Indian Journal of Natural Sciences *www.tnsroindia.org.in ©IJONS*

Vol.15 / Issue 86/ Oct / 2024 *International Bimonthly (Print) – Open Access* *ISSN: 0976 – 0997*

RESEARCH ARTICLE

Optimal Feature Selection using Integration of Firefly Optimization with Random Forest Method

G.Angayarkanni[1]* and M.Rajasenathipathi[2]

[1]Ph.D Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, (Affiliated to Bharathiar University), Coimbatore, Tamil Nadu, India.
[2]Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, (Affiliated to Bharathiar University), Coimbatore, Tamil Nadu, India.

Received: 21 Jun 2024 Revised: 03 Jul 2024 Accepted: 14 Aug 2024

***Address for Correspondence**
G.Angayarkanni
Ph.D Scholar, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi,
(Affiliated to Bharathiar University),
Coimbatore, Tamil Nadu, India.
Email: g.angayarkanni@gmail.com

ABSTRACT

Heart disease, a broad term encompassing various cardiovascular conditions, stands as a formidable global health challenge, representing a leading cause of morbidity and mortality. To increase the efficacy of disease prediction earlier, it is necessary to identify the most relevant subset of features in a given domain. In this paper, the integration of the Firefly Optimization Algorithm (FOA) with Random Forest (RF) is proposed and used to select the optimal feature subset for prediction of Heart disease. The hybridization makes use of the advantages of both FOA and RF methods. Initiating with the data pre-processing step by employing a matching pursuit imputation method to handle missing values, followed by Z-score normalization for feature scaling, and Bray–Curtis feature learning vector quantization is then employed for feature extraction to reduce dataset dimensionality. After that Optimal feature selection is carried out using the Firefly Optimization Algorithm with Random Forest (FOA-RF) method, which minimizes both time and space complexity in disease prediction and then applies the proposed algorithm in five different heart disease datasets to select the best features. The effectiveness of the selected feature set is analyzed using a Support Vector Machine and the experimental results confirm the efficiency of the proposed feature selection approaches rather than the original Firefly Optimization and Random Forest algorithm by way of searching the feature space and selecting the most informative attributes for prediction tasks. Accuracy, Sensitivity, and Specificity have been measured to evaluate the results. Our experimental result demonstrates that the SVM-based model with the FOA-RF achieves a better result than the other feature selector algorithms.

Keywords: Firefly Optimization Algorithm, Random Forest, Support Vector Machine, Heart disease, feature selection, vector quantization, Z-score normalization

INTRODUCTION

Heart Disease is a severe condition that significantly impacts human life, emerging as one of the leading causes of death worldwide. To prevent further damage to patients, accurate diagnosis and early identification of heart disease are essential for effective rehabilitation and treatment. A machine learning model has been trained on medical data to enable efficient heart disease diagnosis, conserving resources while improving accuracy. During the training process, the medical datasets contain both relevant and redundant features about the patients. These unnecessary features do not contribute meaningful information to the disease detection task and lead to the curse of dimensionality. Therefore, significant feature selection techniques are required in heart disease prediction. Feature selection is a crucial step in machine learning and data analysis, particularly in the domain of medical diagnosis such as predicting heart disease. It involves identifying the most relevant features or variables from a dataset that contribute the most to the predictive performance of a model. By selecting the right features, we can improve model accuracy, reduce overfitting, and enhance interpretability. In heart disease prediction, the significance of feature selection cannot be overstated. With a multitude of potential risk factors and diagnostic indicators, identifying the most influential features can lead to more accurate and efficient predictive models. By focusing on relevant features, we can streamline the diagnostic process, potentially leading to earlier detection and intervention, ultimately saving lives.

