feature sets

RFE_ZOA	RFE_NZOA
Age	Age
BP	Smoking

Classification Result

Now, the classification result from the selected feature subset will be examined. A RF classifier with a default hyperparameter setting was used. First the accuracy will be examined, shown in Figure 4. We got the highest accuracy from the RFE_NZOA feature set, 79.63%, and 55.98% accuracy with the RFE_ZOA feature set.

Precision, recall and F1-score values are examined in Fig. 5. As the task is to correctly identify the likelihood of having diabetes or not, models with higher precision, recall and F1-score scores are less likely to misclassify the target label. The F1-score of RFE_NZOA is 81.42%, which indicates a high-performance model. Then, we have a recall score for RFE_NZOA of 88.33% and a

precision of 75.51%. After that, the RFE_ZOA model returned an F1-score of 56.99%, a recall-score of 57.73%, and a precision score of 56.26%. After that, we will look at the Area Under the Curve (AUC), which indicates trade-offs between true positive and false positive rates.

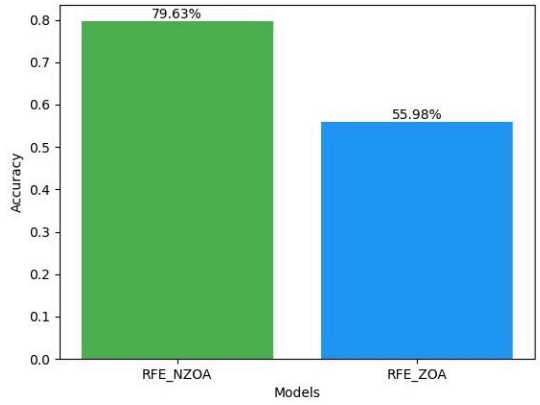

Fig. 4: Accuracy of models on different feature sets

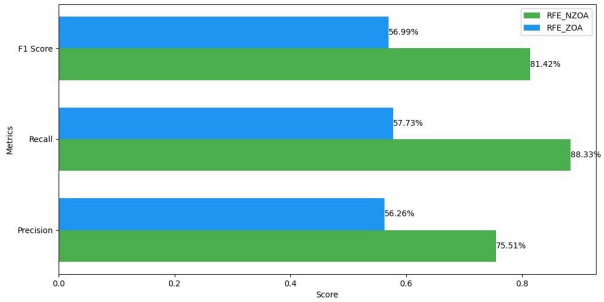

Fig. 5: Performance metrics for the selected feature subset: Precision, Recall, and F1-Score

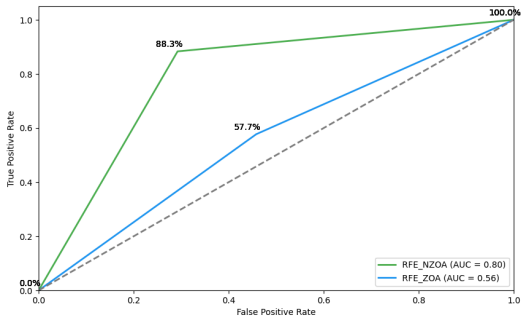

Fig. 6: AUC analysis for selected feature subsets

The RFE_NZOA model achieved an AUC value of 80% shown in Figure 6, which indicates excellent performance with high sensitivity and a low false positive rate. In comparison, the RFE_ZOA model has AUC value of 56%, suggesting poor discriminative ability. The ROC curve of RFE_NZOA model lies significantly above the diagonal reference line (AUC=0.5), showing its effectiveness in classifying positive and negative classes. In contrast, the RFE_ZOA model remains close to the diagonal line, indicating barely minimum performance than random guessing with a score of 57.7%.

Finally, the confusion matrix of the two-feature subset model is shown in Figure 7, which helps evaluate which model performs better in correctly classifying data.

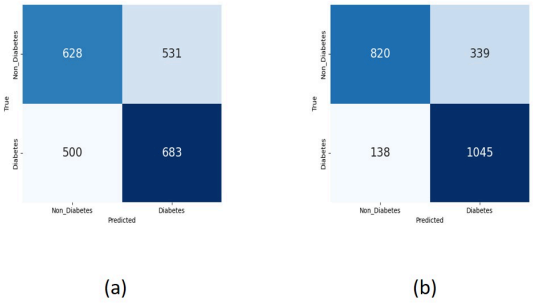

Fig. 7: Confusion matrix of the testing set: (a) RFE_ZOA, (b) RFE_NZOA

Figure 7 shows the performance of RFE_ZOA in classifying diabetes and non-diabetes patients. This model correctly classifies 683 diabetes patients out of 1183, which is 58% accurate in detecting diabetes. Then, from 1159 non-diabetes patients, it correctly classifies 628 patients, which is close to 54% accuracy for non-diabetes patients. In contrast, Figure 7 shows the superior performance of RFE_NZOA model. This model classifies 88% diabetes patient correctly, roughly 820 out of 1159, and achieves 70% accuracy to identify non-diabetes patients.

