Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 2, April 2025, pp: 295-306; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received October 5, 2024; Revised December 1, 2024; Accepted December 12, 2024; date of publication January 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.626

How to cite: Dzira Naufia Jawza, Muhammad Itqan Mazdadi, Andi Farmadi, Triando Hamonangan Saragih, Dwi Kartini, and Vugar Abdullayev, "Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature Selection", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 2, pp. 295-306, April 2025.

Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature Selection

Dzira Naufia Jawza[1], Muhammad Itqan Mazdadi[1], Andi Farmadi[1], Triando Hamonangan Saragih[1], Dwi Kartini[1], and Vugar Abdullayev[2]

[1] Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia
[2] Department of Computer Engineering, Azerbaijan State Oil and Industry University, Baku, Azerbaijan

Corresponding author: Muhammad Itqan Mazdadi (e-mail: mazdadi@ulm.ac.id).

This work was supported by Lambung Mangkurat University for providing valuable resources and support.

ABSTRACT Diabetes is a critical global health challenge, classified as a non-communicable disease, affecting over 422 million individuals worldwide, with prevalence rates continuing to rise annually. This study addresses the need for more accurate diabetes prediction by evaluating the performance of Random Forest and Extreme Gradient Boosting (XGBoost) classification algorithms on a publicly available diabetes dataset from Kaggle. The research aims to improve prediction accuracy through feature selection techniques. The dataset comprises 768 records with 9 attributes, including medical indicators such as pregnancies, glucose levels, blood pressure, and BMI, with the target label categorizing outcomes as diabetic (1) or non-diabetic (0). Preprocessing was conducted to handle missing data, ensuring data reliability. Feature selection methods, namely Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), were employed to identify the most relevant attributes, enhancing the models' efficiency and accuracy. The findings revealed that without feature selection, the Random Forest model achieved an Area Under Curve (AUC) score of 0.8120, while XGBoost scored 0.7666. After applying PSO-based feature selection, the AUC scores increased to 0.8582 and 0.8250 for Random Forest and XGBoost, respectively. GA-based feature selection further improved these scores to 0.8612 for Random Forest and 0.8351 for XGBoost, demonstrating an improvement of up to 8.9%. These results highlight the effectiveness of GA in outperforming PSO for feature selection. This study underscores the significance of integrating feature selection techniques in enhancing classification model accuracy. The findings hold practical implications for developing robust predictive tools for early diabetes detection, which can facilitate timely and precise diagnoses in clinical settings.

INDEX TERMS Random Forest, XGBoost, PSO, GA, Diabetes

I. INTRODUCTION

Chronic hyperglycemia is caused by a lack of insulin or impaired insulin function, affecting the metabolism of carbohydrates, lipids, and proteins. Diabetes primarily impacts tissues like adipose tissue, skeletal muscles, and the liver due to insulin resistance. Symptoms may include increased appetite, polydipsia, weight loss, and vision issues, particularly in children with a complete lack of insulin. Some individuals, especially those with early-stage type 2 diabetes, may not experience symptoms. Without proper treatment, uncontrolled diabetes can lead to severe complications such as coma and death [1]. The history of diabetes dates back to ancient Egypt around 3000 years ago, and its impact remains significant into modern times, continuing to be a serious health concern [2].

Data from the International Diabetes Federation (IDF) indicates that in 2021, approximately 537 million adults were affected by diabetes, with this number projected to rise to 783 million by 2045 (IDF, 2021). Diabetes can cause various serious complications, such as heart disease, kidney damage,

and visual impairment, which ultimately have a negative impact on the quality of life of patients and increase the economic burden on the health system.

Early detection and accurate diagnosis are essential for the management and prevention of this disease. Utilizing machine learning algorithms is one of the effective approaches to identify and predict a disease [3][4][5]. Predicting the risk of diabetes through computer-based models can significantly reduce healthcare costs. Numerous studies have focused on modeling various diseases, including diabetes. Most of these studies train models using different features, such as pregnancy, gender, age, and BMI, employing machine learning algorithms, including classification models like Random Forest and XGBoost, to predict diabetes [6]. Random Forest and XGBoost are commonly used algorithms in medical classification due to their ability to handle complex data effectively.

Random Forest is a development of the Decision Tree method that utilizes many Decision Trees. Each Decision Tree is trained with a different data sample, and at each branch, attribute selection is done from a randomly selected subset of attributes. The Random Forest algorithm has several advantages, such as the ability to increase accuracy when there is missing data, is resistant to outliers, and is efficient in data storage. In addition, Random Forest has a feature selection process that can select the best features, thereby improving the performance of the classification model. With feature selection, Random Forest can work effectively on big data with complex parameters [7].

XGBoost (Extreme Gradient Boosting) is a supervised machine learning method that utilizes an ensemble approach, combining predictions from multiple weak learners, typically decision trees. This algorithm is a development of gradient boosting, equipped with multi-threaded optimization that maximizes the use of CPU cores to accelerate performance and improve efficiency. XGBoost is known for its speed and ability to handle large and complex datasets using parallel processing [8] . Although Random Forest and XGBoost are both powerful models, they may produce overly complex results or perform poorly without appropriate feature selection. Therefore, feature selection is a crucial step to enhance model performance. Optimization methods such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) provide effective approaches for selecting relevant feature subsets. PSO simulates social behavior to find optimal solutions, while GA uses evolutionary mechanisms to solve complex optimization problems. Both methods have the potential to significantly improve the accuracy of diabetes prediction models.

The use of Particle Swarm Optimization (PSO) aims to improve classification results through the selection of relevant features, parameter optimization, prevention of overfitting, increased accuracy, and acceleration of convergence [9] . In a study conducted by Ridho, the heart disease classification model showed an increase in accuracy after the application of feature selection techniques using Particle Swarm Optimization (PSO). This increase indicates that PSO is effective in overcoming the problem of irrelevant features, so that the model's prediction performance can be improved [3] . Genetic Algorithm is a search algorithm based on the mechanisms of natural selection and genetics. This algorithm is very effective in solving various optimization problems, from the simplest to the most complex. In addition, genetic algorithms have been proven efficient in solving Non-Polynomial problems[10]. Thus, the application of PSO and GA feature selection techniques is expected to be able to improve the results of classification performance in predicting diabetes.

This study applies the Random Forest and Extreme Gradient Boosting (XGBoost) classification methods, combined with feature selection techniques such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), to tackle the challenge of irrelevant features. It is expected that integrating PSO and GA will enable the Random Forest and XGBoost algorithms to classify diabetes disease data more accurately and efficiently, thus improving model precision. This study aims to evaluate the effectiveness of PSO and GA methods in enhancing the performance of Random Forest and XGBoost algorithms for diabetes classification. Furthermore, this research contributes to understanding the importance of feature selection in medical data processing, assisting healthcare professionals in decision-making, and optimizing the diagnostic process through the integration of advanced machine learning techniques.

