

CORONARY ARTERY DISEASE PREDICTION BASED ON OPTIMAL FEATURE SELECTION USING IMPROVED ARTIFICIAL NEURAL NETWORK WITH META-HEURISTIC ALGORITHM

D.VETRITHANGAM¹, V. SENTHILKUMAR², NEHA³, A. RAMESH KUMAR⁴, P.NARESH KUMAR⁵, MRADULA SHARMA⁶

¹Associate Professor, Department of Computer Science & Engineering, Chandigarh University, Punjab, India.

²Associate Professor, Department of Mechanical Engineering, SRM TRP Engineering College, Tamilnadu, India.

³Assistant Professor, Department of Computer Science & Engineering, Chandigarh University, Punjab, India.

⁴Professor, Department of Mechatronics Engineering, K.S.Rangasamy College of Technology, Tamil Nadu, India.

⁵Assistant Professor, Department of Computer Science and Engineering, KG Reddy College of Engineering and Technology, Telangana, India.

⁶Assistant Professor, Department of Computer Science and Information Technology, Japjee Institute of Information Technology, Uttarpradesh, India.

E-mail: ¹vetrigold@gmail.com, ²trpvs12@gmail.com, ³neha.arya35@gmail.com,

⁴arameshkumaar@gmail.com, ⁵pnrshkumar@gmail.com, ⁶mradulasharma86@gmail.com

ABSTRACT

Scientific breakthroughs in understanding the etiology of coronary artery disease (CAD) will allow for more accurate coronary artery disease (CAD) diagnosis and treatment techniques. Coronary Artery Disease (CAD) is a type of cardiovascular disease in which atherosclerotic plaques in the coronary arteries cause myocardial infarction or sudden cardiac death. In medicine, disease prediction based on Artificial Neural Networks (ANN) plays a significant role in enhancing the reliability of general population health care. So, our main goal is to propose an improved artificial neural network model in conjunction with a Meta-heuristic algorithm that works with distinct types of CAD datasets with good accuracy. The system selects the most relevant or similar features from the raw dataset; this feature selection is achieved by the Meta-heuristic algorithm. This model uses 18 input nodes, 18 hidden nodes, and 1 output node in an 18-18-1 multilayered feed-forward network architecture, which is the best network for the prediction of CAD with the selected dataset. When using different methodologies on datasets dealing with coronary artery disease (CAD), the results may vary. Efficient medical diagnosis and analysis are important in selecting the important features. The Cleveland Heart Disease dataset, obtained from the UCI repository, was used in this paper; it contains 37079 person data records with 50 attributes. The proposed Improved Artificial Neural Network model with Meta-heuristic Algorithm results in 97.63% Sensitivity, 97.5% Accuracy, and 97.35 % Specificity.

Keywords: *Coronary Artery Disease, Deep Learning, Risk Factor Meta-Heuristic Algorithm And Artificial Neural Networks.*

1. INTRODUCTION

Heart disease (HD) could be an essential term for different types of diseases, factors and abnormalities that affecting the heart and also the blood vessels. The symptoms rely upon the

particular variety of these diseases such as heart failure, hypertensive heart disease, coronary artery diseases, stroke, cardiomyopathy and congenital disease, heart arrhythmia etc. According to the World Health Organization, cardiovascular disease (CVDs) lead to thirty one percentages of all global deaths, with the number of deaths from CVDs

expected to reach thirty million by 2030. Cardiovascular diseases (CVDs) are a class of factors that affect the heart and blood vessels. Cerebrovascular disease, myocardial infarction (MI), coronary artery disease (CAD), rheumatic heart disease, peripheral vascular disease (PVD), stroke, and other conditions are included, but the most prevalent kind of cardiac disease worldwide is coronary artery disease. It is an issue brought on by fat buildup in the blood vessels and veins. Additionally, it hinders blood flow into the heart's veins and vessels, which results in insufficient blood and oxygen reaching the organ's interior. Angina, also known as Angina Pectoris, is a medical term used to describe cardiac pain brought on by insufficient blood flow to the heart. It is the preventative sign of receiving treatment for cardiac issues. This kind of discomfort can last for a few seconds or minutes. The illness known as congestive heart failure is characterized by the heart's inability to adequately pump blood to the body's other organs. The etiology of CAD or atherosclerosis has not been fully understood, besides familial and Meta-heuristic factors, acquired risk factors such as hypertension, cigarette smoking, high-density lipoprotein (HDL), left ventricular hypertrophy, hyperlipidemia, blood pressure, blood cholesterol and diabetes mellitus are associated with this disease. The prevalence of diabetes is increasing globally, and it has reached pandemic levels worldwide. In accordance with Fisher, post-menopausal women, men older than forty five and the person with obesity have a risk factor. At the first plaque grows inside the coronary artery wall till the blood flow to the muscle of the heart is inhibited. The abnormal narrowing of the heart vessels is diagnosed by angiography, which costs high and has more complicated procedures, so researchers are trying to find alternative diagnostic methods [8]. When compared to previous decades, coronary artery disease has significantly increased and has already surpassed all other causes of death. It is extremely difficult for healthcare providers to promptly and accurately identify [33]. A significant obstacle in the field of medical informatics is the variety of issues that might arise with medical data. The following is a summary of these issues: i) Inaccurate, sparse, and temporal information; ii) Small samples; iii) Measuring attribute errors; iv) Inconsistencies in manual data collection; v) Missing values; vi) The skillfulness of medical analysts, practitioners, and technicians in making the diagnosis; and vii) The precision of the machines or instruments used in the diagnosis. The ability to discover new facts in