The integration of Firefly Optimization (FO) and Random Forest (RF) algorithm for feature selection presents a promising approach to addressing the complexity of feature selection problems. Firefly Optimization is a metaheuristic algorithm inspired by the flashing behavior of fireflies, which seeks to optimize solutions in a search space. When combined with Random Forest, a powerful ensemble learning technique known for its robustness and flexibility, this hybrid approach aims to leverage the strengths of both algorithms for more effective feature selection. Several feature selection algorithms exist, each with its strengths and weaknesses. Commonly used methods include Recursive Feature Elimination (RFE), Genetic Algorithms (GA), and Principal Component Analysis (PCA). When compared to these methods, the FO-RF hybrid approach offers distinct advantages. While RFE may struggle with large feature sets and GA can be computationally expensive, the FO-RF hybrid method aims to strike a balance between efficiency and effectiveness. Additionally, the ensemble nature of Random Forest helps mitigate the risk of overfitting, which can be a concern with some feature selection techniques. The hybridization of Firefly Optimization and Random Forest brings several benefits. Firstly, Firefly Optimization enhances the search process by efficiently exploring the feature space, potentially leading to more optimal solutions. Secondly, Random Forest provides robustness and generalization ability, ensuring that the selected features are relevant across different datasets and scenarios. By combining these strengths, the hybrid approach aims to improve feature selection performance, resulting in more accurate and interpretable predictive models for heart disease diagnosis. The subsequent sections of this paper are organized as follows: Section II discusses the related research on feature selection techniques for heart disease prediction and highlights the limitations. Section III briefly explains FO and RF algorithms, and then details how they are integrated for feature selection in the proposed FO-RF approach. Emphasize how FO addresses redundancy issues and how RF contributes to robustness and generalizability. Section IV describes the experimental setup, including data selection, evaluation metrics, and comparison with existing methods. This section should demonstrate the effectiveness of the FO-RF approach in improving model performance. Section V summarizes the findings, emphasizing how the proposed method addresses the initial problem of redundant features and contributes to more accurate heart disease prediction models.

LITERATURE REVIEW

In recent studies aiming to improve heart disease diagnosis accuracy, various feature selection techniques have been employed. Zheng et al. [1] introduce the MPMDIWOA algorithm, which merges the MPMD filter algorithm with the IWOA wrapper algorithm, effectively addressing local optimal values through concepts like maximum value without change (MVWC) and thresholds, resulting in superior classification accuracy and feature subsets. Tubishat et al. [2] present the Dynamic Butterfly Optimization Algorithm (DBOA), an enhanced version of BOA, which resolves

limitations by enhancing solution diversity and mitigating local optima issues. Tested on 20 benchmark datasets, DBOA outperforms other optimization algorithms. Abdel-Basset et al. [3] propose a novel approach by integrating the grey wolf optimizer algorithm with a two-phase mutation strategy, improving feature selection efficiency. The study contributes to optimization methods by leveraging metaheuristic algorithms. Abdelhamid et al. [4] introduce bSCWDTO, a hybrid binary meta-heuristic algorithm for feature selection, showing superiority over existing methods on 30 datasets. A hybrid approach [5] utilizing LGP and BA generates candidate chromosomes and performs neighborhood search, followed by SVM classification, yielding promising results. Another study [6] focuses on a three-step method involving feature selection, clustering, and classification, achieving superior results through various hybrid optimization algorithms. Xiang *et al.* [7] propose a hybrid system merging IGSA with k-NN, enhancing IGSA with PWL and SQP, and extending it to binary space, resulting in superior computational performance. Three hybrid algorithms [8] integrating SOA and TEO are introduced, showing promising results on 20 benchmark datasets. Nemati et al. [9] propose a hybrid feature selection algorithm merging GA and ACO, with low computational complexity and competitive performance. Xi e *et al.* [10] propose IFSFS, a hybrid method merging filter and wrapper methods, yielding optimal feature subsets from the original set. Another framework [11] integrates SACI with SVM, showing promising results in feature selection and model selection. A hybrid FFPSO technique [12] for BBB detection combines Firefly and Particle Swarm Optimization, enhancing classifier performance through optimized features. TRSFFQR [13] combines TRS and FA for feature selection in MRI brain images, addressing the limitations of basic models. HGAWE [14] combines a genetic algorithm with wrapper-embedded feature selection, outperforming existing methods in feature selection and classification accuracy. FCBF is used in conjunction with PSO and recursive FA [15], demonstrating robustness in classification accuracy. In a study on CAD [16], seven feature selection techniques are explored, achieving a classification accuracy of 88.15%.Kabir et al. [17] propose a hybrid ACO algorithm that efficiently explores feature spaces.Yu et al. [18] integrate PSO and GA for feature selection, offering a promising solution for enhancing machine learning model performance, particularly in cyber security applications. A proposed method [19] combines PSO exploration with GWO exploitation, yielding superior performance in feature selection and classification tasks.

