FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 39, Nº 1, March 2026, pp. 55 - 76
https://doi.org/10.2298/FUEE2601055L

Original scientific paper

OPTIMIZING CARDIOVASCULAR DISEASE DIAGNOSIS: A META-HEURISTIC AND FUZZY LOGIC-BASED APPROACH

Geeti Gayatri Lenka[1], Pranati Mishra[1], Jyotirmayee Routray[1], Meenakshi Kandpal[1], Ranjan Kumar Dash[1], Debadutta Mishra[2]

[1]School of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, India
[2]Department of Production Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, India

ORCID iDs:		
	Geeti Gayatri Lenka	https://orcid.org/0009-0005-5516-316X
	Pranati Mishra	https://orcid.org/0000-0001-7698-1149
	Jyotirmayee Routray	https://orcid.org/0000-0003-2747-3919
	Meenakshi Kandpal	https://orcid.org/0000-0002-8974-0229
	Ranjan Kumar Dash	https://orcid.org/0000-0003-3482-465X
	Debadutta Mishra	https://orcid.org/0000-0002-7215-504X

Abstract. *Cardiovascular disease (CVD) is one of the most common and major global health challenges, which requires improved methods for early and precise detection and intervention. So, to recognize these heart problems and avoid sudden cardiac arrest, it is essential to detect abnormal heart conditions early. Machine learning (ML) based medical treatments are being implemented that are very helpful in quickly and effectively diagnosing CVD problems. One method that can offer practical answers to these kinds of problems is a meta-heuristic approach. Owing to its effectiveness, meta-heuristic approaches are presently used with medical data to diagnose conditions more practically and successfully than the traditional ML methods. In this study, we used three different meta-heuristic algorithms which are Genetic Algorithm (GA), Cuckoo Search Algorithm (CSA) and Particle Swarm Optimization (PSO) for diagnosis of the CVD diseases using two different datasets – CVD and Framingham. Finally, various ML classifiers were applied on the best selected features for both the datasets, obtained from the meta-heuristic algorithms for finding efficiency and comparing the results. The results demonstrate that Framingham dataset gives best accuracy of 98.47% by using CSA algorithm for feature selection and Random Forest as classifier whereas for the CVD dataset gives best accuracy of 94.12% by using PSO algorithm and Random Forest as classifier. Then, the best performing model is passed through some fuzzy logic rules to improve the model accuracy and gives better prediction for CVD prediction.*

Key words: *CVD, meta-heuristic algorithm, optimization, feature selection, classifiers, fuzzy logic*

Received March 06, 2025; revised April 27, 2025 and May 14, 2025; accepted May 15, 2025
Corresponding author: Pranati Mishra
School of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, India
E-mail: pranatimishracse@outr.ac.in

1. INTRODUCTION

Heart disease has received a lot of interest in medical research because it is one of the fatal conditions that people around the world have had to deal with over time. It is one of the obvious illnesses that affects a lot of individuals in their mid or elderly years, and in some situations, it finally leads to fatal complications. Heart illness is more common in men than in women. Globally, CVD is the primary reason for demise [1]. The two forms of heart disease that are most common are coronary heart disease and CVD, which are developed when plaque accumulation impairs blood flow inside the arteries that supply the heart with blood. According to a 2019 report by the World Health Organization (WHO), around 1.8 crore of the fatalities worldwide are due to illnesses related to CVD [2], making them the primary cause of mortality. Heart disease is primarily caused by several factors, including elevated blood pressure, high levels of cholesterol, excessive triglycerides, and obesity. Given the rising statistics, it is imperative that this serious sickness be detected and treated promptly to prevent disease and make efficient use of medical resources. Early disease detection can be aided by a routine clinical examination. If detected early enough, coronary disease could be adequately managed or cured, and the disease can be essentially treated with the right diet, medications, and activities [3]. Machine learning prediction representation offers an improved resolution for health diagnosis of patients in the medical field. A forecast that uses machine learning approaches has a high degree of certainty and reliability in detecting CVD. Effective identification of the disease aids in early diagnosis, thereby reducing the mortality rate. Cardiovascular diseases, driven by factors like high blood pressure, genetic factor, tension, age, sex, fat levels, BMI, and unhealthy routines, are now common and overburdening healthcare systems, prompting researchers to propose early diagnosis approaches [4]. It is crucial that this condition should be diagnosed quickly and properly to protect patients' health and extend their lives.

Because of their capacity to manage big and complicated datasets, which are typical in medical diagnostics, meta-heuristic techniques [5] have drawn a lot of interest in the prediction of cardiovascular diseases (CVD). In CVD prediction, meta-heuristic techniques [6] are essential for improving the performance of ML systems by refining parameters, selecting the most relevant features, and addressing challenges like overfitting. Their adaptability and capability to manage uncertainty make them a promising solution for improving accuracy and reliability in real-world healthcare applications. A fuzzy set ensures a smooth transition between member and non-member functions by allowing an element from the supplied set. The fuzzy set offers an excellent model for the medical diagnosis system because of this and the ambiguity in the data. In a medical diagnosis system, fuzzy logic [7] is a crucial methodology in diagnosing disease because of its basic and easy-to-understand structure. Building an optimal initial-stage CVD detection framework based on the most ideal features is the main objective of this research. For this paper CVD and Framingham datasets were used for CVD prediction and comparison. Then for pre-processing, both datasets use various techniques such as normalization, standardization and one-hot encoding to remove noisy and unwanted data, outliers and prepare a pre-processed dataset for further processing. The contribution of this paper is-

1. The meta-heuristic algorithms like CSA, GA and PSO are applied to the pre-processed dataset for best feature selection. Then, this selected set of features for each meta-heuristic approach is only considered for the prediction of CVD detection using various ML classifiers.

2. For each classifier, several evaluation metrics like accuracy, precision, recall and f1-score are calculated and compared for better analysis and finding the best performing meta-heuristic approach and classifier for the two datasets.
3. The best performing model applies 50 fuzzy logic rules to improve the model accuracy and gives better predictions for CVD prediction.

The paper provides a novel contribution by leveraging three powerful meta-heuristic algorithms - GA, CSA, and PSO—for optimal feature selection from two diverse datasets: CVD and Framingham. This strategic feature selection is followed by the application of various machine learning classifiers, where their performance is evaluated using multiple metrics to identify the most effective algorithm-classifier combination. The study further introduces an additional layer of accuracy improvement by incorporating fuzzy logic rules into the best-performing models, showcasing a hybrid and intelligent framework for CVD prediction. The comprehensive comparison between datasets, algorithms, and classifiers, combined with the novel integration of fuzzy logic, offers a valuable insight into developing more precise and interpretable models for early CVD diagnosis—an advancement that holds significant potential for real-world healthcare applications.

The introduction comes first in the paper's setup, and in Section 2, there is a literature survey. Section 3 gives the details about the dataset used and the proposed methodology, along with the various pre-processing techniques. Then, Section 4 contains the description of the meta-heuristic approaches used, Section 5 describes fuzzy logic rules, and the results and discussion of the various classifiers used for CVD prediction and their performance metrics are given in Section 6. Finally, Section 7 is the final section with findings and future work.