the form of patterns from the history of information kept in databases is provided by machine learning or deep learning algorithms. One of these algorithms' main tasks is to predict diseases in their early stages. For patients to examine the signs and symptoms and underlying causes for an accurate disease diagnosis, a variety of expensive tests are necessary. On the other hand, utilizing deep learning or machine learning techniques, this quantity of patient testing can be reduced. This minimizes number of tests which plays considerable outcomes in time and accuracy level of prediction. Coronary artery disease prediction is essential since it permits healthcare professionals to analyze the attributes important for diagnosis like blood pressure, diabetes, age, height, weight, etc., effectively. The main goal of this research is to create a system that will assist doctors, such as cardiologists and lab technicians, in predicting the likelihood of developing coronary heart disease in its early stages. One can also get a precise picture of the situation right away and a quick result. The medical professionals can make judgments right away by using the results as a guide to take preventative steps against heart disease and help save lives.

The specific objectives to achieve the aim are given below in sequence.

- To train and test various machine learning classification algorithms using a sample dataset, and to quantify accuracy, performance, and error rate while taking various evaluation criteria into account.
- Selecting the most effective and pertinent algorithms for algorithm modification.
- Recognizing the shortcomings and making the required adjustments to address them.
- Gathering relevant data from local datasets and the UCI Machine Learning database for efficient mining and prediction processes.
- Create more effective artificial neural network architecture. Using updated data preparation techniques to reduce inconsistencies and prepare the dataset for future mining operations in order to produce better results.
- Using a meta-heuristic algorithm to select the features that are most useful in the diagnosis of CAD.
- Use the suggested modified algorithm to train and test the dataset. to assess and contrast the various degrees of accuracy, error rate, and

- Using the data, determine whether the person is having a heart attack or estimate the likelihood of developing heart disease in its early stages.

To assess the degree to which the individual has been impacted, or to determine the degree to which the individual's condition deviates from that of the ailment from which they are suffering.

2. RELATED WORK

Many researchers are trying to predict coronary artery disease or to extract important risk factors, as it is a very serious disease. Feature selection, ensemble methods, and several classifiers are applied to several CVD datasets for the diagnosis of heart diseases [1][2]. Frantisek Babic et al. used Naïve Bayes, Decision Trees, Neural Networks, and Support Vector Machines to get a classification model that works on different types of heart datasets [3][9]. It is important to pay more attention to choosing the medical dataset that will give the best results in the diagnosis of patients [6][7]. Neeraj Bhargava et al. have suggested a decision tree classifier which performs the classification to make predictions about new data and the method will start at the root node and the intermediate nodes are developed in each step; a leaf node is represented, thus ascertaining the class for the record or ascertaining a probability distribution for the attainable classes [4][5]. El-Bialy et al. utilized a few datasets in their research article. Initially, five general features for every dataset (Hungarian, Cleveland, Statlog project, Long Beach VA, and others) are selected and practiced by various kinds of data mining methods : Fast Decision Tree (improved C4.5 by Harry Zhang and Jiang SU) and C4.5 decision tree [13]. Rather than using a fully connected layer, a Full Convolutional Neural Network (FCN) is proposed by the author, Zhu et al. The FCN handles a convolutional layer and will accept large-sized images as input in the process of network training and combines layers of the feature hierarchy so the spatial precision of the output is refined[10][28]. K. Hornik et al [12] used a feed-forward neural network, where the information flow takes place in a single direction and If the activation function is non-constant, continuous, and bounded, then the learning of continuous mappings takes place uniformly over the exact input sets. Alizadehsani et al. [14] used KNN and Naïve Bayes algorithms on electrocardiography (ECG) data, and symptoms including the examination features and accuracy values of 74.20%, 68.33%,