Training Period

The training period of an ML model for a specific optimization algorithm is crucial for implementing the system in a real-time environment. Our proposed model hasn't just selected important features precisely but also decreased the training period of our proposed ML model. In Table 6, the proposed RFE_NZOA outperformed RFE_ZOA by a 20% faster training period to select features from the dataset. Table 6 depicts ten rounds of the training period of each model.

Table 6: Training period of ML model for selected feature sets

Training Period	RFE_ZOA (second)	RFE_NZOA (second)
Round 1	24462	58170
Round 2	24683	15121
Round 3	23897	15406
Round 4	24114	15523
Round 5	24970	15222
Round 6	24002	14512
Round 7	23869	15122
Round 8	24857	15622
Round 9	24379	15697
Round 10	23950	14206
Average	**24318.3**	**19460.1**

Model Interpretability

The traditional ML model is like a black box. To better understand the black box and its decision-making process, we use SHAP on our model with the base dataset without any feature selection and using the RF classifier. The SHAP values are shown on the x-axis, indicating each feature's impact on model output. The y-axis shows the feature's name, and each feature has a color gradient of blue means low value and red, indicating a high value of impacts. We can see this in Figure 8 smoking and waist have a positive SHAP, indicating a risk of diabetes. While high physical activity and diet show negative SHAP values, suggesting a protective effect against diabetes. Figure 8 depicts the diabetes genre, where high smoking and waist size show stronger positive SHAP values and physical activity and diet show stronger negative SHAP values, highlighting their role in influencing diabetes risk.

Figure 9 shows a single instance observation of a non-diabetes and diabetes patient in order to know which features have a high impact in order to be non-diabetic. Figure 9 shows that the prediction f(x)=0.907 suggests a high probability of being non-diabetic, with physical activity, region and age having the most significant positive impact. Smoking and diet have a negative prediction impact, indicating a protective effect. On the contrary, Figure 9 depicts a diabetes patient's probability of predicting f(x)=1.003, indicating high chances of being detected as a diabetes patient. Here, smoking, physical activity and diet have a substantial impact on output. Family history and gender have negligible negative impact.