2. Related Work

Nandakumar and Narayan in 2022 [8] highlighted the importance of accurate cardiac disease prediction, leveraging advanced methods such as Hamming distance-based feature selection and deep belief networks (DBN) enhanced by bio-inspired algorithms like the cuckoo search. These approaches demonstrated higher accuracy across various datasets, with future efforts directed toward utilizing larger datasets and ECG data for greater precision. Pathan et al. in 2022 [9] emphasized the importance of the ANOVA-F test for identifying key risk factors, improving accuracy from 0.73 to 0.75 for CVD and 0.66 to 0.71 for Framingham datasets. Optimized feature selection enhances model performance and reduces computational complexity in heart disease prediction. Doppala et al. in 2023 [10] suggested a GA-RBF hybrid model for heart disease prediction, achieving improved accuracy from 85.4% to 94.2% by reducing features. In 2023 Ansyari et al. [11] utilized the UCI dataset, applying XGBoost and RF classifiers with PSO for feature selection. PSO improved AUC to 0.913 for XGBoost and 0.918 for RF, compared to baseline values of 0.877 and 0.874. In 2021, Saha et al. [12] used data mining and ML techniques like NB, RF, KNN and DT to analyze medical datasets for heart disease prediction. These methods aim to improve early diagnosis and management by identifying key risk factors. Kaur et al. [13] reviewed ten metaheuristic techniques, like spider monkey optimization and cuckoo search, for disease prediction, emphasizing their role in optimizing feature selection and accuracy. While promising and computationally efficient, these methods require large datasets and careful implementation to minimize errors. Kanagarathinam et al. developed the "Sathvi" dataset by integrating four existing CVD datasets, creating a clean dataset with 531 instances and no missing values. Using six ml

classifiers, the CatBoost model demonstrated superior performance with an average accuracy of 94.34% through 10-fold cross-validation [14]. In their 2023 study, Baghdadi et al. [15] explored the use of diverse information sources, such as images and electronic medical records, to enhance the quick identification and diagnosis of CVD through advanced ML techniques, including SVM and DL. The CatBoost model they developed achieved an F1-score of approximately 92.3% and an accuracy of around 90.94%. In 2021, Hana H. Alalawi [16] presented a model utilizing two datasets. For the cardiovascular disease dataset, the Gradient Boosting Classifier achieved an accuracy of 73%, while the RF reached an accuracy of 94%. Rubini et al. in 2021 [17] examined the relationship between diabetes and heart disease using the RF algorithm, achieving an accuracy of 84.81%. They proposed that developing new algorithms could enhance prediction accuracy even further. Arroyo et al. [18] utilized a CVD dataset with 70,000 instances and 12 variables, optimizing ANN with GA to improve the identification of CVD. The hybrid GA-ANN model outperformed standalone ANN and other ML algorithms, achieving the highest accuracy of 73.43%. Vivekanandan et al. introduced a hybrid model that integrates modified Differential Evolution (DE) for feature selection, fuzzy AHP, and a feed-forward neural network to predict heart disease [19]. With an accuracy of 83%, the model surpasses existing models in both prediction accuracy and processing speed. T. Kasbe and R. S. Pippal [20] proposed a method of handling of patient data is unresolved designed for diagnosing CVD. The process involves fuzzification, rule base, and de-fuzzification, implemented using MATLAB. Testing shows the system achieves 93.33% accuracy, making it a useful instrument for medical practitioners in diagnosing CVD.

This inspires our study to introduce a novel approach that builds upon the rapid advancements in ML and feature selection methods for CVD prediction. By utilizing ML models, bio-inspired algorithms, and refined datasets, significant progress in accuracy has been achieved, providing key insights for creating more accurate, efficient, and scalable diagnostic systems.

3. Implementation Details

3.1. Dataset Description

For our work, we used two datasets: CVD and the Framingham dataset. The CVD dataset comprises 12 columns (11 attributes and one target variable) and 1,189 entries, curated by integrating five CVD datasets (Statlog, Long Beach, VA; Hungarian; Cleveland; and Switzerland). It aims to enhance ML algorithms for computer-aided diagnosis (CAD) by supporting clinical diagnosis and early treatment. Table 1 and Fig. 1 below gives the complete description of the CVD. The Cleveland dataset contains 303 rows and 12 columns. The Hungarian dataset contains 294 rows and 12 columns. The Hungarian dataset contains 294 rows and 12 columns. The Switzerland dataset contains 123 rows and 12 columns. The Long Beach dataset contains 200 rows and 12 columns. The Statlog dataset contains 270 rows and 12 columns. All the five datasets are integrated and downloaded from IEEE Data Port.

Table 1 Description of the CVD Dataset

Attribute	Description	Data type
age	Patients' Years of age	Numeric
gender	Gender of patient, 1 = male, 0= female;	Binary
cp	Types of chest pain	Nominal
trestbp	Patient's blood pressure (measured in mmHg)	Numeric
chol	Fasting blood glucose level of the patient (measured in mg/dl)	Numeric
fbs	Fasting blood sugar (fbs) > 120 mg/dL (1 = yes, 0 = no)	Binary
restecg	Resting electrocardiogram (ECG)	Nominal
thalach	Maximum heart rate (beats per minute)	Numeric
exang	1 = present; 0 = absence of Exercise-induced angina	Binary
oldpeak	Exercise-induced ST depression in comparison to resting levels	Numeric
slope	the slope of peak exercise	Nominal
target	1 = heart disease, 0 = Normal	Binary

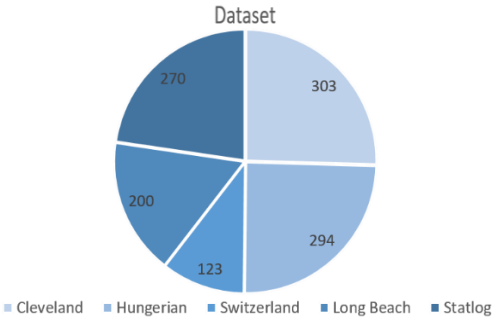

Fig. 1 Distribution of CVD dataset

The second dataset which we have used for our dataset is the Framingham dataset. It has 16 columns (15 attributes and one target variable) and a total of 4240 occurrences. The Framingham dataset, which is available on the Kaggle repository, was formed as part of a CVD study consisting of the people of Framingham, Massachusetts. Table 2 below shows the complete description of the Framingham dataset.

Table 2 Description of the Framingham Dataset

Attribute	Description	Data type
age	Patient's age	Numeric
gender	Gender of patient, 1 = male, 0= female;	Binary
education	Types of education level	Nominal
currentSmoker	Smoking status	Binary
cigsPerDay	Cigarettes smoked per day	Numeric
BPMeds	BP medication use	Binary
prevalentStroke	History of stroke	Binary
prevalentHyp	Hypertension	Binary
diabetes	Diabetes diagnosis	Binary
totChol	Total cholesterol level (mg/dL)	Numeric
sysBP	Systolic blood pressure (mmHg)	Numeric
diaBP	Diastolic blood pressure (mmHg)	Numeric
BMI	Body mass index (weight in kg/height in m^2)	Numeric
heartRate	Heart rate	Numeric
glucose	Glucose level (mg/dL)	Numeric
TenYearCHD	Target variable 1 = Developed coronary heart disease within 10 years, 0 = No	Binary

3.2. Proposed Methodology

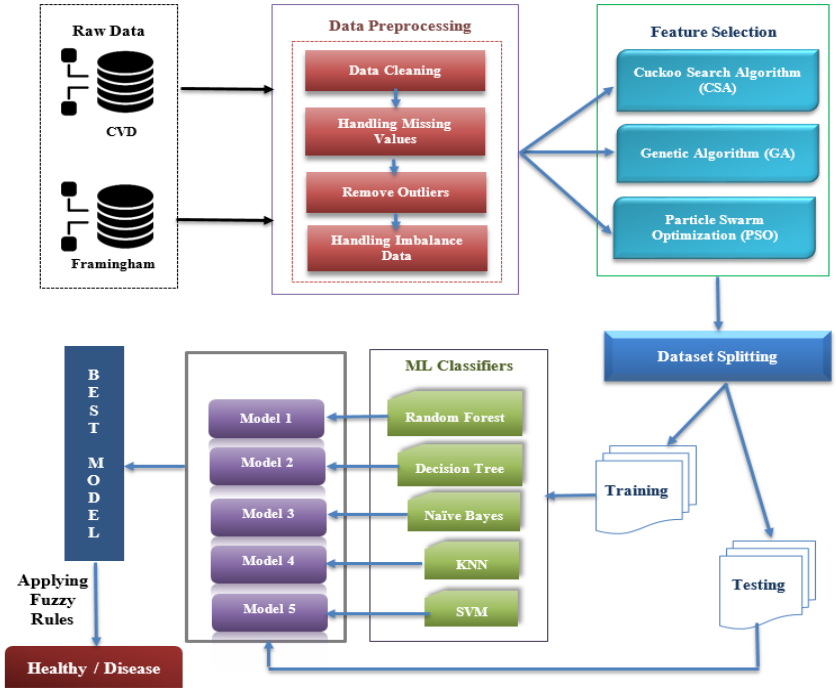

Fig. 2 Proposed Model

As shown in Fig. 2 above, there are two distinct datasets used. One dataset is collected from "IEEE DATA PORT" i.e.; CVD dataset, which has a total of 1190 occurrences. Another dataset is Framingham, which is collected from the Kaggle repository, having a total of 4520 occurrences. After collecting the data, data pre-processing was done. There are various methods used in data pre-processing, i.e. Normalization, Standardization, One hot encoding, Z-score Normalization and SMOTE (Synthetic Minority Oversampling Technique). Then, a meta-heuristic algorithm is used for feature selection, which is applied on each dataset. There are three different meta-heuristic algorithms used in this model, i.e. CSA, GA and PSO. After that, the datasets are divided into two parts- training data and testing data. The training data is used to train different classifiers. The classifiers like RF, DT, SVM, KNN and NB are used in this model. After training the model with classifiers, the testing data is used to test the model and find the performance evaluation metrics like accuracy, precision, recall and f1-score. Then, the model that achieves the highest accuracy is selected as the best model for the CVD prediction and is passed through fuzzy logic rules for achieving much better accuracy in predicting healthy and diseased patients.