and 63.76% were attained to determine the stenosis of the LCX,LAD, and RCA reciprocally. Alizadehsani et al. [15] Utilized the Naïve Bayes algorithms, SMO, and a new integrated method of symptom and ECG values added to examination features and attained an accuracy value of 88.5% to investigate CAD. Roohallah Alizadehsani et al [16]. The significance of dissimilar features on the presence of disease is found to be not uniform and can be quantified with the Gini index. The inequality between the distribution's values is measured using the Gini index. The enhanced Gini index values indicate the existence of disease. Chetna Yadav et al. attained 93.75% accuracy after applying C4.5 Decision Tree, Support Vector Machine (SVM), SMO, and Improved ARM algorithms to the Z-Alizadeh Sani dataset. An artificial neural network-based prediction model for coronary disease (CHD) has been developed, adopting a composite of genetic and traditional aspects of the disease. In CVD diagnosis, several research efforts were done on Hungarian, Cleveland, Switzerland, and Long Beach VA datasets from the Repository of UCI Machine Learning and accuracy of 86% was obtained, and distinctive performance measures such as sensitivity, Area Under Curve (AUC), specificity, accuracy, FMeasure, and running time are considered as essential measures for heart disease. Patients with disease of type two diabetes mellitus (DM) have a high risk of getting CAD than non-diabetic patients [26][27]. The benefit of sequential backward selection (SBS) is used as a standard approach to feature selection. This method begins with the complete feature set and removes one feature at a time until unwanted features are deleted [24]. Most of the classifiers gave very good performance for the subset of the features. In order to measure the feature selection contribution, the experiments of the model were repeated for an N number of the feature subset [23]. Meta-heuristic algorithms (MHAs) are proven as efficient search and feature selection strategies, which are relatively insensitive to noise [25]. It is very crucial that conventional techniques are not robust, and hence, when the structure of a dataset is modified, the attainment of the existing algorithms also changes. Here, feature selection plays a very decisive role in disease prediction and data mining. Moreover, feature selection decreases the dimension without sacrificing uniqueness and improves the correctness of the prediction model. So we have considered the 18 features which will influence the prediction result with good accuracy. The scope of this paper is to show that by developing a model of an

Improved Artificial Neural Network with a Meta-heuristic Algorithm for Coronary Artery Disease Prediction based on Optimal Feature Selection and this proposed methodology, which includes Feature Selection using a Meta-heuristic Algorithm, the structure of an Improved Artificial Neural Network Architecture, the training process, data set selection, and performance results and discussion. Table 1 provides an examination of the approaches currently in use for categorizing and predicting coronary artery disease. It is clear from Table 1 that performance improvement is still required, both in terms of classification performance, accuracy and feature selection.

3. THE PROPOSED METHODOLOGY

In this paper, we used an improved artificial neural network model in conjunction with a meta-heuristic algorithm for CAD datasets, which is the Cleveland dataset, and expect to get more efficient, sensitive, and accurate results. The unimportant data or sets of irrelevant features are eliminated by applying the Meta-heuristic algorithm on the raw dataset, which contains a 37079 count of data. Subsets of 18 attributes are selected as relevant features that are very crucial in predicting coronary heart disease. The proposed methods use an improved artificial neural network model with an 18–18–1 neural network architecture to learn and predict the CAD diseases based on a subset of relevant features. Coronary artery disease typically develops over a period of decades. Sometimes we will not know we have coronary artery disease until we have a heart attack or significant blockage.

3.1 Structure of Improved Artificial Neural Network Architecture

The structure of the Improved ANN model practiced in this paper is the multilayered feed forward network architecture with 18 input nodes, 18 hidden nodes, and one output node as shown in figure 1.

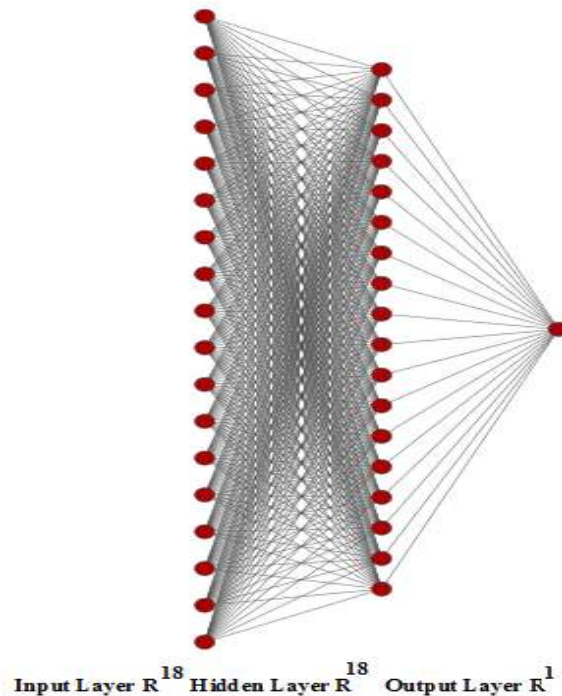


Figure 1 : Proposed System Improved ANN (18-18-1) Architecture for Coronary Artery Disease Prediction

The improved ANN model uses a subset of the input dataset attributes to ascertain the number of neurons in the input layer, and a trial and error method decides the total count of neurons that are concealed to check the integer, which extremely enhances the performance of the improved ANN model. It does prediction, uses the target values, which are continuous, and needs one neuron at the output. The weights and biases are adjusted and the cost function is minimized to improve the learning in the improved ANN model. The cost function contains an error term which is a compute of the proximity of the improved artificial neural network models' outputs to the target values. The binary sigmoidal function is defined as $\text{output} = 1 / (1 + e^{-x})$ with $x = 1$, where x is the total count value of the weighted input value to that specific node. This binary sigmoidal function is committed for each node in the improved ANN model by the activation function.