DATASET

To evaluate the performance of various feature selection methods, we utilized five benchmark datasets from the Kaggle repositories focusing on cardiovascular health. These datasets serve as representative samples of real-world clinical data collected for heart disease prediction. Table 1 summarizes the datasets, including the number of attributes (features) and instances (data points) in each.

DATA PREPROCESSING

Data preprocessing plays a crucial role in ensuring the reliability and generalizability of machine learning models, particularly in the domain of heart disease prediction. Inconsistent data formats, missing values, and outliers can significantly impact model performance. To address these challenges, a comprehensive preprocessing pipeline is essential.

Missing Value Imputation

Missing entries within the datasets require careful handling to minimize their influence on model performance. We employ Matching Pursuit imputation, a machine learning-based approach for robustly imputing missing values. This method leverages the existing data to estimate the conditional statistical mean for each feature (column) with missing values. Here's the formula for calculating the mean:

$$\mu\left(\beta_{fv_i}\right) = \frac{\sum \beta_{fv_i}}{n} \quad (1)$$

Following the mean calculation, Matching Pursuit identifies the optimal value to impute by minimizing the absolute difference between the estimated value and its neighboring entries in the same feature. Essentially, the method seeks the value that creates the least disruption to the existing data pattern within the feature. The formula for absolute difference minimization can be represented as:

$$F = min\left|\beta_{fvD} - \beta_{fv(N)}\right| \quad (2)$$

By minimizing this absolute difference (often referred to as Least Absolute Deviation), Matching Pursuit aims to impute a value that seamlessly integrates with the surrounding data points within the feature.

Feature Scaling

Feature scaling is crucial for ensuring that all features contribute equally to the model's learning process. This is particularly important when dealing with datasets containing features measured in different units. We utilize standardized Z-score normalization for feature scaling. This technique transforms each feature value by subtracting the mean of the feature and then dividing it by its standard deviation. This results in a standardized dataset with a mean of 0 and a standard deviation of 1.

Categorical Feature Encoding

Many real-world datasets, including those used for heart disease prediction, contain categorical features. These features need to be converted into a numerical format suitable for machine learning algorithms. One-hot encoding is employed for this purpose. This technique creates a new binary feature for each unique category within the original categorical feature. The value of each new feature is set to 1 for the corresponding category and 0 for all other categories. For instance, a categorical feature "gender" with values "male" and "female" would be transformed into two new binary features: "gender_male" (1 for male, 0 for female) and "gender_female" (0 for male, 1 for female).

FEATURE EXTRACTION

Following data preprocessing the Bray–Curtis feature learning vector quantization is utilized for feature extraction, aiming to reduce the dataset's dimensionality. This process enhances computational efficiency and mitigates the curse of dimensionality, transforming the original dataset into a lower-dimensional representation for modeling. Feature learning vector quantization employs winner-take-all training algorithms, identifying significant features based on the Bray–Curtis index and mapping input features to a smaller set through quantization. For each feature in the vector 'β_{fi}' and β_{fj}, find the closet features using the Bray-Curtis Similarity Index.

$$\omega_{ij} = 1 - \frac{2\,(M_{ij})}{|\beta_{fi}| + |\beta_{fj}|} \quad (3)$$

In (3), ω_{ij} denotes a Bray-Curtis Similarity coefficient, M_{ij} denotes a mutual dependence between the two features, $|\beta_{fi}|$ and $|\beta_{fj}|$ represents the cardinalities of the two sets (i.e. number of values in each feature)**.**The Bray–Curtis similarity coefficient is bounded between 0 and 1. The winner-take-all training algorithm is exploited in learning vector quantization to identify winning features in the vector (i.e. high similarity).

$$H = \begin{cases} 1, & winning\ features\ in\ vector \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

FEATURE SELECTION

After the feature extraction, feature selection is performed to choose a subset of optimal relevant features from extracted features. The main aim of feature selection is to improve model performance and reduce error, by focusing on the most informative and discriminative features. Therefore, the integration ofthe firefly optimization algorithm with random forest is introduced for optimal feature selection.