3.3. Data Preprocessing

The dataset is an essential initial phase to make sure that the data is clean, regular and prepared for evaluation on ML models. This process ensures structured data, enhances model performance and enables reliable, accurate results.

The datasets for CVD prediction are subjected to pre-processing and we used different techniques like normalization, standardization, one-hot encoding and outlier filtering. Normalization rescales numerical data to a fixed range, typically between 0 and 1, by using Minmax scaling as shown in equation (1), which makes the data easier to compare and suitable for models sensitive to varying scales. Standardization transforms data to a mean of 0 and a standard deviation of 1 by using Z-score scaling as given in equation (2). One-hot encoding converts categorical values into binary columns (0s and 1s) to make them suitable for ML models. Outliers are extreme values that can distort analysis and ML models. The Z-score method is used here to detect and remove such values by using a threshold value of 3 and values with a Z-score > 3 are considered outliers. In this study, the Z-score method is used during the data pre-processing phase to clean the dataset by removing outliers, which are extreme values that could negatively impact the training of machine learning models. For each feature in the dataset, the Z-score is calculated for each data point. A threshold of 3 is applied, meaning that any data point with a Z-score greater than +3 and less than -3 is considered an outlier. These values are removed from the dataset to ensure that the remaining data is more representative of the typical pattern, improving model performance and generalization.

$$X_{min-max-scalling} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

$$X_{std} \ or \ Z - score = \frac{X - \mu}{\sigma} \tag{2}$$

Where μ is the mean and σ is the standard deviation.

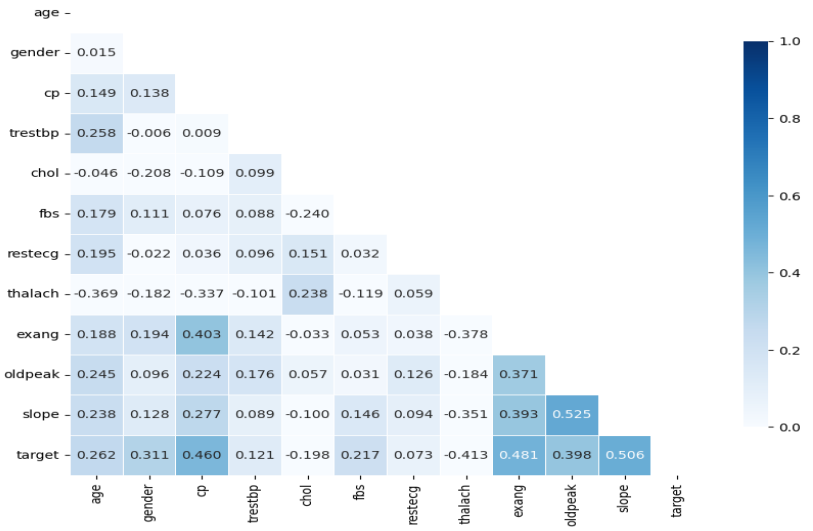

Fig. 3 (a) Feature Correlation Heatmap with Heart Disease in CVD Dataset

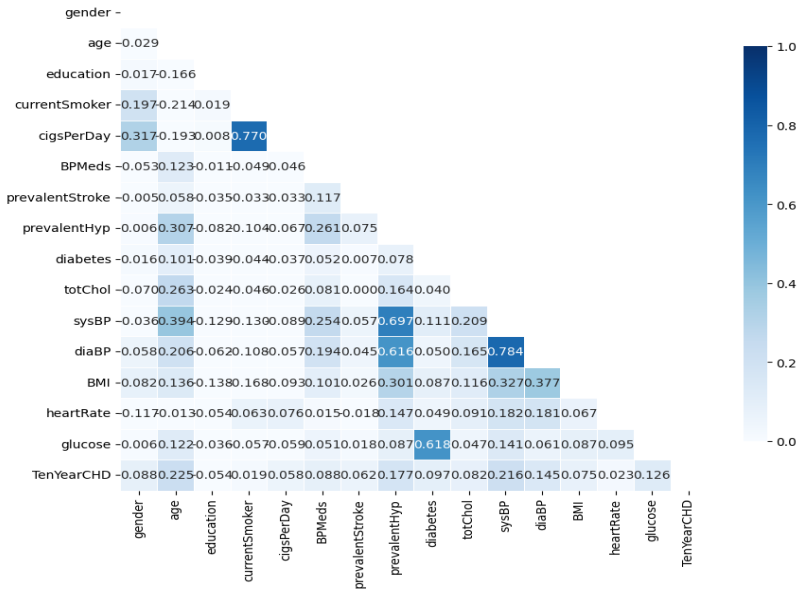

Fig. 3 (b) Feature Correlation Heatmap with Heart Disease in Framingham Dataset

SMOTE (Synthetic Minority Oversampling Technique) is a method for oversampling that addresses class imbalance in a dataset. It works by identifying minority class examples that are near each other in the feature space. Exploratory Data Analysis (EDA) involves evaluating and visualizing datasets to uncover patterns, relationships, anomalies, and missing values. The processed numerical (normalized) and one-hot encoded categorical data are combined to create a clean dataset for ML models. Fig. 3 (a) and 3 (b) represent the positive correlated heatmap for CVD and Framingham dataset respectively.

3.4. Feature Selection

Finding a subset of the attributes that are most significant from a bigger set that has the biggest influence on the result is known as feature selection. Selection of features using meta-heuristic algorithms is an advanced technique to identify the dataset's most pertinent attributes for improved model performance [21]. These methods explore and evaluate combinations of features to reduce redundancy, improve accuracy, and prevent overfitting in machine learning models. Unlike traditional techniques, meta-heuristics can handle datasets with high dimensions and complicated interactions among attributes, making them powerful for feature selection in real-world problems where exhaustive search is computationally impractical.

In this work, we have used meta-heuristic algorithms [22] like CSA, GA and PSO for the best feature selection (the features which are required to give best prediction for the CVD) from the available set of features from the dataset and by using the best selected features only, different ML classifiers were used for CVD prediction and the results are compared. The detailed explanation and working of all the meta-heuristic algorithms used are given in Section 4.

3.5. Dataset Splitting

It is possible to divide the completely known dataset into training and test sets. The "CVD" dataset's training and testing datasets are divided in a 70:30 ratio, with 833 and 357 instances respectively. The target '0' has 561 instances out of 1190, and the target '1' has 628 instances. All five classifiers were evaluated using this 70:30 dataset. Similarly, the Framingham dataset is also divided into the same ratio with 2968 instances taken for training and the rest of the instances are taken for testing purposes for predicting the target variable.

3.6. Machine Learning Classifiers

Several ML classifiers are used for making predictions, including DT, RF, KNN, SVM and NB.

Decision Tree -It is a supervised learning method to structure like a tree, with internal decision nodes standing in for features and leaf nodes for results. Decision nodes make choices that lead to branches, whereas leaf nodes deliver final outputs without any additional branches.