3.2 Dataset and Feature Selection using a Meta-heuristic Algorithm

This dataset consists of 37079 data points with 50 attributes, which is collected from the UCI repository. In particular, most of the authors or published experiments referred to a subset of 13 of the Cleveland datasets. To date, only the Cleveland dataset has been used by machine learning and deep learning researchers. The existence or absence of

coronary artery disease (CAD) is indicated by 1(presence) and 0 (absence). Many characteristics are more difficult to recognize. There will be inadequate features and minor changes in features in use [1]. It is true that selecting a set of useful features plays an important role in the prediction result.

Table 2: The proposed method's best selected attributes for the prediction of coronary heart disorders

S.No	Variables
1	Age in (years)
2	Gender : Female=0, Male=1
3	Smoking: Present=1, Absent=0
4	Family history of CAD : Present , Absent
5	Diabetes : Type1- Type 2
6	Trestbps : Blood Pressure at Resting (mm/hg)
7	Chol : (mg/dl of serum cholesterol) 1.53 - 14.09
8	fbs : fasting glucose levels (Greater than 120 mg/dl): [1 = yes, 0 = no]
9	HDL : High-density lipoprotein cholesterol (Above 60 mg/dL)
10	thalach (reached the maximum heart rate)
11	exang : Activity induced Chest Pain : 0 = no,[1 = yes]
12	oldpeak (Relative to rest and activity ST depression)
13	slope (the ST section's slope ((2: upsloping, 1: flat, 0: downsloping) of the high point exercise)
14	ca : major vessels (0 to 3)
15	Hypertension systolic over 180 mm Hg
16	Bilirubin levels : 0.3 to 1.2 mg/dL
17	Sleep apnea : Sudden drops in blood oxygen levels
18	LDH :Serum lactate dehydrogenase 140 (U/L) to 280 U/L U/L : units per liter
19	Target : Prediction of Heart Disease for coronary artery disease Present=1, Absent=0

The datasets that are useful and enhance the neural network's performance are estimated by feature selection methods, which select only relevant features. The best algorithm for selecting the feature set is the Meta-heuristic algorithm. It is a stochastic method for heuristic feature set optimization based on natural genetics and biological evolution. It is a stochastic method for heuristic feature set optimization that depends on the mechanism of biological evolution and natural genetics. Meta-heuristic algorithms consider and use a population of individuals to produce the best approximations. Table 2 represents the best selected features for the prediction of coronary heart diseases by our proposed method. At each generation, a new population is created by

choosing individuals in accordance with the variable degree of fitness in the chosen data set or in the problem domain, and the recombination of variables together using operators acquired from natural genetics is also considered, and the offspring might also undergo mutation.

Algorithm

Input D_n :Training Set

Output

Start the algorithm

Measure the Feature relevance with fitness

score

Calculate feature similarities

Generate the initial population of variables

Do

Calculate Fitness Score values

Perform Crossover & Mutation operation

While (Stopping criterion is met)

Select the best features (R_m : Selected features

i.e. Feature set)

End algorithm

3.3 Training Process

When a machine learning system is in capable of recognizing the fundamental trend in the data, i.e., under fitting may happen when a model only works well on training data while failing miserably on testing data. When a model fails to correctly predict outcomes based on test data, it may be over fitted. When a model is calibrated using such a large amount of data, it starts to learn from the noise and inaccurate data values in our data set, and considerable deviation is produced when test results are used for assessment. As a result, the best network for predicting Coronary Artery Disease (CAD) using the chosen data set was found to be a multilayered feed-forward network with an 18-18-1 architecture. The optimal and best network in this research is made up of 18 input neurons, 18 hidden neurons, and one output neuron since it delivers the best fitness in terms of weights and error function. During analysis, the last column is deliberate as the predicted value or target value, and the other columns are considered as input columns. The weights are adjusted in the training phase of an improved ANN model, and the obtained network's output is as close to the target value or predicted value for many examples in the training set. After the training and validation of the improved ANN model on the chosen dataset, the test set is used for obtaining the best output.

4. RESULTS AND DISCUSSION

The primary goal of this paper is to predict

patients with coronary artery disease, and sensitivity is considered a significant performance metric to predict real patients.

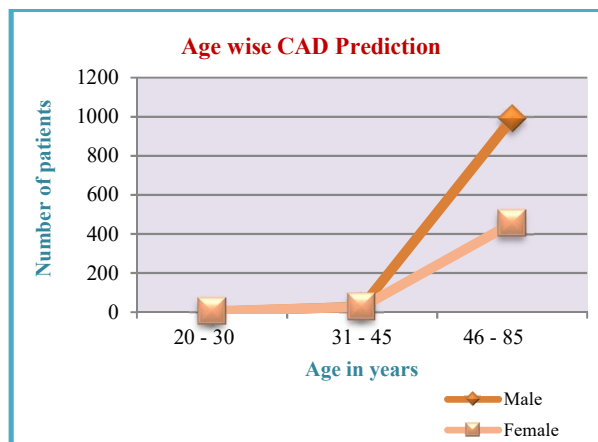


Figure 2: Coronary artery disease among distinct age groups

The present data analysis established unique characteristics among the coronary artery disease patients. Figure 2 shows that the presence of hypertension, the presence of coronary artery disease in the family, and vigorous work increases the risk of coronary artery disease in 8 patients aged 20 to 30 years, 54 patients aged 31 to 45 years, and 324 patients aged 46 to 85 years. It is very clear that males show a significantly higher risk for coronary artery disease than females.