The anticipated contributions of this research are: a. expanding the understanding of the application of feature selection and classification techniques in the context of health data, especially related to diabetes; b. helping medical professionals optimize the decision-making process through more in-depth analysis; c. enhancing the accuracy of data evaluation by combining the Random Forest and XGBoost algorithms with PSO and GA feature selection methods; d. Provide knowledge about the effectiveness of two feature selection methods, namely PSO and GA, in improving the accuracy of two classification algorithms, Random Forest and XGBoost, on diabetes classification problems.

A limitation of this study is the use of a relatively small and homogeneous dataset, which may limit the generalizability of the results. Additionally, the default parameters used in the model have not been optimized for other datasets. This study is expected to serve as a foundation for future research with larger and more diverse datasets, as well as the testing of other algorithms such as LightGBM or CatBoost.

II. METHOD

The research procedures are illustrated in FIGURE 1, which presents the research flowchart, starting with dataset loading, data preprocessing, and initial model evaluation, followed by feature selection using PSO and GA, and concluding with performance and computational efficiency analysis using AUC and Confusion Matrix. This study begins with loading the Pima Indians Diabetes dataset from Kaggle, which consists of 768 samples with 9 relevant attributes, including Glucose, BloodPressure, BMI, and Outcome. Data preprocessing is performed by filling missing values with 0

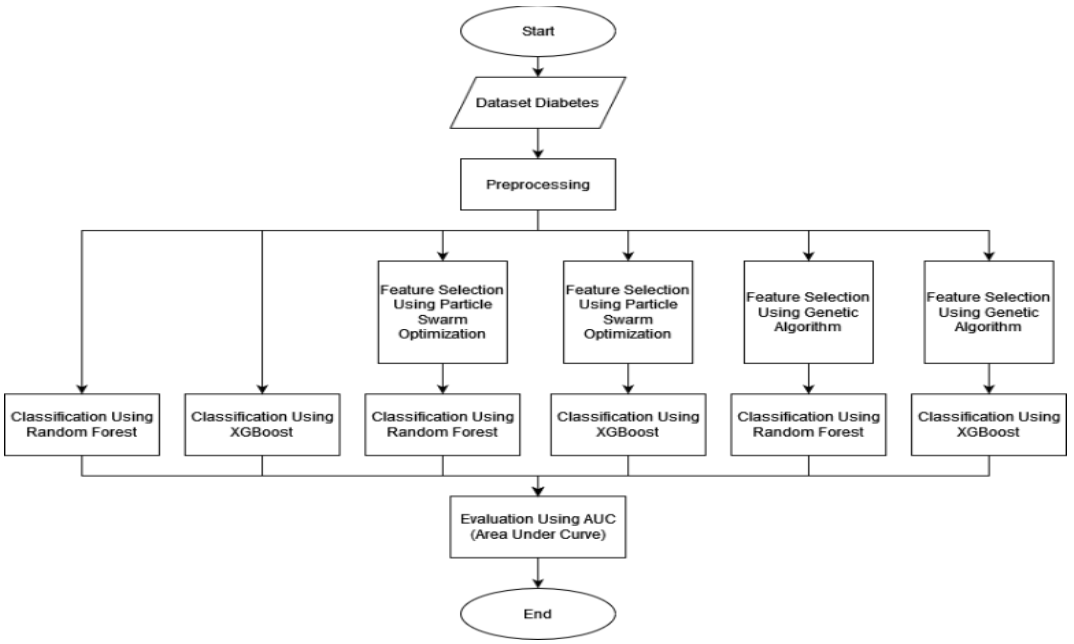

FIGURE 1. Research Flowchart

and splitting the dataset into 80% training data and 20% testing data using the train_test_split function with a fixed random state (random state = 42) to ensure reproducibility of the results. The first step involves an initial evaluation of the Random Forest and XGBoost models without feature selection to obtain a baseline performance, measured using the Area Under Curve (AUC). Next, the Particle Swarm Optimization (PSO) feature selection method is applied, where PSO selects the optimal feature subset to improve model accuracy, and the models are retrained using the selected feature subset. Subsequently, feature selection using the Genetic Algorithm (GA) is applied to compare its effectiveness against PSO, using the same models, Random Forest and XGBoost. Evaluation is conducted using AUC and Confusion Matrix to analyze classification performance. In addition to analyzing classification performance using AUC and the Confusion Matrix, this study also evaluates the computational efficiency of the feature selection algorithms employed. The evaluation focuses on several key aspects, including execution time, computational complexity, memory usage, convergence rate, and scalability. Regarding execution time, PSO is generally faster due to its simpler optimization process, while GA requires more time because of additional operations such as mutation and crossover. In terms of computational complexity, PSO exhibits lower complexity as it only updates particle positions and velocities, whereas GA involves more intricate population manipulation. PSO also demonstrates more efficient memory usage, as it only tracks particle positions and velocities, unlike GA, which demands additional memory to store populations and newly generated individuals. When it comes to convergence, PSO tends to converge more quickly; however, GA offers better solution exploration, reducing the risk of getting trapped in local minima. Scalability is another important consideration, with PSO being more efficient for small to medium-sized datasets, whereas GA is better suited for larger datasets or more complex search spaces. This evaluation aims to assess the suitability of these algorithms for practical applications, particularly those requiring high computational efficiency.

A. DATA COLLECTION

The research was conducted using the publicly accessible Pima Indians Diabetes Dataset from Kaggle. This dataset can be viewed at the following link https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database which consists of 768 data with 9 medical attributes relevant for diabetes diagnosis. These attributes include number of pregnancies, glucose levels, diastolic blood pressure, triceps skinfold thickness, insulin levels, body mass index (BMI), Diabetes Pedigree Function, age, and an outcome variable indicating whether the patient has diabetes or not. This dataset has several zero values in attributes such as insulin and skinfold thickness, which need to be considered in the analysis process. The information below provides an overview of the features and descriptions of the Pima Indian Diabetes dataset, as shown in TABLE 1.

TABLE 1
Surgery data attribute description

No	Attribute	Description	Category
1	Pregnancies	Number of pregnancies	Numeric
2	Glucose	Plasma glucose concentration after 2 during an oral glucose tolerance test	Numeric
3	BloodPressure	Diastolic blood pressure in millimeters of mercury (mm Hg)	Numeric
4	SkinTickness	Triceps skinfold thickness in millimeters (mm)	Numeric
5	Insulin	Serum insulin level after 2 hours in micro-units per milliliter (mu U/ml)	Numeric

6	BMI	Body mass index (BMI), calculated as weight in kilograms divided by height in meters squared (kg/m²)	Numeric
7	DiabetesPedigreeFunction	Diabetes pedigree function score	Numeric
8	Age	Age in years	Numeric
9	Outcome	Class label or outcome variable	Binary

B. RANDOM FOREST

Breiman first introduced the Random Forest method in 2001 . The Random Forest method has two functions for solving a case, namely classification and prediction [11]. Random Forest is an algorithm described as an ensemble method that uses many decision trees to improve prediction accuracy. Each tree is created from a subset of the data, and the final result is determined by majority voting from all trees [12][13][14]. This technique helps reduce the risk of overfitting that often occurs in the Single decision tree method [15] . The Random Forest approach involves creating multiple decision trees, with the final prediction determined by a majority vote of each individual prediction.