Random Forest - A tree-based technique called RF uses several decision trees to provide predictions. The model chooses the most popular class as its final forecast after each tree casts a vote on a class.

K-Nearest Neighbor - The KNN classifier is a supervised ML methodology used for categorizing and regression, which categorizes data points according to the classes of their nearest neighbor. Known as a lazy learner, KNN does not train a model but instead stores data and only calculates distances when queried, making it suitable for data mining.

Support Vector Machine –SVM works by finding the best hyper plane to divide data points into discrete groups while optimizing the margin between them.

Naïve Bayes- Based on Bayes' theorem, NB frequently used for classification tasks. It is simple to build and computationally effective because it works under the premise that characteristics are independent of conditions provided the category name.

3.7. Performance Evaluation

Model performance is determined by using various performance metrics, including accuracy, precision, F1-score, and recall (sensitivity). The confusion matrix helps identify the performance of each model by focusing on four essential components: True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP).

Accuracy: The percentage of patients who get an accurate diagnosis of heart disease is known as accuracy. It is computed as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision: The percentage of individuals having heart disease who are predicted to have it is known as precision. It's determined:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Re-call: It is the measure indicating the ratio of patients predicted to have heart disease to those who have it and is expressed as:

$$recall = \frac{TP}{TP + FN} \tag{5}$$

F1-score: It determines the test's accuracy. It is computed as follows:

$$f1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6}$$

3.8. Fuzzy Logic

Fuzzy logic is designed to handle reasoning that involves uncertainty and imprecision, mimicking human decision-making processes. Unlike traditional binary logic, which operates with strict binary values, fuzzy logic employs degrees of truth, allowing for more nuanced and flexible modeling of complex systems. By incorporating linguistic variables and membership functions, it is particularly well-suited for applications requiring approximate reasoning, such as control systems, data classification, and decision-making under ambiguity. Its adaptability makes it a valuable tool for addressing real-world problems where exact values are difficult to ascertain. Fig. 4(a-e) represents membership functions of different attributes of Framingham dataset. Table 3 below represents some of the fuzzy logic rules which we have applied to our best performing model.

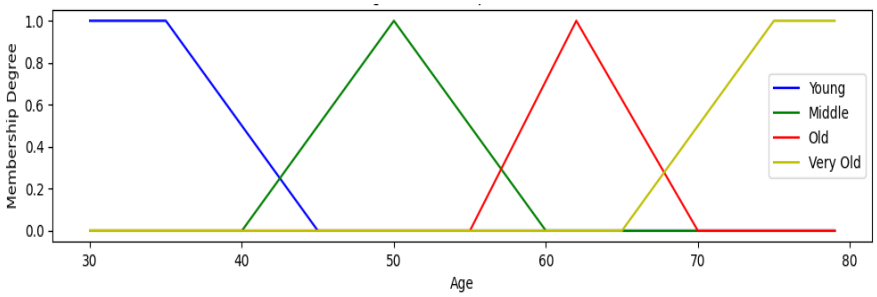

Fig. 4(a) Membership Function of age

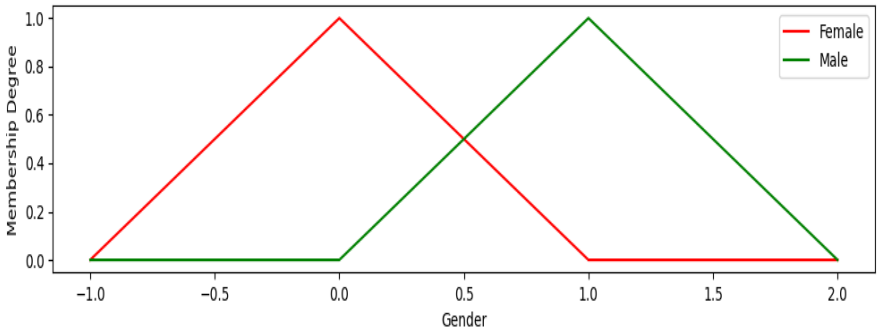

Fig. 4(b) Membership Function of gender

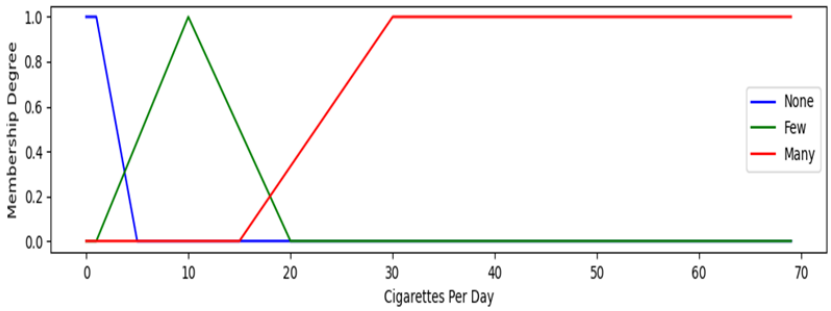

Fig. 4(c) Membership Function of cigPerDay

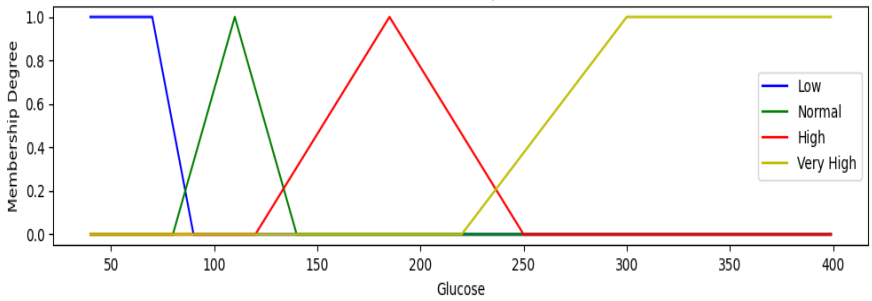

Fig. 4(d) Membership Function of glucose

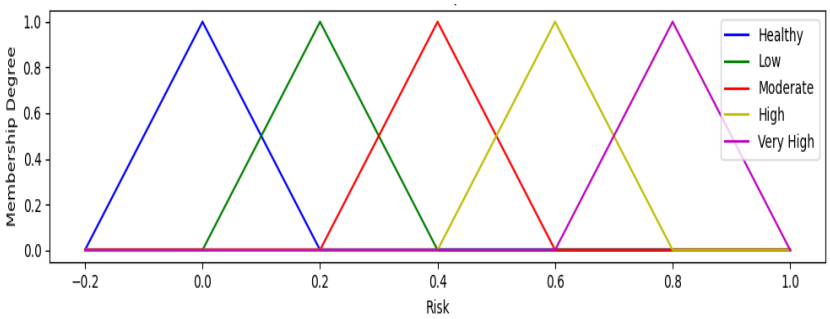

Fig. 4(e) Membership Function of risk

4. META-HEURISTICS ALGORITHMS

Meta-heuristic algorithms are high-level optimization techniques designed to solve complex problems where traditional approaches like brute-force or exhaustive search are inefficient [23]. These algorithms draw inspiration from processes and events seen in nature, taking cues from how nature finds efficient and effective solutions. In summary, meta-heuristic algorithms mimic natural processes to optimize complex problems efficiently.

An optimization algorithm is a procedure used to analyze different solutions to find the best or most practical one. The search and optimization techniques offered by several iterations of nature-inspired algorithms significantly influence the diagnostic procedure

and results. This has an impact on the algorithm's capacity to navigate across complex search spaces and its convergence time to the optimal answer. Now, we will look at some of the meta-heuristic based optimization algorithms used in our study for feature selection. Table 3 below shows the feature selected by the three meta-heuristic algorithms for both the datasets.

4.1. Cuckoo Search Algorithm (CSA)

One significant population-based optimization technique is Cuckoo Search. Smearing their eggs in the shelters of various feeding birds is the basic idea behind it. Inspired by breeding behavior, Cuckoo Search is used to solve several optimization problems. The algorithm relies on two main concepts:

Levy Flight: A random walk, used to search for optimal solutions, which is inspired by the movement patterns of the cuckoo. u and v are random variables drawn from normal distributions, and their ratio defines the step size and value of 1.5 is set to beta.

$$step = \frac{u}{|v|^{\frac{1}{beta}}} \tag{7}$$

Nest Evolution: The algorithm uses a population of solutions, or "nests," and iteratively improves them by searching for better solutions and replacing the worst ones.