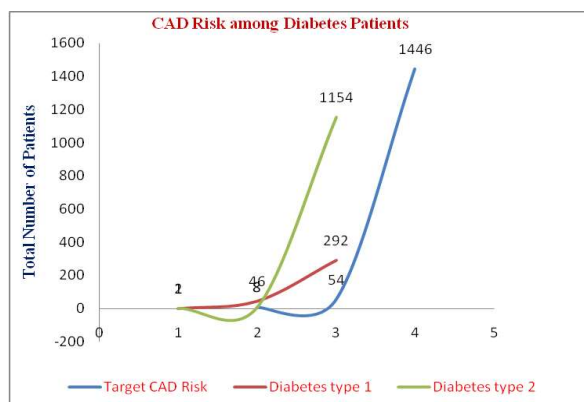


Figure 3: Coronary artery disease prediction in diabetes type 1 and type 2 patients

Table 3 shows that diabetes has a negative effect on a patient's cholesterol level. It ranges from 126.06 to 176.99mg/dL for the 8 patients over age group (20–30) years; it ranges from 99.00 to 259.01mg/dL for the 54 patients over age group

(31–45) years; and it ranges from 86.00 to 404.87mg/dL for the 1446 patients over age group with the presence of family history of CAD as shown in figure 3 & figure 4.

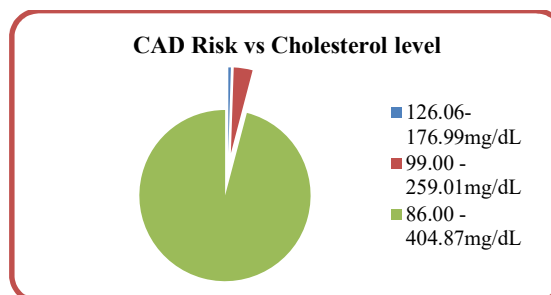


Figure 4: Cholesterol levels in patients with coronary artery disease

As shown in table 2, lactate dehydrogenase (LDH) values in the typical range are 140 units per litre (U/L) to 280 U/L, so the indication of high levels of lactate dehydrogenase (LDH) and isoenzyme lead to more than one cause of severe disease, tissue damage or multiple organ failure.

Table 4 shows that the normal range of bilirubin values is 0.3 to 1.2 milligrammes per deciliter. The range of bilirubin values is 0 to 124.2 mg/dL and the LDH value falls between 35 U/L and 577 U/L for the 8 patients over the age group (20–30) years with the presence of hypertension (systolic over 180 mm Hg). The range of bilirubin values is 0 to 61.6 mg/dL and the LDH value falls between 32 and 159 U/L for the 54 patients over the age group (31–45) years with the presence of hypertension (systolic over 180 mm Hg). The range of bilirubin values is from 0 to 121.41 mg/dL and the LDH value falls between 4 and 1092 U/L for the 1446 patients over the age group (46–85) years with the presence of hypertension (systolic over 180 mm Hg).

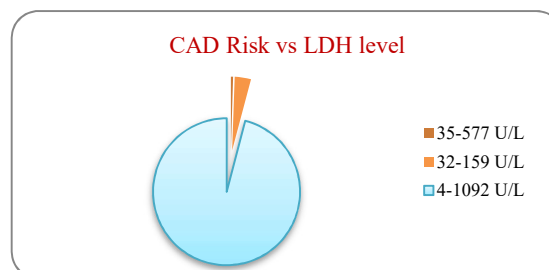


Figure 5: Coronary artery disease prediction in patients with LDH

Figure 5 shows that the bilirubin level is higher than

normal. An abnormal bilirubin level varies by age group and sex and may be different for those who have coronary artery disease or related heart disease. FS: Feature Selection ACC: Accuracy BABC: Binary Artificial Bee Colony MHA: Meta-heuristic algorithm IANN: Improved ANN model.

The following values are obtained as the predicted output values after the training and testing of the improved ANN model with the Meta-heuristic algorithm.