Proposed Hybrid FOA-RF Algorithm

Firefly Optimization [20] is a metaheuristic algorithm inspired by the flashing light behavior of fireflies. Fireflies are generally unisexual. Each firefly is attracted to others based on its light intensity, as the attractiveness of a firefly is directly proportional to its brightness. Fireflies with lower intensity are attracted to other fireflies that emit brighter light. Here the firefly is related to the number of features in the given dataset. The population offireflies (i.e., features) $\beta_{fi} = [\beta_{fv_1}, \beta_{fv_2}, \cdots, \beta_{fv_n}]$is initialized in the search space. For every firefly, the fitness value is calculated based on the objective function.

$$f(x) = arg\,max\,Acc\ (DD) \quad (5)$$

Where $f(x)$ denotes a fitness function, $arg\,max$ denotes an argument of the maximum function, $Acc(DD)$ denotes an accuracy of the disease diagnosis (i.e. target output). The attractiveness of a firefly is determined by its fitness

value. Fireflies are attracted to other fireflies with higher fitness values and move towards them. In other words, the firefly 'ff_i'with less fitness is moved towards to fireflies with high fitness. Otherwise, the fireflies move in a random position. The position of the firefly is updated as follows,

$$X_{t+1} = X_i^{\,t} + a_{ij}e^{[-\rho D_{ij}{}^2]}\left(X_i^{\,t} - X_j^{\,t}\right) + q_t\epsilon_t \quad (6)$$

From (11), X_{t+1} represents an updated light intensity of the firefly,$X_i^{\,t}$ denotes a current light intensity of firefly 'i', $X_j^{\,t}$ denotes a current light intensity of firefly 'j', a_{ij} attractiveness of the firefly, q_t denotes a parameter controlling step size and ϵ_t represents a vector drawn from a Gaussian or other distribution, 'ρ' denotes a light absorption coefficient, D_{ij} distance between the fireflies 'i' and 'j'. Then the fitness is again computed based on the updated position of the firefly. Followed by, the fireflies are ranked to determine the more optimal solution with the help of a random Forest decision tree. Random Forest is an ensemble learning algorithm that utilizes multiple decision trees to find more optimal features. The proposed ensemble learning technique first constructs the set of the weak learner as a decision stump to find the optimal features through the ranking process. The Random Forest ensemble technique utilizes a multi-iterative decision tree for weak learners and gives the input of fireflies with the fitness value. A decision stump is a basic decision tree comprising a root node directly linked to a leaf node, employing a decision rule. The root node performs decision-making by establishing the decision rule. Based on the decision-making process, optimal features are identified, and results are obtained at the leaf nodes. The root node makes a rule as given below

$$R = if\ (f_i(x) > (f_j(x)) \quad where\ i = 1,\quad j = 2,3,4,5\ldots.n \quad (7)$$

Where R denotes a rule, $f_i(x)$ denotes the fitness of one firefly, $f_j(x)$ denotes the fitness of other fireflies. If the fitness of the 'i^{th}' firefly is greater than the fitness of the other 'j^{th}' fireflyf_j, then the root tree node rank the features. Followed by, other features are ranked.

$$Q = \sum_{i=1}^{n} W_i \quad (8)$$

From (13), Qdenotes strong output results by combining all the weak learner results 'W_i'. Finally, the preferential voting scheme is applied to find the final optimal results.

$$Z = arg\ max\,\vartheta(W(ff_i)) \quad (9)$$

From (14), 'Z' denotes the final random forest decision tree result. '$arg\ max\,\vartheta(W(ff_i))$' represents the majority votes of weak learner results are obtained as a final result. In this way, optimal features are selected for disease diagnosis to minimize complexity and improve accuracy.