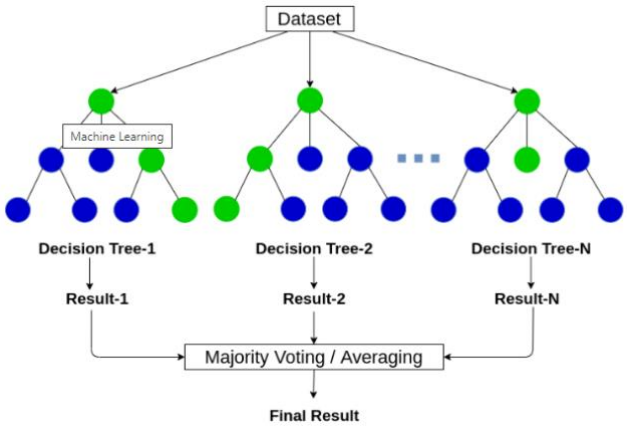

FIGURE 2. The structure of the Random Forest algorithm, where multiple decision trees are trained on different subsets of data. The final prediction is determined through majority voting or averaging.

This approach effectively addresses the problems that may arise when performing classification using only one decision tree, which often does not provide optimal results [16]. Random Forest is a method that can improve accuracy by randomly generating attributes for each node. This method consists of a number of decision trees that are used to classify data into a class. The decision tree is built by determining the root node and ending with several leaf nodes to reach the final result [17]. The process of forming a decision tree in the Random Forest method is similar to that carried out in the Classification and Regression Tree (CART), but in Random Forest there is no pruning stage. FIGURE 2 shows the mechanism Random Forest Structure. The process of forming each decision tree in Random Forest is as follows:

1. Random Sample Selection: From the available training dataset, data samples are taken randomly with replacement (bootstrap) to form a dataset of the same size as the original training dataset.
2. Random Feature Selection: From the number of available features, a subset of features is randomly selected for use in building the decision tree. Typically, the number of features selected each time is much smaller than the total number of available features.
3. Building a Decision Tree: Using a sample of data and a selected subset of features, a decision tree is built using algorithms such as ID3, C4.5, or CART. This tree is formed by splitting the data based on the most informative features, with the aim of minimizing the variance in each node of the tree.
4. Ensemble Formation: Steps 1 to 3 are repeated several times to form an ensemble of decision trees. Each decision tree in the Random Forest votes or predicts a desired class or regression value.
5. Majority Decision: The final prediction in Random Forest is obtained by taking the majority vote or average of the predictions from all decision trees in the ensemble.

C. EXTREME GRADIENT BOOSTING (XGBoost)

An ensemble boosting algorithm known as extreme gradient boosting (XGBoost) creates a more robust model by integrating multiple less effective models, or weak learners. In its attempt to improve model performance, XGBoost uses the gradient of the loss function as a reference. By paying special attention to incorrect samples in subsequent iterations, this approach focuses on correcting the prediction errors of the pre-existing model [18][19][20]. The XGBoost algorithm is described as a very efficient and effective method in handling large and complex data [21].

D. PARTICLE SWARM OPTIMIZATION (PSO)

Particle Swarm Optimization (PSO), developed by Kennedy and Eberhart in 1995, is based on the concept of simulating a basic social system. This system mirrors the behavior of a flock of birds flying toward an uncertain destination in search of food in nature [3] [9] . The movement of a flock of birds can be accurately simulated by maintaining a specific distance between each bird and its closest neighbor. This distance may vary depending on the size of the flock and the desired behavior. Particle Swarm Optimization (PSO) involves learning and applies this learning to solve optimization problems. In PSO, each individual solution, or bird in the search space, is referred to as a particle. Each particle has a fitness value determined by the objective function to be optimized, as well as a velocity that guides the particle's movement [22]. Each particle (potential solution) flies around the search space, updating its position based on the best position ever reached by the particle itself (pbest) and the best position ever reached by the entire flock (gbest). PSO is used to find the best combination of features that can produce a classification model with the highest accuracy. Each particle in PSO represents a combination of features. The position of the particle indicates the selected features [23]. The steps of

the Particle Swarm Optimization (PSO) algorithm in solving a problem are as follows:

1. Determine the number of particles to be used.
2. Randomly initialize the position and velocity of the particles.
3. Assess the fitness value of each particle based on its position using a predefined formula.
4. Identify the particle with the best fitness to serve as the Gbest.
5. The initial Pbest is identical to the initial position.
6. Update the particle velocity using the existing Pbest and Gbest with the following formula (Eq. (1)) [23]:

$$V_i^{t+1} = V_i^t + C_1 \times R_1 \times (X_L - X_i^t) + C_2 \times R_2 \times (X_G - X_i^t) \quad (1)$$

In Particle Swarm Optimization (PSO), several key variables define the algorithm's functionality. The velocity of a particle is represented by **V**, while $\mathbf{V}_i$ refers to the velocity of a particle at a specific index. The variable **t** indicates the current iteration, and **i** denotes the particle's index. The position of each particle is denoted by **X**, and its velocity is updated using random values, $\mathbf{R}_1$ and $\mathbf{R}_2$, which are generated within the range of 0 to 1. Two positive constants, $\mathbf{C}_1$ and $\mathbf{C}_2$, often referred to as learning factors, play a crucial role in influencing the particle's movement. Each particle also maintains a local best position, represented by $\mathbf{X}_L$, which is the best solution it has encountered individually. Additionally, the global best position, $\mathbf{X}_G$, represents the best solution identified by the entire swarm. These variables work together to guide the particles in exploring the search space effectively, balancing individual exploration and collective optimization.

7. Update the position of each particle with the following formula (Eq. (2)) [23]:

$$C_i^{t+1} = C_i^t + V_i^{t+1} \quad (2)$$

8. Re-evaluate the fitness value of each particle
9. Identify the particle with the best fitness to be used as GBest. For each particle, update PBest by comparing the current position with the previous iteration's PBest.
10. Verify if all particle positions have converged. If so, stop. A particle is considered to have converged if all positions are either identical or very close to the gBest particle, with no further position changes in the next iteration. Converged particle positions signify that the solution has been reached.