$N = 37079$, $TP = 18716$, $FP = 475$, $FN = 454$, $TN = 17434$;

Sensitivity(SN) : $TP / (TP + FN)$
 $= 18716 / (18716 + 454) = 97.63$;

Specificity(SP) : $TN / (TN + FP)$
 $= 17434 / (17434 + 475) = 97.35$;

Accuracy(ACC) : $(TP + TN) / (TP + FN + FP + TN)$
 $= (18716 + 17434) / (18716 + 454 + 475 + 17434) = 97.5$;

Where, N is the count of data available in the dataset; TP = True Positive; FN=False Negatives; FP= False Positives and TN =True Negatives;

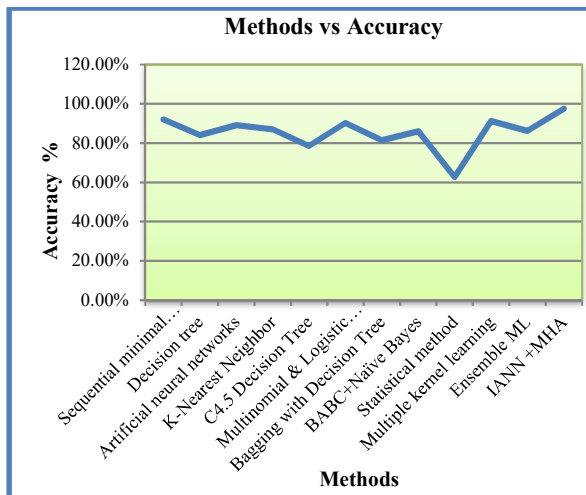


Figure 6: Comparison of Different Prediction Methods for CAD based on the Accuracy Metric

In Table 5, the effectiveness of each prediction method and the proposed method, which is based on an improved artificial neural network model in conjunction with a meta-heuristic algorithm, can be seen. The high accuracy of 97.5% can be obtained when we use selected features (18) from the Cleveland data set as shown in Figure 6. Table 6 shows the comparison of Different Prediction Methods for Coronary Artery Disease with

Sensitivity and Specificity Measures

Table 6: Comparison of Different Prediction Methods for Coronary Artery Disease with Sensitivity and Specificity Measures

Author/System	Dataset	Method	FS	SN	SP
Alizadehsani <i>et al</i> [3]	ZAliza dehsani	SMO	Yes	97.2	79.3
Shouman <i>et al</i> .[29]	Cleveland	DT	No	77.9	85.2
Resul Das <i>et al</i> [30]	Cleveland	ANN	No	80.9	95.9
Kemal Polat <i>et al</i> .[21]	Cleveland	KNN	No	92.3	92.3
Proposed method	Cleveland	IAN+MHA	Yes(18)	97.6	97.3

Our improved artificial neural network model with a meta-heuristic algorithm outperformed well on the chosen features from the Cleveland datasets, hence we obtained 97.63% sensitivity and 97.35 % specificity values. Our improved artificial neural network model with a meta-heuristic algorithm outperformed well on the chosen features from the Cleveland datasets, hence we obtained 97.63 % of people with the CAD disease who were correctly predicted and 97.35 % of people without the disease who were correctly identified by the experiment, as shown in Figure 7.

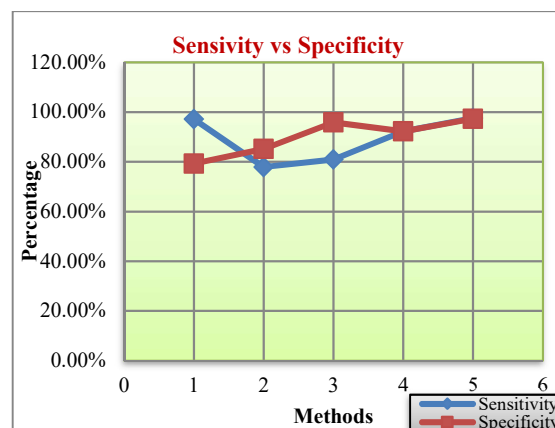
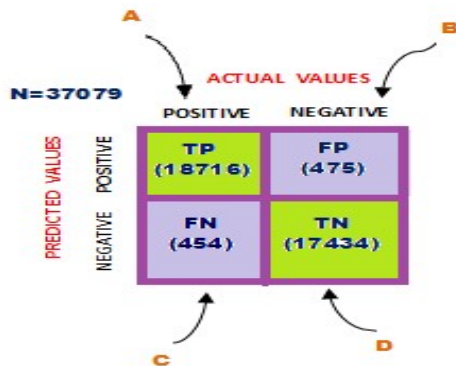


Figure 7: Comparison Of Different Prediction Methods With The Proposed Method In Terms Of Sensitivity And Specificity

4.1 Confusion Matrix for the Proposed Method

The confusion matrix of the proposed method with the above mentioned number of data is computed using the predefined formula after

training, validating, and testing as shown in figure 3. The proposed Improved ANN model results in 97.35 % Specificity, 97.63% Sensitivity, and 97.5% Accuracy. The proposed method outperformed with high values of SP, SN, and AC as shown in figure 8.



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP (18716)	FP (475)
	NEGATIVE	FN (454)	TN (17434)

Figure 8 : Confusion Matrix for Prediction of Coronary Artery Disease using the Proposed Method (IANN+MHA)

TP:(A) is the total count of people correctly predicted as healthy, TN:(D) is the count of people correctly predicted as unhealthy; FN:(C) is the count of people incorrectly predicted as unhealthy when really healthy, and FP:(B) is the count of people incorrectly predicted as healthy when really unhealthy.