// Algorithm Integration of Firefly optimization with Random Forest for feature selection
Input: Number of fireflies (i.e. number of features) $\beta_{fi} = \beta_{fv_1}, \beta_{fv_2}, \ldots, \beta_{fv_n}$,
Output: Optimal feature selection
Begin Generate an initial population of fireflies i.e. number of features $\beta_{fi} = \beta_{fv_1}, \beta_{fv_2}, \ldots, \beta_{fv_n}$ **For each** firefly Compute fitness '$f(x)$' **while**(t < $terminationismet$) **for**i = 1: n (all n fireflies) **for**j = 1: n (all n fireflies) **if** ($f(X_i)$ < $f(X_j)$)**then** Move firefly ff_itowards firefly ff_j Evaluate new solutions and update light intensity using (11) **End if** **end for*j*** **end for*i*** **Go to step 3** **for each**firefly with the fitness value Construct a set of weak classifiers **End for** $\boldsymbol{if}\ (f_i(x) > f_j(x))\ \boldsymbol{then}$ Rank the features with a high position

```
        End if
        Combine all weak classifiers
         for each weak classifier
        Assign the voting scheme
        Find weak learner resultswith majority votes
              Return (optimal k features)
         End for
        End while
    End
```

The algorithm describes an approach for optimal feature selection by integrating the firefly algorithm with random forest to improve the accuracy of disease diagnosis and minimize the time. The algorithm begins by randomly initializing the population of fireflies, which corresponds to the number of features. For each feature, fitness is measured, and the light intensity of the firefly is determined based on the fitness function. If the light intensity of one firefly is higher than that of another, it moves and is attracted to the other firefly. After the movement of fireflies, the light intensity of all fireflies is updated. Then, the fitness is re-evaluated based on the updated light intensities of the fireflies. To rank the features, a random forest decision tree is constructed. Initially, a set of weak learners is created based on the fitness of each firefly. The root node determines a firefly with a higher fitness is ranked first, followed by the others in ascending order. Finally, the results of the weak learners are combined using a preferential voting scheme to create a strong output. The majority votes of the weak learner results determine the final output. This process is repeated until the algorithm reaches the maximum number of iterations. Finally, the optimal set of features gets selected for accurate disease diagnosis.

EXPERIMENTS AND RESULTS

This research investigates the predictive effectiveness of Random Forest (RF), Firefly Optimization Algorithm (FOA), and their combination FOA-RF, across five datasets: Framingham, Heart Disease, Cardiovascular Disease Dataset, Cardio train, and Heart Attack, procured from Kaggle repository. Under the experiment's framework, each dataset undergoes particular preprocessing and feature extraction, followed by model training and evaluation using an 80-20 training-testing split. Performance metrics encompassing accuracy, specificity, and sensitivity are precisely recorded. The results indicated that FOA-RF consistently outperformed RF and FOA across all datasets, demonstrating superior feature selection capabilities and classification accuracy. Statistical analysis confirmed the significance of the performance difference between FOA-RF and the other algorithms. Overall, the integration of Firefly Optimization with Random Forest emerged as a promising approach for accurately diagnosing cardiovascular diseases, leveraging the complementary strengths of both algorithms.

Accuracy

In the prediction of heart disease, accuracy is a crucial performance metric that measures the proportion of correctly classified instances among all instances evaluated. It provides an overall assessment of how well a predictive model performs in correctly identifying individuals with or without heart disease. The formula for accuracy is:
Accuracy = (True Positives + True Negatives) / Total Instances

Sensitivity

Sensitivity, also known as recall, is a vital metric in feature selection for heart disease prediction. It measures the proportion of true positive cases correctly identified by the model. A high sensitivity indicates effective identification of individuals with the disease, minimizing false negatives. In our comparison of Random Forest (RF), Firefly Optimization Algorithm (FOA), and the integrated FOA-RF for feature selection, FOA-RF demonstrated superior sensitivity. Leveraging both FOA's efficient exploration of the search space and RF's robust classification, FOA-RF optimizes feature selection to maximize sensitivity, enhancing the accuracy of heart disease prediction models. The

Formula for Sensitivity
Sensitivity = True Positives / (True Positives + False Negatives)

Specificity
Specificity, along with sensitivity (recall), is a vital metric for evaluating the effectiveness of a model in disease prediction. It measures the proportion of true negative cases that the model correctly identifies. In heart disease prediction, specificity signifies the model's ability to accurately classify individuals who do not have heart disease.