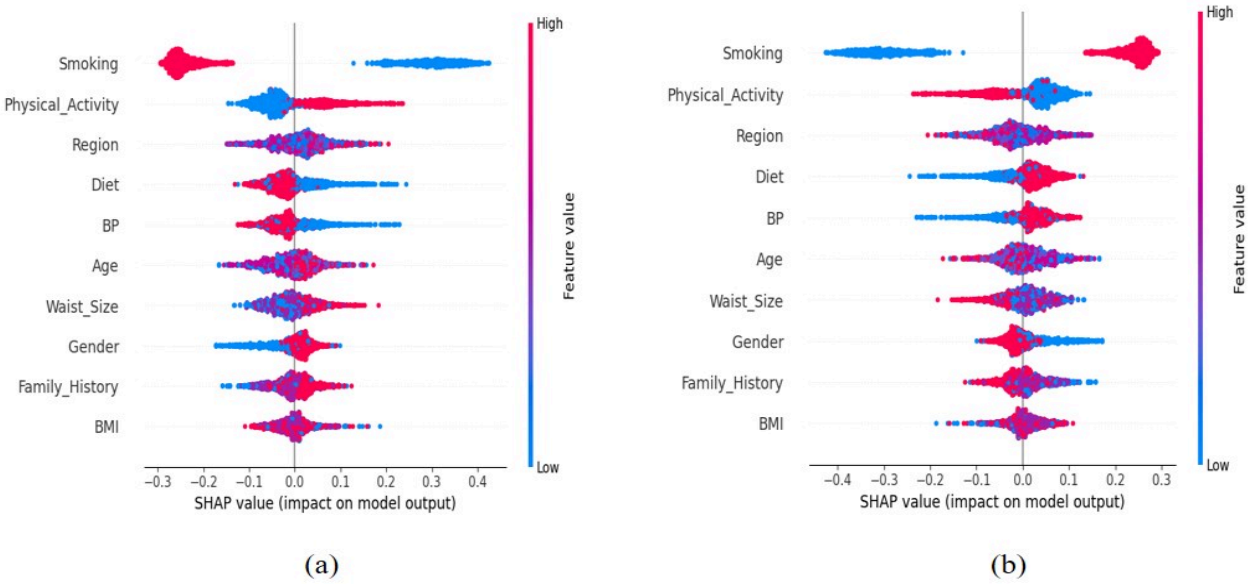

Fig. 8: Summary plot of SHAP analysis for (a) non-diabetes and (b) diabetes

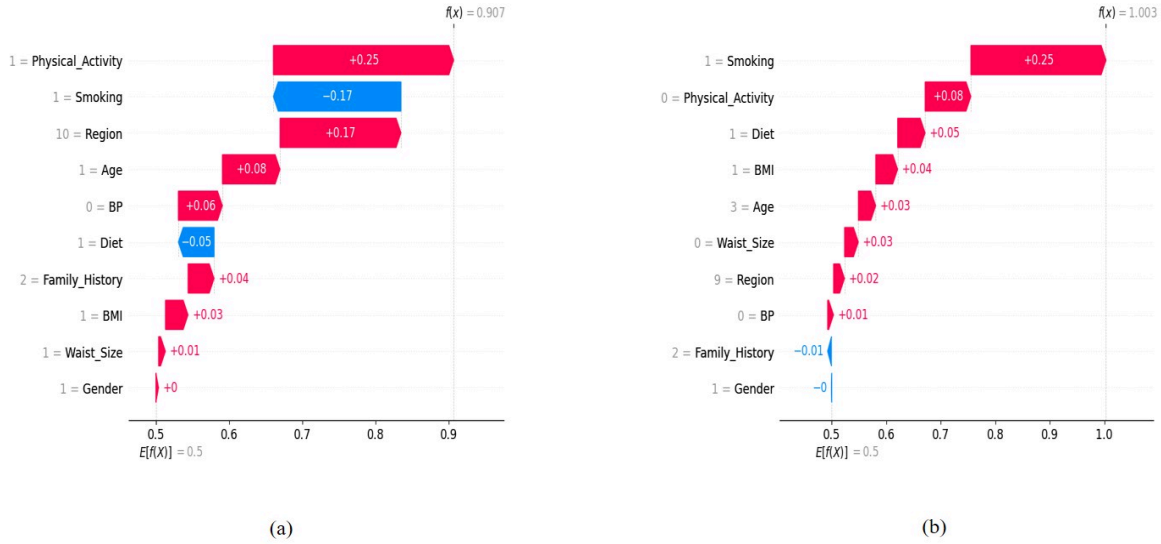

Fig. 9: Single instances observation for (a) non-diabetes and (b) diabetes

Discussion

In this study, we have proposed an innovative approach to the feature selection technique in the high-dimensional retrospective dataset. This technique focuses on phenotype characteristics based on data related to diabetes. Our proposed method, the NZOA, has outperformed the ZOA regarding feature selection for binary classification. The RFE technique discards irrelevant features in each iteration of the proposed ML model. We use an RF classifier to make a well-established model from selected feature sets from the feature selection stage. We get the highest accuracy of 79.63% from RFE_NZOA feature sets, Smoking and Age, which is notable in disease prediction using just two features. The model interpretation has been made using SHAP, which also shows that smoking is a high-impact factor for identifying whether an individual is diabetic or not. The study findings emphasize the effectiveness of NZOA in identifying valuable features associated with diabetes diseases. The research presented here improves our understanding of phenotype characteristics that may cause Diabetes illnesses, providing essential insight that could guide future diagnostic and medical approaches.

Conclusion

Diabetes is recognized as a chronic disease and often contributes to all other diseases. By predicting diabetes in the early stage, the mortality rate can be reduced. However, it is very challenging to identify at an early stage because of a lack of experienced clinicians. In such scenarios, ML can play a crucial role where there are challenges in collecting sufficient patient samples. This paper proposes an ML model capable of filtering out essential features relevant to diabetes prediction. We have successfully implemented the NZOA algorithm and extracted the most influential two features related to these chronic illnesses. The proposed model achieved an accuracy of 79.86% with a precision of 75.51%, recall of 88.33% and F1-score of 81.42%. We have demonstrated how each feature contributes to model prediction using SHAP, offering valuable information for the physician from the diagnostic point of view as well as for the patients to understand. Overall, this system is very user-friendly for inexperienced health workers or patients.

Funding Information

This Research is supported by Research Management Centre, Multimedia University.