In this work, CSA is applied for feature selection, where each solution (or "nest") indicates a portion of the features that have been chosen. The algorithm explores the search space using Levy flight, which allows it to make large and small random jumps, improving the exploration of potential feature subsets. The approach balances exploration of new solutions with exploitation of good solutions, enabling effective feature selection for classification tasks.

Fitness function: The feature selection is done by converting the solution vector into a binary decision - If the numerical value of an element is larger than 0.5, the corresponding feature is selected. If no features are selected, assign a high penalty (1e6) to discourage this solution. Since the goal is to maximize accuracy, the fitness function returns the negative accuracy to fit the minimization framework of the algorithm.

Algorithm 1 CUCKOO SEARCH ALGORITHM

1. Start the process
2. Initialize the parameters: num_nests, dim, pa, max_iter, lower_bounds, upper_bounds
3. Initialize nests with random solutions within bounds
4. Evaluate fitness of all nests using the objective function
5. For t in range (max_iter):
6. For each nest:
   6.1. Generate a new solution using Levy flight
   6.2. Clip the solution to stay within bounds
   6.3. Evaluate the objective function for the new solution
   6.4. If the new approach is more fit:
7. Replace the current nest with the novel approach
8. Replace worst nests with random solutions within bounds
9. Return the nest with the highest fitness as the optimum solution
10. End of Process

4.2. Genetic Algorithm (GA)

GA is a technique for optimization inspired by environment that replicates the evolutionary ecology and biological selection processes. Darwin's theory of "survival of the fittest," according to which the strongest individuals have a higher chance of surviving and procreating, so transferring on their characteristics to the following generation, is the foundation of this theory. In GA, chromosomes (or people) are used to represent possible solutions to an optimizing issue, which are typically encoded as binary strings or real-valued vectors. The technique starts with a sample of arbitrarily produced solutions and uses crossover, mutation, and selection operations to iteratively improve them.

Selection: The best-performing individuals (based on a fitness function) are selected to create offspring for the following generation. This guarantees that superior solutions are more likely to spread their characteristics.

Crossover (Recombination): To create children that combine the strengths of both parents, pairs of parent solutions trade chromosomes.

Mutation: Small random changes are introduced in the offspring's chromosomes to maintain diversity and stay clear of local optima traps.

Fitness Function: It calculates the negative accuracy on the test set. The negative sign is used because GA minimizes the objective function, but accuracy is maximization metric. A large penalty (1e6) is imposed if no features are selected, discouraging invalid solutions. The fitness function for GA is given as:

$$\text{Fitness (solution)} = \begin{cases} -\text{accuracy}, & \text{if any features are selected} \\ 10^6, & \text{if no features are selected} \end{cases}$$

Algorithm 2 GENETIC ALGORITHM

1. Start the process
2. Initialize the population size and set algorithm parameters like mutation, crossover, etc
3. Evaluate initial population:
   - 3.1. For each solution in the population:
     - 3.1.1. Identify selected features based on binary representation
     - 3.1.2. Train a classifier on selected features
     - 3.1.3. Compute the negative accuracy as the fitness value
4. Perform iterative optimization:
5. While stopping criteria (max iterations or no improvement) is not met:
   - 5.1. Selection: select a subset of the population as parents based on fitness values
   - 5.2. Crossover: perform uniform crossover on selected parents to create offspring
   - 5.3. Mutation: randomly mutate bits in the offspring with a given mutation probability
   - 5.4. Evaluation: calculate fitness of the offspring using the same step as in step 3
   - 5.5. Replacement: replace the least-fit individuals in the population with the offspring
6. Return the best solution (feature subset) found during optimization
7. End the process

4.3. Particle Swarm Optimization (PSO)

PSO is an optimizing technique inspired by nature based on the collective nature of groups observed in nature, such as flocks of birds or schools of fish. Based on the population stochastic optimizing method it finds the best answer by mimicking the motion and collaboration of particles (potential solutions) in a multidimensional searching area. It

iteratively updates the position and velocity of particles, adjusting the feature subset based on their personal most well-known solution and the most well-known solution in the swarm.

Fitness Function: It quantifies how good a subset of features is for building a predictive model. It ensures that the PSO algorithm maximizes the model's classification accuracy by selecting relevant features and penalizes solutions that select no features, encouraging meaningful feature subsets. The fitness function is set to be negative of accuracy other than any values of the condition given below. It is mathematically represented as:

$$F(x) = C \text{ if } \sum_{i=1}^{n} (x_i > 0.5) = 0 \tag{8}$$

Algorithm 3 PARTICLE SWARM OPTIMIZATION ALGORITHM

1. Start the process
2. Initialize the swarm - randomly set positions and velocities of all particles
3. Evaluate the fitness using objective function
4. Update personal best (pBest):
   4.1. Compare the current position of each particle with its optimal position
   4.2. Update the personal record if the current position is superior
5. Update global best (gBest):
   5.1. Find the best position among all particles
6. Update velocity:
   6.1. Calculate each particle's new velocity based on inertia, pBest and gBest
7. Update position by using its updated velocity
8. Check stopping criteria stop if max iteration is reached or go back to step 3
9. Output the global best solution
10. End the process

Table 3 Feature Selected by Meta Heuristic Algorithm for both the datasets

Dataset	Meta Heuristic Algorithm	Feature Selected
CVD	CSA	trestbp, chol, oldpeak, cp, slope
	GA	thalach, chol, oldpeak, cp, restecg, slope
	PSO	trestbp, chol, oldpeak, cp, slope, restecg
Framingham	CSA	gender, age, education, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, glucose
	GA	gender, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, BMI
	PSO	gender, age, education, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, diaBP, BMI, heartRate

5. FUZZY LOGIC RULES

The figure shown in Fig. 5 represents a fuzzy logic system for risk prediction. It processes inputs such as gender, age, education, cigarettes per day, history of stroke, hypertension and blood pressure medications through a fuzzifier, which converts these crisp inputs into fuzzy values. These fuzzy values are then evaluated using some inference rules as shown below in Table 4 to generate fuzzy outputs based on predefined rules. The de-fuzzifier translates these fuzzy outputs back into a crisp value, which

represents the predicted risk level. The system combines linguistic reasoning and quantitative data to deliver a risk prediction outcome. De-fuzzification process transforms abstract, rule-based fuzzy results into a clear and usable prediction value. It plays a vital role in making the fuzzy logic system practical and meaningful for CVD risk prediction, bridging the gap between human-like reasoning and machine-based precision.

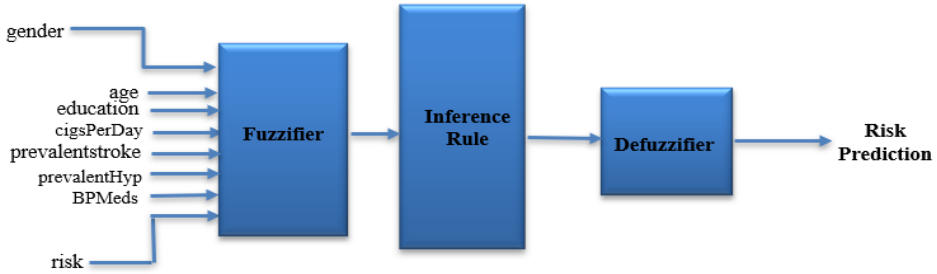

Fig. 5 Fuzzy Inference System

Table 4 Fuzzy Logic Rules

Rule No.	Gender	Age	education	cigsPerDay	BPMed	Preval-stroke	Prevalent Hyp	Glucose	Risk
1	-	young	high	none	no	no	no	normal	healthy
2	female	-	Very high	none	no	no	-	low	healthy
3	-	young	medium	few	no	no	-	normal	low
4	female	young	medium		no	no	-	normal	Low
5		middle	medium	few	no	-	-	high	moderate
6	male	middle	-	many	no	-	-	normal	moderate
7	-	old	low	many	yes	-	-	very high	High
8	male	Very old	low	-	yes	-	-	Very high	High
9	male	old	-	many	yes	yes	-	-	Very high
10	-	Very old	low	-	yes	-	yes	Very high	Very high

6. Results and Discussions

The results of various performance evaluation metrics for different classifiers for both the datasets are shown above. Table 5 and Fig. 6 represent the evaluation of various metrices using CVD dataset and similarly Table 6 and Fig. 7 represent the evaluation of various evaluated metrices using Framingham dataset. As shown in Fig.6 for the CVD dataset the PSO algorithm by using the RF classifier achieves the highest accuracy of 94.12% and similarly in Fig.7 for the Framingham dataset the CSA algorithm by using the RF classifier achieves the highest accuracy of 98.47%.