The formula for Specificity:
Specificity = True Negatives / (True Negatives + False Positives)

CONCLUSION

This study investigated FOA-RF, a novel approach combining Random Forest and Firefly Optimization Algorithm, for selecting optimal features in heart disease prediction was conducted on five heart disease datasets from Kaggle: Framingham, Heart Disease, Cardiovascular Disease Dataset, Cardio train, and Heart Attack. FOA-RF outperformed RF and FOA individually in selecting optimal features for accurate diagnosis. Evaluation metrics such as Accuracy, Sensitivity, and Specificity showed FOA-RF's superior performance over individual algorithms. The hybrid method efficiently searched feature space, selecting relevant attributes for prediction tasks. SVM-based models with FOA-RF feature selection achieved better results than other selectors, streamlining diagnostics and enabling earlier intervention. The integration of Firefly Optimization and Random Forest addresses feature selection complexity in heart disease prediction, aiming to improve performance and model interpretability. Overall, experiments validate FOA-RF's effectiveness in feature selection for heart disease prediction, highlighting its importance in enhancing accuracy and healthcare outcomes.

REFERENCES

1. Zheng, Yuefeng, et al. "A novel hybrid algorithm for feature selection based on whale optimization algorithm." IEEE Access 7 (2018): 14908-14923.
2. Tubishat, Mohammad, et al. "Dynamic butterfly optimization algorithm for feature selection." IEEE Access 8 (2020): 194303-194314.
3. Abdel-Basset, Mohamed, et al. "A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection." Expert Systems with Applications 139 (2020): 112824.
4. Abdelhamid, Abdelaziz A., et al. "Innovative Feature Selection Method Based on Hybrid Sine Cosine and Dipper Throated Optimization Algorithms." IEEE Access (2023).
5. Hasani, Seyed Reza, Zulaiha Ali Othman, and Seyed Mostafa Mousavi Kahaki. "Hybrid feature selection algorithm for the intrusion detection system." Journal of Computer Science 10.6 (2014): 1015.
6. Li, Xiaohua, Jusheng Zhang, and Fatemeh Safara. "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm." Neural processing letters (2021): 1-17.
7. Xiang, Jie, et al. "A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method." Applied Soft Computing 31 (2015): 293-307.
8. Jia, Heming, Zhikai Xing, and Wenlong Song. "A new hybrid seagull optimization algorithm for feature selection." IEEE Access 7 (2019): 49614-49631.
9. Nemati, Shahla, et al. "A novel ACO–GA hybrid algorithm for feature selection in protein function prediction." Expert systems with applications 36.10 (2009): 12086-12094.
10. Xie, Juanying, and Chunxia Wang. "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythematous-squamous diseases." Expert Systems with Applications 38.5 (2011): 5809-5815.

11. Aladeemy, Mohammed, Salih Tutun, and Mohammad T. Khasawneh. "A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence." Expert Systems with Applications 88 (2017): 118-131.
12. Kora, Padmavathi, and K. Sri Rama Krishna. "Hybrid firefly and particle swarm optimization algorithm for the detection of bundle branch block." International Journal of the Cardiovascular Academy 2.1 (2016): 44-48.
13. Jothi, G. "Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification." Applied Soft Computing 46 (2016): 639-651.
14. Liu, Xiao-Ying, et al. "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection." IEEE Access 6 (2018): 22863-22874.
15. Sahu, Bibhuprasad, et al. "A hybrid cancer classification based on SVM optimized by PSO and reverse firefly algorithm." International Journal of Control and Automation 13.4 (2020): 506-517.
16. Angayarkanni, G., and S. Hemalatha. "Selection of features associated with coronary artery diseases (CAD) using feature selection techniques." Journal of Xi'an University of Architecture & Technology (2020): 686-699.
17. Kabir, Md Monirul, Md Shahjahan, and Kazuyuki Murase. "A new hybrid ant colony optimization algorithm for feature selection." Expert Systems with Applications 39.3 (2012): 3747-3763.
18. Yu, Xue, et al. "A hybrid algorithm based on PSO and GA for feature selection." Journal of Cybersecurity 3.2 (2021): 117.
19. Al-Tashi, Qasem, et al. "Binary optimization using hybrid grey wolf optimization for feature selection." IEEE Access 7 (2019): 39496-39508.
20. Emary, Eid, et al. "Firefly optimization algorithm for feature selection." Proceedings of the 7th balkan conference on informatics conference. 2015.