Authors Contributions

All authors contributed equally to this study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Abdel-Basset, M., Abdel-Fatah, L., & Sangaiah, A. K. (2018). Metaheuristic Algorithms: A Comprehensive Review. *Elsevier*, *10*, 185-231. https://doi.org/10.1016/b978-0-12-813314-9.00010-4

Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019). *IEEE Access*, *9*, 26766-26791. https://doi.org/10.1109/access.2021.3056407

Alamro, H., Bajic, V., Macvanin, M. T., Isenovic, E. R., Gojobori, T., Essack, M., & Gao, X. (2023). Type 2 Diabetes Mellitus and its comorbidity, Alzheimer's disease: Identifying critical microRNA using machine learning. *Frontiers in Endocrinology*, *13*, 1-15. https://doi.org/10.3389/fendo.2022.1084656

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, *97*, 00133. https://doi.org/10.1051/bioconf/20249700133

Alqushaibi, A., Hilmi Hasan, M., Jadid Abdulkadir, S., Muneer, A., Gamal, M., Al-Tashi, Q., Mohd Taib, S., & Alhussian, H. (2023). Type 2 Diabetes Risk Prediction Using Deep Convolutional Neural Network Based-Bayesian Optimization. *Computers, Materials & Continua*, *75*(2), 3223-3238. https://doi.org/10.32604/cmc.2023.035655

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & da Fontoura Costa, L. (2014). A Systematic Comparison of Supervised Classifiers. *PLoS ONE*, *9*(4), e94137. https://doi.org/10.1371/journal.pone.0094137

Budreviciute, A., Damiati, S., Sabir, D. K., Onder, K., Schuller-Goetzburg, P., Plakys, G., Katileviciute, A., Khoja, S., & Kodzius, R. (2020). Management and Prevention Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors. *Frontiers in Public Health*, *8*, 1-11. https://doi.org/10.3389/fpubh.2020.574111

Bulbul, A. A.-M., Hossain, Md. B., Labib, M. I., & Nahid, A.-A. (2023). Classification of ECG Arrhythmias Using Conventional Tree-Based Machine Learning Approaches. *Computational Vision and Bio-Inspired Computing (Conference Proceedings)*, *1439*, 729-741. https://doi.org/10.1007/978-981-19-9819-5_52

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40(1)*, 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Cheng, Y.-L., Wu, Y.-R., Lin, K.-D., Lin, C.-H., & Lin, I.-M. (2023). Using Machine Learning for the Risk Factors Classification of Glycemic Control in Type 2 Diabetes Mellitus. *Healthcare*, *11*(8), 1141. https://doi.org/10.3390/healthcare11081141

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1-4), 131-156. https://doi.org/10.1016/s1088-467x(97)00008-5

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant Colony Optimization. *IEEE Computational Intelligence Magazine*, *1*(4), 28-39. https://doi.org/10.1109/ci-m.2006.248054

Droste, S., Jansen, T., & Wegener, I. (2002). Optimization with randomized search heuristics-the (A)NFL theorem, realistic scenarios, and difficult functions. *Theoretical Computer Science*, *287*(1), 131-144. https://doi.org/10.1016/s0304-3975(02)00094-4

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863-905. https://doi.org/10.1613/jair.1.11192

Holland, J. H. (1984). Genetic Algorithms and Adaptation. *Adaptive Control of Ill-Defined Systems*, *16*, 317-333. https://doi.org/10.1007/978-1-4684-8941-5_21

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01-11. https://doi.org/10.5121/ijdkp.2015.5201

Kakoly, I. J., Hoque, Md. R., & Hasan, N. (2023). Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique. *Sustainability*, *15*(6), 4930. https://doi.org/10.3390/su15064930

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Proeeding of the Science and Information Conference*, 372-378. https://doi.org/10.1109/sai.2014.6918213

Kumar, S., & Sahoo, G. (2017). A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis. *International Journal of Engineering*, *30*(11), 1723-1729.

Le, T.-T.-H., Kim, H., Kang, H., & Kim, H. (2022). Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method. *Sensors*, *22*(3), 1154. https://doi.org/10.3390/s22031154

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS Proceedings*, 4775.