E. GENETIC ALGORITHM (GA)

The Genetic Algorithm (GA) is an optimization technique that refines potential solutions within a binary search space by modifying them. This space is represented by chromosomes, which are made up of a finite sequence of "0"s and "1"s. GA operates on a population of candidate solutions, gradually increasing the number of candidates in search of an optimal solution. The population evolves through genetic operators, such as selection, crossover (inheritance), and mutation. The process begins by defining a set of hyperparameters for the population, which represent possible solutions. In the context of feature selection, chromosomes act as feature masks represented as binary strings, where "1" indicates a selected feature and "0" indicates an unselected feature [12] [24]. The fitness of each chromosome is evaluated using a fitness function, often based on classification accuracy or another performance metric (Eq. (3)) [24]:

$$F(i) = \frac{Corect\ Prediction\ on\ Validation\ Set}{Total\ Predictions} \quad \forall\ i \in Population \quad (3)$$

Once fitness values are calculated, the genetic operators are applied:

1. Selection : Selects parent chromosomes based on their fitness values, often using methods like roulette wheel selection or tournament selection.

2. Crossover : Combines two parent chromosomes to produce offspring. A common formula for single-point crossover is (Eq. (4)) [24]:

$$C_{offspring} = P_1[:c] + P_2[c:] \quad (4)$$

Where P_1 and P_2 are parent chromosomes, and c is the crossover point.

3. Mutation : Introduces random changes in the chromosome to maintain diversity, typically expressed as (Eq. (5)) [24]:

$$C_{mutated}[i] = \begin{cases} 1 - C[i] \\ C[i] \end{cases} \quad (5)$$

Where $C[i]$ is the i-th gene pf the chromosome.

F. AREA UNDER CURVE (AUC)

Area Under Curve (AUC) is a method used to evaluate the area beneath the Receiver Operating Characteristic (ROC) curve. AUC serves as a metric to determine the likelihood that a classification method will assign a higher score to a positive instance than to a negative one, after selecting one instance from each category. Consequently, a higher AUC value reflects a more effective classification method. AUC is used to determine the area under the ROC curve. The AUC value is calculated by summing the trapezoidal areas of the AUC measure. The value of AUC ranges from 0 to 1, with a value closer to 1 indicating a better model performance in classifying the data. The categorization of AUC values can be seen in TABLE 2 as follows [25][26].

TABLE 2
Accuracy of Classification Results Based on AUC Values

AUC Values	Category
0,90 – 1,00	Excellent Classification
0,80 – 0,90	Good Classification
0,70 – 0,80	Fair Classification
0,60 – 0,70	Poor Classification
0,50 – 0,60	Failure

G. CONFUSION MATRIX

The performance of a classification model can be assessed based on its accuracy, which is determined using the confusion matrix. The confusion matrix is an important tool for evaluating how effectively the classifier differentiates between classes. TP and TN represent correct classifications, while FP and FN reveal errors made by the classifier. Confusion matrix is a tool used to evaluate the performance of a classification algorithm. It summarizes the model's predictions in comparison to the actual labels of the data. The confusion matrix offers insights into the model's classification accuracy by showing the counts of true positives, true negatives, false positives, and false negatives [26][27].

1) DATA COLLECTION

This study uses secondary data taken from the Pima Indians Diabetes Database, which is publicly available on Kaggle and can also be accessed through the UCI Machine Learning Repository. This dataset was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) to predict the likelihood of diabetes in Pima Indian women based on a number of health factors. This dataset consists of 768 observations with 8 predictor variables and one attribute that acts as the target variable. The target variable in this dataset is Putcome, which indicates whether the subject is diagnosed with diabetes (1) or not (0). This data was chosen because of its relevance in diabetes prediction research using machine learning techniques. The use of this dataset allows for accurate research because the dataset has been widely used in similar studies and is freely available to the public. This study uses the data to evaluate the performance of the Random Forest and Extreme Gradient Boosting (XGBoost) algorithms with Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) for feature selection.

2) PREPROCESSING

In the preprocessing stage of this study, data preprocessing was conducted to handle missing values by replacing them with zero. The dataset was then divided into features (X) and labels (y), where the features represent independent variables, while the label serves as the target variable. Next, the data was split into a training set and a testing set using the train_test_split function with an 80:20 ratio, allocating 80% for training and 20% for testing. The selected features were then processed using Particle Swarm Optimization (PSO) and the Genetic Algorithm (GA) to identify the most relevant features before being applied to the Random Forest and XGBoost models.

3) FEATURE SELECTION

Feature selection becomes more prominent, especially in datasets with many variables and features. Feature selection will eliminate unimportant variables and improve classification accuracy and performance [28]. In this study, feature selection was carried out using two optimization methods, namely Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). Both of these methods were applied to improve the accuracy of the diabetes classification model built using the Random Forest and XGBoost algorithms.

a. Feature Selection Using PSO

PSO method is applied to identify the most relevant features from the dataset. The objective function is formulated to minimize the AUC (Area Under Curve) value of the Random Forest model. In this objective function, the PSO parameter will indicate which features are selected (values greater than 0.5 are considered selected features). If no features are selected, a penalty is applied by returning an AUC value of 1. The following are the steps taken in feature selection using PSO:

1. Load the dataset and separate the features (X) from the target (y).
2. Applying PSO to find the optimal feature combination.
3. Train a Random Forest model using selected features and evaluate the model performance using AUC.

b. Feature Selection Using GA

The GA method is also used for feature selection. In this method, each individual in the population represents a subset of selected features. The fitness function is designed in the same way as in PSO, with a penalty if no features are selected. The Random Forest model is then trained on the selected features, and performance is evaluated using the AUC value. The steps taken in feature selection using GA are as follows:

1. Set GA parameters, including population size and mutation probability.
2. Running the GA algorithm to find the optimal feature combination.
3. Train a Random Forest model with selected features and evaluate its performance.

After feature selection using PSO and GA, Random Forest and XGBoost models are employed on the chosen feature subset. The results of these two feature selection methods are then compared based on the AUC value to determine which method is more effective in improving the accuracy of diabetes prediction.

4) CLASSIFICATION

The process of data analysis that involves determining a model or function to represent a concept or class of data is called classification [3] . Classification is defined as the process of grouping data based on relevant features to predict whether an

individual has a disease or not. In this study, an experiment was conducted to classify diabetes using two commonly used algorithms, namely *Random Forest* and *Extreme Gradient Boosting* (XGBoost). In data classification using the Random Forest algorithm, the Gini Index is utilized as a measure to evaluate the diversity or impurity of the nodes created at each branch of the decision tree. The Gini Index assists the algorithm in partitioning the data into more homogeneous groups, aiming to achieve more accurate classification outcomes. The Gini Index is calculated using (Eq. (6)) [29]:

$$Gini = 1 - \sum_{i=1}^{c} (pi)^2 \quad (6)$$

(Eq. (7)) [29] is is employed to determine the entropy value:

$$Enteropy = \sum_{i=1}^{c} - pi * log_2(pi) \quad (7)$$

In this case, the variable "pi" indicates the proportion of a specific class within the dataset, whereas "c" refers to the overall count of distinct classes. These two factors play an important role in statistical analysis by aiding in the understanding and interpretation of data distributions.