5. CONCLUSION AND FUTURE WORK

The main intention of this paper is to develop an Improved Artificial Neural Network model in conjunction with Meta-heuristic algorithm for prediction of Coronary Artery Disease. The Improved ANN model is trained, validated and tested after selecting a subset of 18 features of the Cleveland heart disease dataset which contains the 37079 number of data. In this research paper, the total performance of our improved Artificial Neural Network model with architecture 18-18-1 is evaluated based on Specificity, Sensitivity and Accuracy measures. The proposed method shows the following results as the overall predictive 97.35 % specificity, 97.63% Sensitivity and 97.5% Accuracy. In conclusion, Improved Artificial Neural Network (ANN) models with Meta-heuristic algorithm gives the best and reliable prediction metrics in comparison with the other state of the Deep learning and machine

learning algorithms. The present research issue is that time complexity occurs in selecting the best set feature, and this research paper did not use real-time datasets available in medical labs. In future work, the performance of the proposed system will be tested on different datasets with different numbers of attributes, and a solution will be found to reduce the time complexity.

REFERENCES:

- [1] Kolukisa, B. Hacilar, H. Goy, G Kus, M., Bakir- Gungor, B. Aral, A & Gungor, V. C. "Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease". Proceedings of 2018 IEEE International Conference on Big Data, 2018, pp. 2232-2238.
- [2] Babič F et al. "Predictive and descriptive analysis for heart disease diagnosis", Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE Vol 11, 2017, pp .155-163
- [3] Alizadehsani, R.Hosseini, M. J., Sani, Z. A Ghandeharioun, A and Boghrati, R, "Diagnosis of coronary artery disease using cost-sensitive algorithms" Proceedings of 2012 IEEE 12th International Conference on Data Mining Workshops IEEE, 2012, pp. 9-16.
- [4] Neeraj Bhargava, Girjas Sharma, Ritu Bhargava, Manish Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, 2013, p. 114-1119.
- [5] Nikita Jain, Vishal Srivastava" Data Mining Techniques: a survey paper", IJRET: International Journal of Research in Engineering and Technology, 2013, 2321-7308, p. 116-119
- [6] Inci Aksoy, Bertan Badur, Sona Mardikyan,"Finding hidden patterns of hospital infections on newborn: A data mining approach", Journal of Istanbul University Journal of the School of Business Administration, 2010, 1303-1732;p. 210-226.
- [7] V.Speckauskiene, A. Lukosevicius, "Methodology of Adaptation of Data Mining Methods for Medical Decision Support: Case Study", Journal of Electronics and Electrical Engineering, 2009 , 1392-1215;p.

- 25-28.
- [8] M Cengiz Colak ,Cemil Colak, Hasan Kocatürk, Seref Sağiroğlu, Irfan Barutçu "Predicting coronary artery disease using different artificial neural network models" Department of Cardiovascular Surgery, Faculty of Medicine University of Firat, Elazığ, Turkey, National library of medicine, 2008 , 8(4);P.249-254
- [9] Oleg Yu. Atkov, Svetlana G. Gorokhova, Alexandr G. Sboev ,Eduard V. Generozov , "Elena V. Muraseyeva, Svetlana Y. Moroshkina, Nadezhda N. Cherniy "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters" Journal of Cardiology 2012 (59);p.190-194.
- [10] Y. Zhu and N. Zabarar, "Bayesian deep convolutional encoder_decoder networks for surrogate modeling and uncertainty quantification," J. Comput. Phys., vol. 366, pp. 415-447, Aug. 2018.
- [11] A. M. Elbir, K. V. Mishra, and Y. C. Eldar, "Cognitive radar antenna selection via deep learning," IET Radar, Sonar Navigat., vol. 13, no. 6, pp. 871_880, Jun. 2019.
- [12] K. Hornik, "Approximation capabilities of multilayer feed forward networks," Neural Networks, vol. 4, 1991, pp. 251-257, doi: 0.1016/0893-6080(91)90009-T.
- [13] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, p. 459-468, doi: 10.1016/j.procs.2015.09.132.
- [14] Alizadehsani Roohallah, Habibi Jafar, Bahadorian Behdad, Mashayekhi Hoda, Ghandeharioun Asma, Boghrati Reihane, et al. "Diagnosis of Coronary Arteries Stenosis Using Data Mining" Journal of Medical Signals and Sensors. 2012; 2(3): p.57-65.
- [15] Alizadehsani Roohallah, Habibi Jafar, Hosseini Mohammad Javad, Boghrati Reihane, Ghandeharioun Asma & Bahadorian Behdad "Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features" European Journal of Scientific Research. 2012;82(4):p.542-553.
- [16] Roohallah Alizadehsani, Jafar Habibi ,Zahra Alizadeh sani, Hoda mashayekhi, Reihane boghrati, Asma Ghandeharioun, Fahime Khozeimeh & Fariba Alizadehsani "Diagnosing coronary artery via Data mining algorithms by considering Laboratory and Echocardiography Features" Research in Cardiovascular Medicine official journal of Rajaie Cardiovascular Medical and Research Center, 2013;2(3):p .133-139.
- [17] Yadav, C., Shrikant, L. & Manish, K. S. 2014 "Predictive analysis for the diagnosis of coronary artery disease using association rule mining" International Journal of Computer Applications, 87(4),p.9 to13
- [18] Das, R., Ibrahim, T., & Abdulkadir, S. 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, 36(4), p.7675-7680
- [19] Verma, L., Sangeet, S., & Negi, P. C. 2016 "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data", Journal of medical systems, 40(7), 178 doi: 10.1007/s10916-016-0536-z.
- [20] Nahar, J., et al. 2013 "Association rule mining to detect factors which contribute to heart disease in males and females" Expert Systems with Applications, 40(4) p. 1086-1093.
- [21] Polat, K. n, Şahan, S.& Güneş, S. 2007 "Automatic detection of heart disease using an artificial immune recognitio system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing "Expert Systems with Applications 32(2), p.625-631.
- [22] Takci H. Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences. 2018 ;Jan 27;26(1):p.1-10.
- [23] Armin Attara, Arman Mehrzadehb, Mohsen Fouladb, Davar Aldavoodb, Mohammad Amin Fallahzadehb, Mohammad Assadian Radc& Shahdad Khosropanahd "Accuracy of exercise tolerance test in the diagnosis of coronary artery disease in patients with left dominant coronary circulation" Indian Heart Journal 69 (2017) p; 624-627.
- [24] Devijver, P and Kittler J. "Pattern Recognition: A Statistical Approach," Prentice Hall, 1982.
- [25] Martin Jung and Jakob Zscheischler " A guided hybrid genetic algorithm for feature selection with expensive cost functions" International Conference on Computational