Table 5 Performance metrics of different classifiers using CVD Dataset

Meta Heuristic Algorithm	ML Classifiers	Accuracy	Precision	Recall	F1-score
CSA	RF	93.28	93.66	94.58	94.12
	DT	91.88	93.94	91.63	92.77
	KNN	88.80	90.55	89.66	90.10
	SVM	83.75	88.77	81.77	85.13
	Naïve Bayes	87.67	89.34	86.70	88.00
GA	RF	92.44	93.80	92.37	93.08
	DT	88.24	92.56	85.50	88.89
	KNN	85.71	88.19	85.50	86.82
	SVM	84.03	86.61	83.97	85.27
	Naïve Bayes	82.35	86.78	80.15	83.33
PSO	**RF**	**94.12**	**94.66**	**94.66**	**94.50**
	DT	89.08	90.70	89.31	90.00
	KNN	87.82	89.84	87.89	88.80
	SVM	86.13	87.69	87.02	87.36
	Naïve Bayes	84.87	85.71	87.02	86.36

The best results are in bold

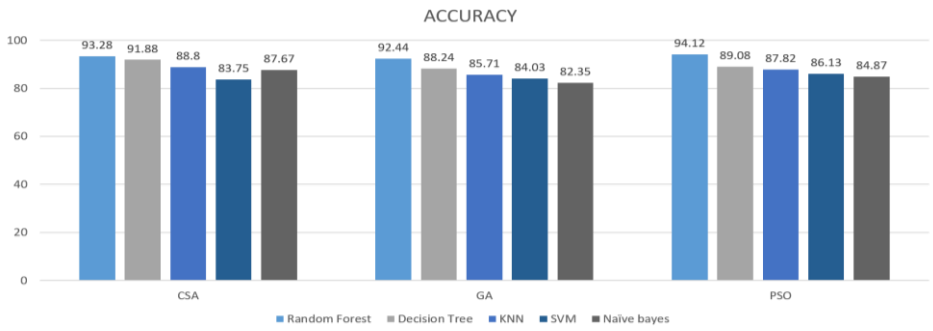

Fig. 6 Accuracy graph of different classifiers using CVD Dataset

Table 6 Performance metrics of different classifiers using Framingham Dataset

Meta Heuristic Algorithm	ML Classifiers	Accuracy	Precision	Recall	F1-score
CSA	**RF**	**98.47**	**98.57**	**98.29**	**98.42**
	DT	91.84	86.21	98.24	92.27
	KNN	79.56	73.32	91.68	81.48
	SVM	69.18	66.78	73.91	70.17
	Naïve Bayes	67.66	66.57	68.34	67.44
GA	RF	90.65	88.34	90.48	90.98
	DT	81.94	78.40	84.24	81.63
	KNN	80.41	72.78	95.37	82.60
	SVM	67.71	64.59	71.23	67.39
	Naïve Bayes	61.80	67.54	38.43	49.15
PSO	RF	89.66	89.28	89.86	89.66
	DT	79.81	77.20	80.48	79.54
	KNN	83.53	75.50	96.30	85.21
	SVM	67.13	65.31	71.24	68.96
	Naïve Bayes	61.38	63.20	47.84	54.48

The best results are in bold

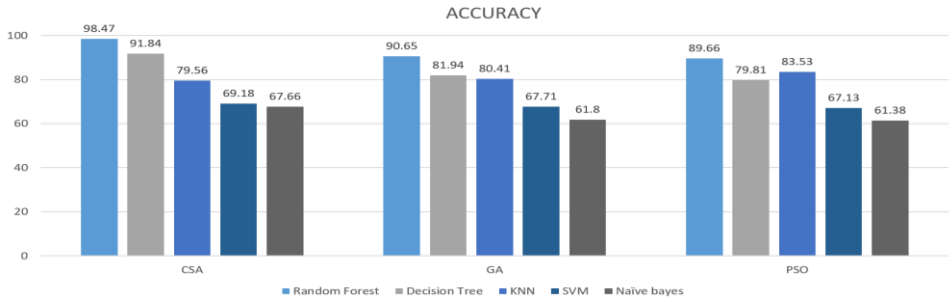

Fig. 7 Accuracy graph of different classifiers using Framingham Dataset

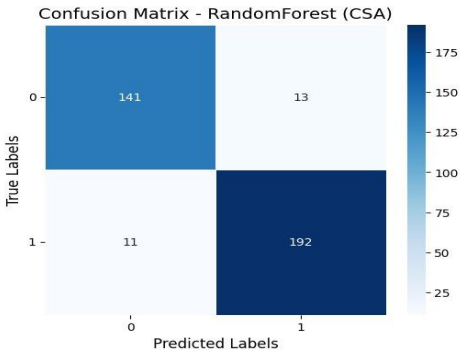

Fig. 8 (a)

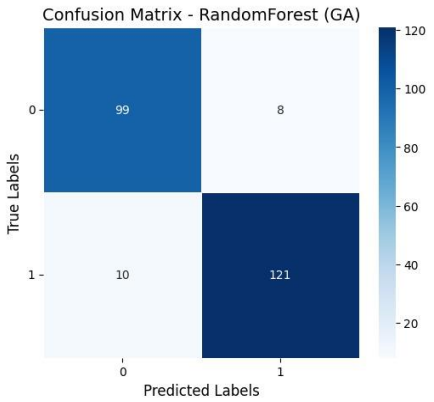

Fig. 8 (b)

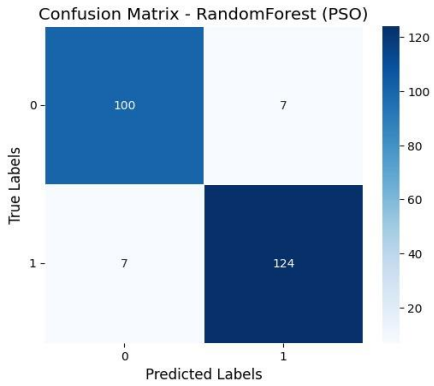

Fig. 8 (c)

Fig. 8 (a) represents a confusion matrix for RF classifier using CSA technique for CVD dataset. Fig. 8 (b) represents a matrix for RF classifiers using GA technique for CVD dataset. Fig. 8 (c) represents the matrix for RF classifier using PSO technique for CVD dataset.