In addition, features selected through two *feature selection methods* , namely *Particle Swarm Optimization* (PSO) and *Genetic Algorithm* (GA), were used to improve classification performance. The classification process in this study includes the following steps:

a. Data Preprocessing

The diabetes dataset used in this study was obtained from Kaggle. This dataset consists of 768 samples with 8 feature attributes and 1 output label (Outcome). Before the classification process, data pre-processing was carried out to overcome *missing values by replacing missing values with zero. The data was then divided into training data (80%) and test data (20%) using the train-test split* technique with *random state* 42 to ensure replicable results.

b. Random Forest and XGBoost Models without feature selection

Initial evaluation was conducted to obtain a baseline of model performance, namely how well Random Forest and XGBoost can classify diabetes data without feature selection. The model was trained using all features in the dataset, and AUC was calculated as a benchmark. This baseline will later be compared with the accuracy after feature selection using PSO and GA.

1. **Random Forest** : Random Forest is an ensemble algorithm that combines multiple decision trees to improve prediction accuracy. In the evaluation stage, this model is trained using all training data and tested on test data. The evaluation results show an AUC score of 0.8120 and a confusion matrix is used to analyze the model's classification performance.
2. **XGBoost** : XGBoost is an efficient gradient boosting algorithm that is often used in machine learning competitions. The evaluation of this model was done by training on all training data and testing it on test data without cross-validation, resulting in an AUC score of 0.7666. The prediction results were tested with a confusion matrix to assess the model's classification performance.

c. Feature Selection Using PSO and GA

In addition to using native features, Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) methods are employed for feature selection with the aim of identifying relevant attributes and improving model accuracy.

1. **PSO + Random Forest** : PSO is used to select the best feature subset, which is then used to train a Random Forest model. The results show improved performance, with an AUC score of 0.8582 .
2. **PSO + XGBoost** : In this experiment, PSO is used with the XGBoost algorithm. The resulting model achieves an AUC score of 0.8250, showing a performance improvement compared to XGBoost without feature selection.
3. **GA + Random Forest** : Genetic algorithm (GA) is used to identify the most optimal features through the evolution process. The Random Forest model built using the selected features shows significant results with an AUC score of 0.8612.
4. **GA + XGBoost** : The same process is applied to XGBoost, resulting in a model with an AUC score of 0.8351. This result shows that the combination of GA with XGBoost produces better classification performance than the model without feature selection.

d. Confusion Matrix

To thoroughly evaluate the classification performance, a confusion matrix is applied to each trained model. This matrix provides insights into the number of correct and incorrect predictions, which are then visualized using a heatmap. This visualization enhances the understanding of how effectively the model differentiates between classes. Furthermore, it helps identify potential misclassification patterns that may affect overall performance. Analyzing these patterns allows for necessary adjustments to improve the model's predictive accuracy. The visualization on run 3 is shown in the FIGURE below.

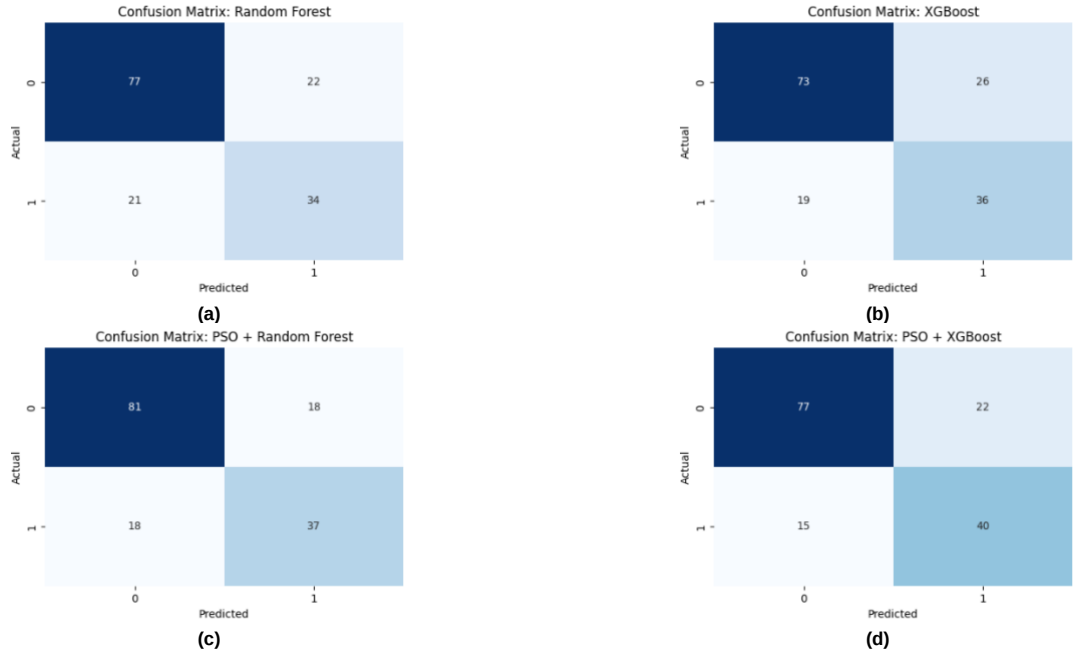

FIGURE 3. Confusion matrices and AUC values for different classification methods applied to the diabetes dataset. (a) Random Forest confusion matrix and AUC score. (b) XGBoost confusion matrix and AUC score. (c) Confusion matrix and AUC for Random Forest with PSO-based feature selection. (d) Confusion matrix and AUC for XGBoost with PSO-based feature selection.

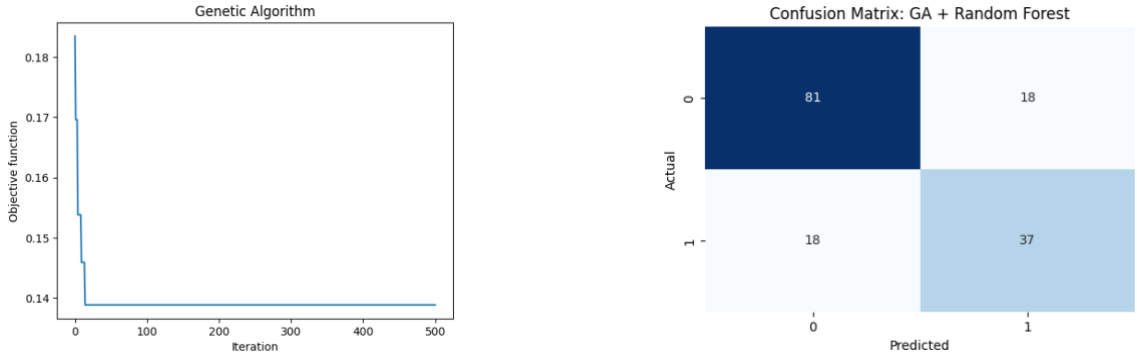

FIGURE 4. Genetic Algorithm (GA) feature selection results and classification performance. (a) Optimization process of GA showing the objective function value across iterations. (b) Confusion matrix and AUC score (0.8612) for Random Forest with feature selection using Genetic Algorithm (GA).