Table 1: Attribute Names

Dataset	Attribute Names	No. of Attributes	No. of Instance
Framingham (Kaggle)	Male, age, education, current Smoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, TenYearCHD	16	4240
Heart Disease (Kaggle)	Gender, age, education, currentSmoker, cigsPerDay, BPMeds , revalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, Heart_ stroke	16	4238
Cardiovascular_Disease_Dataset (Kaggle)	patientid, age, gender, chestpain, restingBP, serumcholestrol, fasting blood sugar, resting relectro, max heart rate, exercise angia, old peak, slope, noofmajorvessels, target	14	1000
Cardio_train (Kaggle)	age, height, weight, gender, ap_hi, ap_lo, cholestrol, gluc, smoke, alco, active, cardio	12	70000
Heart attack (Kaggle)	Heart Diseaseor Attack, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, Diabetes. Phys Activity Fruits, Veggies, HvyAlcoholConsump, Any Healthcare, NoDocbcCost, GenHlth, MentHlth PhysHlth, DiffWalk, Sex, Age, Education, Income	22	253661

Table 2: Comparison of Accuracy

Algorithms / Dataset	All + SVM	Firefly Optimization (FO) + SVM	Random Forest (RF) + SVM	FO-RF + SVM
Framingham	78.4	83.4	84.2	86.0
Heart Disease (Kaggle)	78.2	79.7	85.9	89.0
Cardiovascular_Disease_Dataset	80.5	81.4	81.8	85.5
Cardio_train	79.1	85.9	87.1	89.1
Heart attack	83.3	85.0	86.4	91.3

Table 3: Comparison of Sensitivity

Algorithms / Dataset	All + SVM	Firefly Optimization (FO) + SVM	Random Forest (RF) + SVM	FO-RF + SVM
Framingham	82.6	87.2	90.3	91.6
Heart Disease (Kaggle)	93.0	90.6	92.1	92.3
Cardiovascular_Disease_Dataset	76.3	75.4	77.5	76.4
Cardio_train	81.1	75.9	82.9	82.6
Hheart attack	90.1	91.3	96.1	97.3

Table 4: Comparison of Specificity

Algorithms / Dataset	All + SVM	Firefly Optimization (FO) + SVM	Random Forest (RF) + SVM	FO-RF + SVM
Framingham	89.5	90.9	89.5	91.2
Heart Disease (Kaggle)	84.3	81.0	82.3	85.7
Cardiovascular_Disease_Dataset	76.1	77.3	76.8	79.5
Cardio_train	77.0	78.5	82.4	86.7
Heart attack	94.3	95.5	93.2	96.4

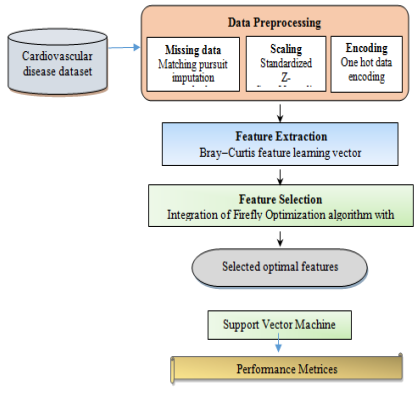

Figure 1: Architecture diagram of Proposed model

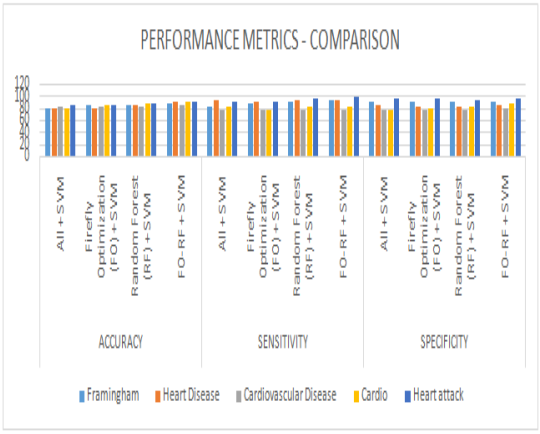

Fig 2: Accuracy, Sensitivity and Specificity Comparison Chart