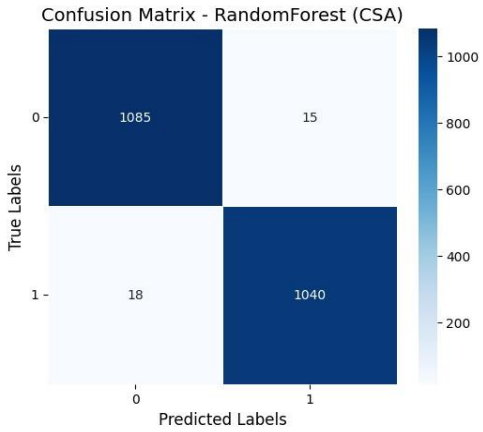

Fig. 9 (a)

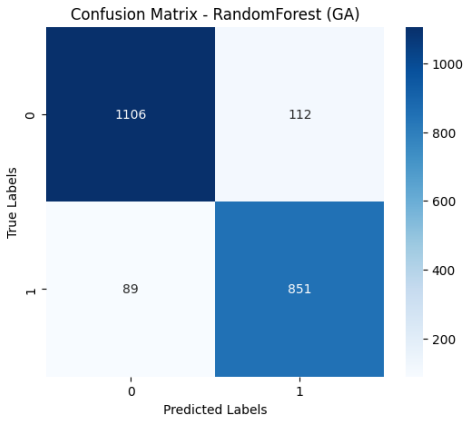

Fig. 9 (b)

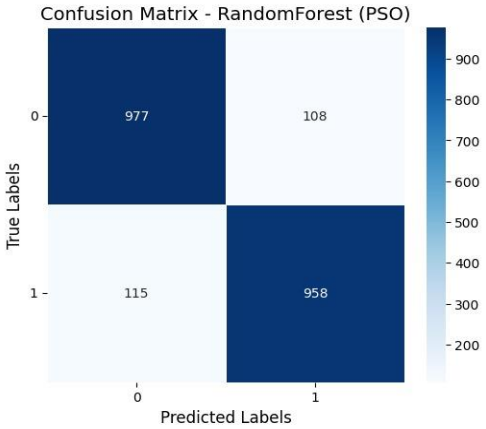

Fig. 9 (c)

Fig. 9 (a) represents the matrix for RF classifier using CSA technique for Framingham dataset. Fig. 9 (b) represents the matrix for RF classifier using GA technique for Framingham dataset. Fig. 9 (c) represents the matrix for RF classifier using PSO technique for Framingham dataset.

Fig. 10 (a) shows how training accuracy and loss evolve over 50 iterations using Particle Swarm Optimization (PSO). Accuracy (green line) steadily increases, while loss (red line) consistently decreases, indicating that the model is learning and improving its performance over time.

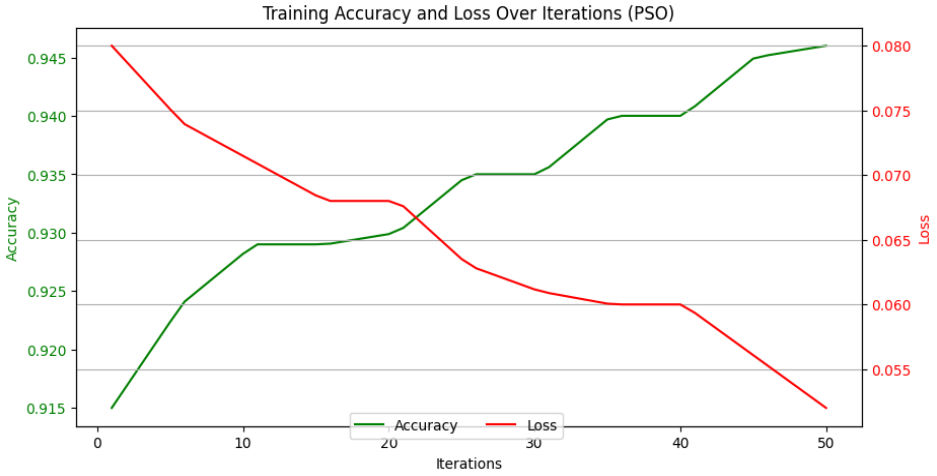

Fig. 10 (a) Training accuracy and loss for PSO on CVD

Fig. 10 (b) shows how training accuracy and loss evolve over 50 iterations using Cuckoo Search Algorithm (CSA). Accuracy (green line) steadily increases, while loss (red line) consistently decreases, indicating that the model is learning and improving its performance over time.

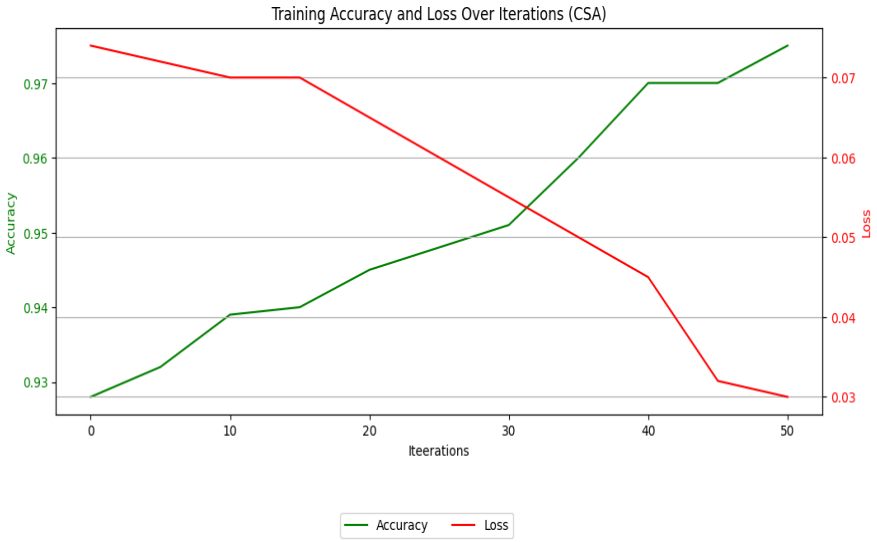

Fig. 10 (b) Training accuracy and loss for CSA on Framingham

Fig. 11 (a) represents ROC curve which compares five classifiers for CVD dataset. The Random Forest model (AUC = 0.98) outperforms the others, while Decision Tree, KNN, SVM, and Naive Bayes have similar but slightly lower AUC values (~0.93–0.94), indicating strong but slightly less accurate performance.

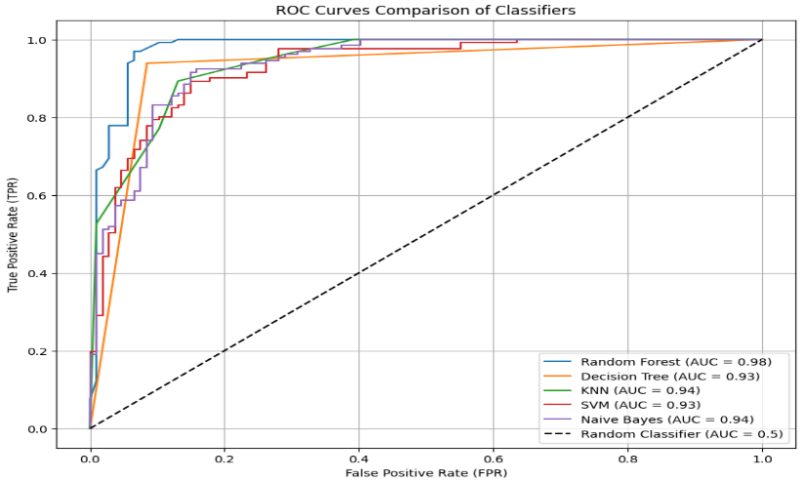

Fig. 11 (a) ROC Curve for all classifiers using PSO for CVD

Fig 11 (b) represents ROC curve which compares several classifiers for Framingham dataset, showing that Random Forest (AUC = 0.99) performs the best, followed by Decision Tree (AUC = 0.90) and KNN (AUC = 0.85). SVM and Naive Bayes perform comparatively worse with lower AUC values (around 0.72), indicating less effective classification.

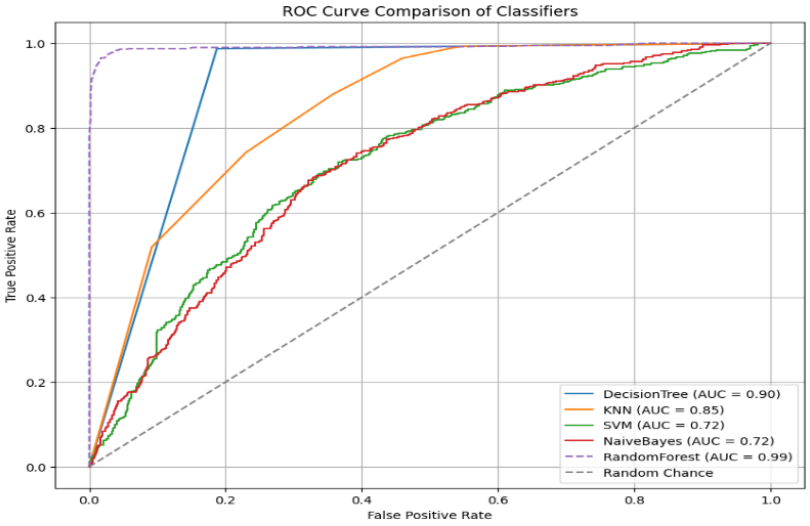

Fig. 11 (b) ROC Curve for all classifiers using CSA for Framingham

Table 7 represents the predicted output after applying the fuzzy inference rules and Table 8 represents the comparison of our work with various papers and shows that our work performs better and outperforms the previous models.