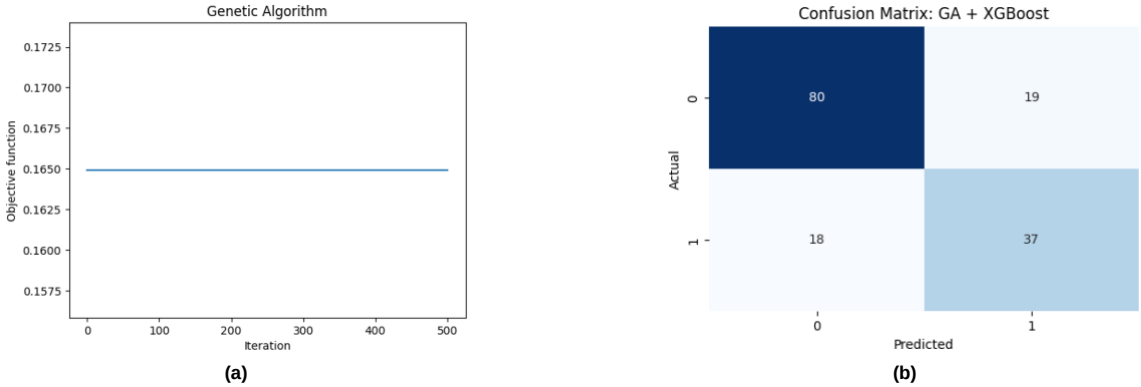

FIGURE 5. Genetic Algorithm (GA) feature selection results and classification performance. (a) Optimization process of GA showing the objective function value across iterations. (b) Confusion matrix and AUC score (0.8351) for XGBoost with feature selection using Genetic Algorithm (GA).

5) EVALUATION

Area Under the Curve (AUC) is a metric used to measure the performance of a classification model. Specifically, AUC evaluates the model's ability to distinguish between positive and negative classes, indicating how well the model can predict outcomes across various thresholds. An AUC value of 1 indicates perfect classification, while a value of 0.5 indicates no discrimination power, similar to random guessing [30]. TABLE 3 displays the classification accuracy quality based on the AUC values from the testing.

TABLE 3
Accuracy of Classification Results Based on AUC Values

AUC Values	Category
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

III. RESULTS

This study will showcase the results of AUC assessment in predicting diabetes using the diabetes dataset. The models used to predict diabetes include Random Forest using AUC without feature selection, Random Forest using AUC with PSO feature selection, Random Forest using AUC with GA feature selection, XGBoost using AUC without feature selection, XGBoost uses AUC with PSO feature selection, XGBoost uses AUC with GA feature selection. The results of the Average AUC Evaluation on Random Forest and XGBoost with and without Feature Selection (PSO and GA) in Five Tests are shown in TABLE 4.

TABLE 4
Average AUC Evaluation of Random Forest and XGBoost with and without Feature Selection (PSO and GA) Across Five Test Runs

	Run1	Run2	Run3	Run4	Run5	Rata-Rata
AUC (Random Forest)	0.8120	0.8120	0.8120	0.8120	0.8120	0.8120
AUC (XGBoost)	0.7666	0.7666	0.7666	0.7666	0.7666	0.7666
AUC (PSO+Random Forest)	0.8462	0.8612	0.8612	0.8612	0.8612	0.8582
AUC (PSO+XGBoost)	0.8351	0.8351	0.8099	0.8099	0.8351	0.8250
AUC (GA+Random Forest)	0.8612	0.8612	0.8612	0.8612	0.8612	0.8612
AUC (GA+XGBoost)	0.8351	0.8351	0.8351	0.8351	0.8351	0.8351

The evaluation results include the performance values shown in TABLE 5.

TABLE 5
AUC result values for diabetes disease

Model	AUC Values
Random Forest	0.8120
Random Forest + PSO	0.8582
Random Forest + GA	0.8612
XGBoost	0.7666
XGBoost + PSO	0.8250
XGBoost + GA	0.8351

The results showed that the Random Forest model without feature selection produced an AUC value of 0.8120, while XGBoost obtained an AUC value of 0.7666. After feature selection using Particle Swarm Optimization (PSO), the AUC increased to 0.8582 for Random Forest and 0.8250 for XGBoost. In addition, the Genetic Algorithm (GA) as a feature selection method showed better results, with an AUC of 0.8612 for Random Forest and 0.8351 for XGBoost. These results indicate that the increase in accuracy after feature selection using PSO ranged from 5.7% to 7.6%, while the increase with GA ranged from 6.1% to 8.9%, with GA providing more significant results.

IV. DISCUSSION

This study aims to evaluate the performance of the Random Forest and XGBoost algorithms in classifying diabetes using feature selection techniques such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). The results demonstrate that feature selection significantly enhances prediction accuracy in both algorithms, with Random Forest outperforming XGBoost both before and after feature selection was applied. The GA method yielded the highest AUC among all tested methods, showcasing its efficiency in search space exploration to select relevant features. However, the study also highlights certain limitations, such as sensitivity to the relatively small and homogeneous dataset, which may restrict the model's ability to generalize to a broader population. Additionally, the default parameters used may not be optimal for other datasets, suggesting opportunities for further exploration of parameter tuning.

The findings of this study have significant implications for clinical applications, particularly in supporting physicians during the diabetes diagnosis process. Machine learning algorithms like Random Forest and XGBoost, when combined with effective feature selection techniques such as GA, can provide more accurate predictions, aiding doctors in identifying high-risk patients more quickly and accurately, thereby enabling earlier medical interventions. Moreover, improved prediction accuracy can reduce diagnostic costs by eliminating irrelevant data, making the diagnostic process more efficient. However, it is crucial to validate these models on larger and more diverse datasets before widespread clinical implementation.

This research aligns with previous findings, such as the study by Ansyari et al., which demonstrated that feature selection using PSO could enhance the accuracy of Random Forest and XGBoost models in classifying heart disease [3]. And the study by Roy et al., which showed that the results without feature selection achieved good accuracy and other metrics, but there was still room for improvement, while the results with feature selection using CNN and Random Forest demonstrated a significant performance improvement in terms of accuracy, sensitivity, specificity, and AUC [31].