Mao, Y., He, Y., Liu, L., & Chen, X. (2020). Disease Classification Based on Eye Movement Features With Decision Tree and Random Forest. *Frontiers in Neuroscience*, *14*, 798. https://doi.org/10.3389/fnins.2020.00798

Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *Proceeding of the SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340-347. https://doi.org/10.1109/sibgrapi51738.2020.00053

Mauricio, D., Alonso, N., & Gratacòs, M. (2020). Chronic Diabetes Complications: The Need to Move beyond Classical Concepts. *Trends in Endocrinology & Metabolism*, *31*(4), 287-295. https://doi.org/10.1016/j.tem.2020.01.007

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*(1), 213. https://doi.org/10.1186/1471-2105-10-213

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919-926. https://doi.org/10.1016/j.procs.2016.07.111

Mirjalili, S., & Lewis, A. (2016). The Whale Optimization Algorithm. *Advances in Engineering Software*, *95*, 51-67. https://doi.org/10.1016/j.advengsoft.2016.01.008

Octavially, R. P., Riskiana, R. R., Laksitowening, K. A., Kusumo, D. S., Adrian, M., & Selviandro, N. (2022). Test Case Analysis with Keyword-Driven Testing Approach on Angkasa Website Using Katalon Studio Tools. *Ultimatics: Jurnal Teknik Informatika*, *13*(2), 134-141. https://doi.org/10.31937/ti.v13i2.2391

Oreski, D., Oreski, S., & Klicek, B. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, *52*, 109-119. https://doi.org/10.1016/j.asoc.2016.12.023

Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. *Advanced Computing and Intelligent Technologies*, *218*, 219-244. https://doi.org/10.1007/978-981-16-2164-2_19

Raihan, Md. J., Khan, Md. A.-M., Kee, S.-H., & Nahid, A.-A. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, *13*(1), 6263. https://doi.org/10.1038/s41598-023-33525-0

Sakri, S. B., Abdul Rashid, N. B., & Muhammad Zain, Z. (2018). Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. *IEEE Access*, *6*, 29637-29647. https://doi.org/10.1109/access.2018.2843443

Samee, N. A., El-Kenawy, E.-S. M., Atteia, G., Jamjoom, M. M., Ibrahim, A., Abdelhamid, A. A., El-Attar, N. E., Gaber, T., Slowik, A., & Shams, M. Y. (2022). Metaheuristic Optimization Through Deep Learning Classification of COVID-19 in Chest X-Ray Images. *Computers, Materials & Continua*, *73*(2), 4193-4210. https://doi.org/10.32604/cmc.2022.031147

Sathi, N. J., Islam, Md. A., Ahmed, Md. S., & Islam, S. M. S. (2022). Prevalence, trends and associated factors of hypertension and diabetes mellitus in Bangladesh: Evidence from BHDS 2011 and 2017-18. *PLOS ONE*, *17*(5), e0267243. https://doi.org/10.1371/journal.pone.0267243

Shrestha, M., Alsadoon, O. H., Alsadoon, A., Al-Dala'in, T., Rashid, T. A., Prasad, P. W. C., & Alrubaie, A. (2023). A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes. *Multimedia Tools and Applications*, *82*(4), 6221-6241. https://doi.org/10.1007/s11042-022-13582-9

Sonia, J. J., Jayachandran, P., Md, A. Q., Mohan, S., Sivaraman, A. K., & Tee, K. F. (2023). Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm. *Diagnostics*, *13*(4), 723. https://doi.org/10.3390/diagnostics13040723

Syed, A. H., & Khan, T. (2020). Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. *IEEE Access*, *8*, 199539-199561. https://doi.org/10.1109/access.2020.3035026

Tomic, D., Shaw, J. E., & Magliano, D. J. (2022). The burden and risks of emerging complications of diabetes mellitus. *Nature Reviews Endocrinology*, *18*(9), 525-539. https://doi.org/10.1038/s41574-022-00690-7

Trojovska, E., Dehghani, M., & Trojovsky, P. (2022). Zebra Optimization Algorithm: A New Bio-Inspired Optimization Algorithm for Solving Optimization Algorithm. *IEEE Access*, *10*, 49445-49473. https://doi.org/10.1109/access.2022.3172789

Uddin, J., Ahamad, M., Hoque, N., Walid, A. A., Aktar, S., Alotaibi, N., Alyami, S. A., Kabir, M. A., & Moni, M. A. (2023). A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh. *Information*, *14*(7), 376. https://doi.org/10.3390/info14070376

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 1-8. https://doi.org/10.14569/IJACSA.2021.0120670

Wolpert, D. H., & Macready, W. G. (1997). *No free lunch theorems for optimization*. IEEE Transactions on Evolutionary Computation. https://doi.org/10.1109/4235.585893

Wong, W., & Ming, C. I. (2019). A Review on Metaheuristic Algorithms: Recent Trends, Benchmarking and Applications. *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1-5. https://doi.org/10.1109/icscc.2019.8843624