Table 7 Predicted output after applying Fuzzy if-then rules

Features								Predicted Output
Gender	Age	Education	cigs Per Day	BPMeds	Prevalent Stroke	Prevalent Hyp	Glucose	
1	39	4	0	0	0	0	77	Healthy
0	46	2	0	0	0	0	76	Healthy
1	70	1	30	1	1	1	250	Disease
0	61	3	30	0	0	1	103	Healthy
0	38	2	20	0	0	1	70	Disease

Table 8 Comparison of different techniques based on accuracy

Author	Classifiers	Accuracy (%)
Tiwari et al. [3]	RF+ET+XGB+GB	92.34
Ghosh et al. [4]	Random forest	92.00
Khan et al. [5]	Random forest	85.01
Nandakumar et al. [8]	DBN+CSA	91.27
Doppala et al. [10]	GA+RBF	94.20
Proposed Model	**CSA+RF**	**98.47**

7. Conclusion and Future Work

CVD is the most fatal condition that is rapidly expanding and is currently one of the main causes of death globally. If appropriate treatment measures are implemented in the early phases of the disease, the damage it causes can be greatly decreased. The study of metaheuristic algorithms presents a fantastic chance for models to identify illnesses, benefiting patients. In our work, several meta-heuristic techniques are employed to find out and categorize a range of health-related conditions according to their optimal outcomes, for feature selection. Meta-heuristic algorithms are computationally efficient compared to other approaches, making them suitable for critical medical applications. These methods generally offer approximations, direct the searching method, explore the searching area effectively to identify optimal solutions, and are versatile across different problem domains. Here, two different datasets are used, CVD and Framingham, and by using various meta-heuristic algorithms for feature selection various classifiers are implemented to determine the model's optimal performance. Among all the classifiers Random Forest gives the highest accuracy of 94.12% for CVD for PSO, 98.47% for Framingham dataset for CSA and outperforms other ML classifiers for both the datasets. We have also applied fuzzy logic rules on our best performing model for better prediction of CVD. For our future work, we would like to implement deep learning classifiers for performance evaluation of CVD prediction. We would also like to use an ensemble model for feature selection and compare it with our results. The limitations of the paper are no deep learning models are used, lack of real time dataset, fixed fuzzy rule set and computational cost of meta heuristic approaches.

REFERENCES

[1] H. V. Bhagat and M. Singh, "A Machine Learning Model for the Early Prediction of Cardiovascular Disease in Patients", In Proceedings of the 2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC), 2023, pp. 1–5.

[2] G. S. Raksha, R. Hegde, M. N. Shivani, P. S. Shrinidhi, M. T. Monnappa and S. M. Soumyasri, "A Novel Technique for Prediction of Cardiovascular Disease", In Proceedings of the 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022, pp. 1–5.

[3] A. Tiwari, A. Chugh and A. Sharma, "A Ensemble Framework for Cardiovascular Disease Prediction", *Comput. Biol. Med.*, vol. 146, p. 105624, 2022.

[4] S. Ghosh and M. A. Islam, "Performance Evaluation and Comparison of Heart Disease Prediction Using Machine Learning Methods With Elastic Net Feature Selection", *Am. J. Appl. Math. Stat.*, vol. 11, no. 2, pp. 35–49, 2023.

[5] A. Khan, M. Qureshi, M. Daniyal and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction", *Health Soc. Care Community*, vol. 2023, p. 1406060, 2023.

[6] N. M. Lutimath, H. V. Ramachandra, S. Raghav and N. Sharma, "Prediction of Heart Disease Using Genetic Algorithm", In Proceeding of the 2nd Doctoral Symp. Comput. Intell. (DoSCI), Springer Singapore, 2022, pp. 49–58.

[7] T. K. Tanmay, "FRBF: A Fuzzy Rule-Based Framework for Heart Disease Diagnosis"," *Inteligencia Artificial*, vol. 25, no. 69, pp. 122–138, 2022.

[8] P. Nandakumar and S. Narayan, "Cardiac Disease Detection Using Cuckoo Search Enabled Deep Belief Network", *Intell. Syst. Appl.*, vol. 16, p. 200131, 2022.

[9] M. S. Pathan, A. Nag, M. M. Pathan and S. Dev, "Analyzing the Impact of Feature Selection on the Accuracy of Heart Disease Prediction", *Healthcare Anal.*, vol. 2, p. 100060, 2022.

[10] B. P. Doppala, D. Bhattacharyya, M. Chakkravarthy and T. H. Kim, "A Hybrid Machine Learning Approach To Identify Coronary Diseases Using Feature Selection Mechanism on Heart Disease Dataset", *Distrib. Parallel Databases*, vol. 41, pp. 1–20, 2023.

[11] 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. Inform.*, vol. 5, no. 4, pp. 250–260, 2023.

[12] D. Shah, S. Patel and S. K. Bharti, "Heart Disease Prediction Using Machine Learning Techniques", *SN Comput. Sci.*, vol. 1, no. 6, p. 345, 2020.

[13] S. Kaur, Y. Kumar, A. Koul and S. K. Kamboj, "A Systematic Review on Metaheuristic Optimization Techniques for Feature Selections in Disease Diagnosis: Open Issues and Challenges", *Arch. Comput. Methods Eng.*, vol. 30, no. 3, pp. 1863–1895, 2023.

[14] K. Kanagarathinam, D. Sankaran and R. Manikandan, "Machine Learning-Based Risk Prediction Model for Cardiovascular Disease Using a Hybrid Dataset", *Data Knowl. Eng.*, vol. 140, p. 102042, 2022.

[15] N. A. Baghdadi, S. M. F. Abdelaliem, A. Malki, I. Gad, A. Ewis and E. Atlam, "Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis", *J. Big Data*, vol. 10, no. 1, p. 144, 2023.

[16] H. H. Alalawi and M. S. Alsuwat, "Detection of Cardiovascular Disease Using Machine Learning Classification Models", *Int. J. Eng. Res. Technol.*, vol. 10, no. 7, pp. 151–157, 2021.

[17] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar and T. M. Nithya, "A Cardiovascular Disease Prediction Using Machine Learning Algorithms", *Ann. Romanian Soc. Cell Biol.*, vol. 25, no. 2, pp. 904–912, 2021.

[18] J. C. T. Arroyo and A. J. P. Delima, "An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction", *J. Adv. Inf. Technol.*, vol. 13, no. 1, pp. 95–99, 2022.

[19] T. Vivekanandan and N. C. S. N. Iyengar, "Optimal Feature Selection Using a Modified Differential Evolution Algorithm and Its Effectiveness for Prediction of Heart Disease", *Comput. Biol. Med.*, vol. 90, pp. 125–136, 2017.

[20] T. Kasbe and R. S. Pippal, "Design of Heart Disease Diagnosis System Using Fuzzy Logic"," In Proceedings of the 2017 Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS), 2017, pp. 3183–3187.

[21] N. V. MahaLakshmi and R. K. Rout, "An Intelligence Method for Heart Disease Prediction Using Integrated Filter-Evolutionary Search Based Feature Selection and Optimized Ensemble Classifier", *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 39841–39865, 2024.

[22] J. Priyadarshini, M. Premalatha, R. Čep, M. Jayasudha and K. Kalita, "Analyzing Physics-Inspired Metaheuristic Algorithms in Feature Selection With K-Nearest-Neighbor", *Appl. Sci.*, vol. 13, no. 2, p. 906, 2023.

[23] S. K. Sharma, L. Goel and N. Mittal, "Nature-Inspired Optimization Techniques for Cardiovascular Disease Detection: A Comprehensive Survey", *Neural Comput. Appl.*, vol. 37, pp. 1839–1874, 2024.