Similarly, Mahmud et al., integrated the C5.0 algorithm with the Chi-Square feature selection technique to enhance the accuracy of Hepatitis C classification. The results demonstrated that applying Chi-Square feature selection

significantly improved the effectiveness of the C5.0 algorithm, achieving a classification accuracy of 96.75%, surpassing previous research benchmarks. This integration highlights the substantial potential to improve the precision of Hepatitis C diagnosis through a more efficient machine learning approach. [32]. This demonstrates that the use of feature selection can improve classification accuracy, which aligns with the findings of this study. The research found that after applying feature selection, classification accuracy improved, with GA providing a more significant performance boost compared to PSO, particularly in terms of AUC. This finding supports the literature highlighting GA's superior search space exploration capabilities, especially when applied to high-dimensional datasets.

For future development, this study suggests several approaches, including using data balancing methods like SMOTE or ADASYN to address data imbalance, testing alternative algorithms such as LightGBM or CatBoost, and applying hybrid feature selection approaches that combine PSO and GA to enhance model performance. Additionally, validating the models on larger and more diverse datasets is essential to improve their generalization capability.

The application of machine learning in the medical field requires cross-disciplinary collaboration among data scientists, healthcare practitioners, and software engineers. Such collaboration is critical to ensuring that developed models are not only technically accurate but also clinically relevant and practical for real-world implementation. For instance, the features selected by the algorithms must be validated by medical experts to confirm their clinical relevance. With a multidisciplinary approach, future research can produce models that are not only technically advanced but also make a tangible impact on improving healthcare quality. As shown in TABLE 6, a comparison of the current research and similar studies highlights the differences and similarities in the methodologies and results.

TABLE 6
Comparison of Current Research and Similar Studies

Research	Algorithm	Feature selection	Evaluation	Results without feature selection	Results with feature selection
[3]	Random Forest. XGBoost	PSO	AUC	Random Forest : AUC 0.874 XGBoost : AUC 0.877	Random Forest : AUC 0.918 XGBoost : AUC 0.913
[31]	Convolutional Random Forest (Conv-RF)	Squeeze Net (CNN-based)	Accuracy	The accuracy is good, but it is lower compared to Conv-RF.	98.65%
[32]	C5.0 Decision Tree	Chi-Square	Accuracy	It is not explicitly mentioned but feature selection has been shown to improve the model's performance.	96.75%
Current research	Random Forest. XGBoost	PSO, GA	AUC, Confusion Matrix	random forest: AUC 0.8120 XGBoost : AUC 0.7666	Random Forest + PSO : AUC 0.8582 Random Forest + GA : AUC 0.8612 XGBoost + PSO : AUC 0.8250 XGBoost + GA : AUC 0.8351

V. CONCLUSION

This study aimed to evaluate the performance of the Random Forest and XGBoost algorithms in classifying diabetes using the Kaggle diabetes dataset. It also explored the impact of feature selection methods, specifically Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), on improving model accuracy.

The results revealed that applying feature selection significantly enhanced the accuracy of both models, with GA yielding a more substantial improvement compared to PSO. Specifically, PSO led to an accuracy increase of 5.7% to 7.6%, while GA achieved an increase of 6.1% to 8.9%. Furthermore, the Random Forest model consistently outperformed XGBoost in terms of prediction accuracy after feature selection was applied.

This study highlights that machine learning algorithms such as Random Forest and XGBoost, when combined with feature selection techniques like GA, can enhance predictive accuracy in diabetes diagnosis, enabling healthcare professionals to identify high-risk patients more quickly and efficiently. Further research is essential, particularly through interdisciplinary collaboration among data scientists, healthcare professionals, and software engineers, to develop more advanced AI-based applications in the future. These developments could include AI systems for hospitals that improve the speed and accuracy of diabetes diagnosis. Ethical considerations, especially regarding patient data privacy, must also be addressed in these applications. Moreover, future studies could explore hybrid feature selection techniques that integrate PSO and GA and apply these methods to larger, more diverse datasets or other non-communicable diseases.

REFERENCES

[1] S. A. Antar *et al.*, "Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments," *Biomed. Pharmacother.*, vol. 168, p. 115734, 2023, doi: 10.1016/j.biopha.2023.115734.

[2] N. F. Idris, M. A. Ismail, M. I. M. Jaya, A. O. Ibrahim, A. W. Abulfaraj, and F. Binzagr, "Stacking with Recursive Feature Elimination-Isolation Forest for classification of diabetes mellitus," *PLoS One*, vol. 19, no. 5, pp. 1–18, 2024, doi: 10.1371/journal.pone.0302595.

[3] M. R. Ansyari, M. I. Mazdadi, F. Indriani, D. Kartini, and T. H. Saragih, "Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease using Particle Swarm Optimization Feature Selection," *J. Electron.*

Electromed. Eng. Med. Informatics, vol. 5, no. 4, pp. 250–260, 2023, doi: 10.35882/jeeemi.v5i4.322.

[4] A. R. Kulkarni *et al.*, "learning algorithm to non- - invasively detect diabetes and pre- - diabetes from electrocardiogram," pp. 32–42, 2023, doi: 10.1136/bmjinnov-2021-000759.

[5] W. Feng *et al.*, "Automated segmentation of choroidal neovascularization on optical coherence tomography angiography images of neovascular age-related macular degeneration patients based on deep learning," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00757-w.

[6] M. A. Hama Saeed, "Diabetes type 2 classification using machine learning algorithms with up-sampling technique," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, 2023, doi: 10.1186/s43067-023-00074-5.

[7] H. Ghinaya, R. Herteno, M. R. Faisal, A. Farmadi, and F. Indriani, "Analysis of Important Features in Software Defect Prediction using Synthetic Minority Oversampling Techniques (SMOTE), Recursive Feature Elimination (RFE) and Random Forest," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 276–288, 2024.

[8] A. M. Akbar, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "Enhancing Software Defect Prediction through Hybrid Optimization for Feature Selection and Gradient Boosting Classification," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 169–181, 2024, doi: 10.35882/jeeemi.v6i2.388.

[9] A. P. Ariyanti, M. I. Mazdadi, A.- Farmadi, M. Muliadi, and R. Herteno, "Application of Extreme Learning Machine Method With Particle Swarm Optimization to Classify of Heart Disease," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 3, p. 281, 2023, doi: 10.22146/ijccs.86291.

[10] C. Mangla, M. Ahmad, and M. Uddin, "Optimization of complex nonlinear systems using genetic algorithm," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1913–1925, 2021, doi: 10.1007/s41870-020-00421-z.

[11] T. H. S. Li, H. J. Chiu, and P. H. Kuo, "Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm," *IEEE Access*, vol. 10, no. September, pp. 91045–91058, 2022, doi: 10.1109/ACCESS.2022.3202295.

[12] V. Maulida, R. Herteno, D. Kartini, F. Abadi, and M. R. Faisal, "Feature Selection Using Firefly Algorithm With Tree-Based Classification In Software Defect Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 223–230, 2023, doi: 10.35882/jeeemi.v5i4.315.

[13] S. S. Alahmari, D. Cherezov, D. B. Goldgof, L. O. Hall, R. J. Gillies, and M. B. Schabath, "Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening," *IEEE Access*, vol. 6, pp. 77796–77806, 2018, doi: 10.1109/ACCESS.2018.2884126.

[14] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

[15] M. K. Suryadi, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "A Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 137–147, 2024, doi: 10.35882/jeeemi.v6i2.375.

[16] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P. H. Huynh, "Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 125–136, 2024, doi: 10.35882/jeeemi.v6i2.382.

[17] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[18] E. Ismanto, A. Fadlil, A. Yudhana, and K. Kitagawa, "A Comparative Study of Improved Ensemble Learning Algorithms for Patient Severity Condition Classification," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 3, pp. 312–321, 2024, doi: 10.35882/jeeemi.v6i3.452.

[19] A. J. Weiss *et al.*, "Machine learning using institution-specific multi-modal electronic health records improves mortality risk prediction for cardiac surgery patients," *JTCVS Open*, vol. 14, no. June, pp. 214–251, 2023, doi: 10.1016/j.xjon.2023.03.010.

[20] Syarifah Aini, Wisnu Ananta Kusuma, Medria Kusuma Dewi Hardhienata, and Mushthofa, "Network-Based Molecular Features Selection to Predict the Drug Synergy in Cancer Cells," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 3, pp. 168–176, 2023, doi: 10.35882/jeemi.v5i3.307.

[21] D.- Andriansyah and Eka Wulansari Fridayanthie, "Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance," *J. Informatics Telecommun. Eng.*, vol. 6, no. 2, pp. 484–493, 2023, doi: 10.31289/jite.v6i2.8373.

[22] A. G. Gad, *Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review*, vol. 29, no. 5. Springer Netherlands, 2022. doi: 10.1007/s11831-021-09694-4.

[23] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle Swarm Optimization: A Comprehensive Survey," *IEEE Access*, vol. 10, pp. 10031–10061, 2022, doi: 10.1109/ACCESS.2022.3142859.

[24] K. A. Putri and W. F. Al Maki, "Enhancing Pneumonia Disease Classification using Genetic Algorithm-Tuned DCGANs and VGG-16 Integration," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 1, pp. 11–22, 2024, doi: 10.35882/jeeemi.v6i1.349.

[25] S. Napi, T. Hamonangan Saragih, D. Turianto Nugrahadi, D. Kartini, and F. Abadi, "Implementation of Monarch Butterfly Optimization for Feature Selection in Coronary Artery Disease Classification Using Gradient Boosting Decision Tree," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 314–323, 2023.

[26] A. A. Abdillah and Suwarno, "Diagnosis of diabetes using support vector machines with radial basis function kernels," *Int. J. Technol.*, vol. 7, no. 5, pp. 849–858, 2016, doi: 10.14716/ijtech.v7i5.1370.

[27] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[28] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.

[29] M. K. Rezki, M. I. Mazdadi, F. Indriani, H. Saragih, and V. A. Athavale, "Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5 . 0 , Random Forest , and Support Vector Machine," vol. 6, no. 4, pp. 343–354, 2024.

[30] N. Z. Al Habesyah, R. Herteno, F. Indriani, I. Budiman, and D. Kartini, "Sentiment Analysis of TikTok Shop Closure in Indonesia on Twitter Using Supervised Machine Learning," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 148–156, 2024, doi: 10.35882/jeeemi.v6i2.381.

[31] T. Sinha Roy, J. K. Roy, and N. Mandal, "Conv-Random Forest-Based IoT: A Deep Learning Model Based on CNN and Random Forest for Classification and Analysis of Valvular Heart Diseases," *IEEE Open J. Instrum. Meas.*, vol. 2, no. September, pp. 1–17, 2023, doi: 10.1109/ojim.2023.3320765.

[32] M. Mahmud *et al.*, "Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 116–124, 2024, doi: 10.35882/jeeemi.v6i2.384.

AUTHOR BIBLIOGRAPHY

Dzira Naufia Jawza is originally from Banjarbaru, South Kalimantan. Since 2021, she has continued her academic studies as a student in the Department of Computer Science, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Lambung Mangkurat (ULM). At ULM, she has focused on studying data science and various other fields of computer science. She has shown a strong interest in technology and data processing. Her deep curiosity about data analysis and information technology has led her to explore further research in the field of computer science. Her current research area is Data Science, and her final project involves research focused on diabetes classification. Email: dzirajawza81@gmail.com.

Muhammad Itqan Mazdadi is a lecturer in the Computer Science Study Program at Lambung Mangkurat University, South Kalimantan. He completed his undergraduate education in Computer Science at Lambung Mangkurat University in 2008-2013. After earning a bachelor's degree, he continued his master's education at the Indonesian Islamic University, Yogyakarta, from 2013 to 2017, with a focus on Computer Science. As an educator, he not only plays a role in transferring knowledge to students, but is also active in research and development in the fields of data centers and computer networks. His contributions in this field have had a positive impact on improving the quality of education and information technology practices. Email: mazdadi@ulm.ac.id.

Andi Farmadi a senior lecturer in the Computer Science program at Lambung Mangkurat University. He has been teaching since 2008 and currently serves as the Head of the Data Science Lab since 2018. He completed his undergraduate studies at Hasanuddin University and his graduate studies at Bandung Institute of Technology. His research area, up to the present, focuses on Data Science. One of his research projects, along with other researchers, published in the International Conference of Computer and Informatics Engineering (IC2IE), is titled "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," and this research was published in 2021. Email: andifarmadi@ulm.ac.id.

Triando Hamonangan Saragih, currently holding the position of a lecturer within the Department of Computer Science at Lambung Mangkurat University, is heavily immersed in the realm of academia, with a profound focus on the multifaceted domain of Data Science. His academic pursuits commenced with the successful completion of his bachelor's degree in Informatics at the esteemed Brawijaya University, located in the vibrant city of Malang, back in the year 2016. Building upon this foundational achievement, he proceeded to further enhance his scholarly credentials by enrolling in a master's program in Computer Science at Brawijaya University, Malang, culminating in the conferral of his advanced degree in 2018. The research field he is involved in is Data Science. Email: triando.saragih@ulm.ac.id.

Dwi Kartini is a lecturer with a strong basis in computer science. Having obtained both her bachelor's and master's degrees from the Faculty of Computer Science at Putra Indonesia "YPTK" Padang, Indonesia, she has dedicated her career to teaching and research in this field. As a lecturer in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia, She imparts her knowledge to students through courses such as linear algebra, discrete mathematics, and research methodologies. Her scholarly interests converge on the practical applications of Artificial Intelligence and Data Mining, making her a valuable asset to the academic community. She can be contacted at email: dwikartini@ulm.ac.id.