- Science, ICCS 2013 , Procedia Computer Science 18 (2013) p: 2337 – 2346.
- [26] Newman AB, Siscovick DS, Manolio TA, Polak J, Fried LP & Borhani NO “Ankle-arm index as a marker of atherosclerosis in the Cardiovascular Health Study. Cardiovascular Heart Study (CHS)” Collaborative Research Group Circulation 1993, 88(3): p; 837-845
- [27] Himmelmann A, Hansson L, Svensson A, Harmsen P, Holmgren C & Svanborg A. “Predictors of stroke in the elderly” Acta Med Scand 1988; 224(5): p.439 to 443.
- [28] Vaanathi Sundaresan, Christopher P. Bridge, Christos Ioannou, J. Alison Noble “Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks” ,IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) ISSN: 1945-8452 p. 671 to 674.
- [29] Shouman, M., Turner, T & Stocker, R. 2011. “Using decision tree for diagnosing heart disease patients “ Proceedings of the Ninth Australasian Data Mining Conference Volume Australian Computer Society, Inc 121 (p. 23 to 30).
- [30] Das, R. Turkoglu, I & Sengur, A. 2009” Effective diagnosis of heart disease through neural networks ensembles” Expert systems with applications, 36(4), p.7675 to 7680.
- [31] My Chau Tu, Dongil Shin & Dongil Shin “A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms” Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC 2009 p.12-14.
- [32] R R Rajalaxmi “A Novel Feature Selection Algorithm for Heart Disease Classification” International Journal of Computational Intelligence and Informatics, Vol. 4: No. 2 2014 ,p. 117 to 124.
- [33] Long, N. C., Meesad, P., & Unger, H. “A highly accurate firefly based algorithm for heart disease prediction” Expert Systems with Applications, 42(21) 2015, 8221-8231.
- [34] Akanksha Pathak., Akanksha Pathak , Kayapanda Mandana, and Goutam Saha” Ensembled Transfer Learning and Multiple Kernel Learning for Phonocardiogram Based Atherosclerotic Coronary Artery Disease Detection” IEEE journal of biomedical and health informatics vol. 26, no. 6, 2022 2804 – 2813.
- [35] Ankush D. Jamthikar, Deep Gupta , Laura E. Mantella , Luca Saba , Amer M. Johri , and Jasjit S. Suri “Ensemble Machine Learning and Its Validation for Prediction of Coronary Artery Disease and Acute Coronary Syndrome Using Focused Carotid Ultrasound”, IEEE Transactions on Instrumentation and Measurement , Vol. 71, 2022 , 1557-9662.

Table 1: Literature on coronary artery disease disease prediction and classification

Method	Performance metrics(Accuracy)	Findings	Gap identified
Sequential Minimal Optimization [3]	92.09%	Distinguish CAD patients from healthy individuals.	The cost matrix was set with no difference between the two classes; it affected the result.
ANN[8]	81%	Predicting CAD patients from healthy individuals	Low performance
Fast decision tree and pruned C4.5 tree [13]	78.06%	Different data sets were used to target the CAD disease.	Low performance and No proper preprocessing and feature selection were used.
C 4.5 [14] Naïve Bayes[14] KNN[14]	74.20% 62.73% 61.4%	Diagnosing LAD stenosis	Only age and typical chest pain were considered.
SMO and Naïve Bayes [15]	88.52%	Early diagnosis of CAD based on Symptoms	Tested with a single dataset
Bootstrap Aggregating &C4.5[16]	79.54% 61.46%	Diagnosing the LAD stenosis	Tested with a single dataset and Low performance
Apriori algorithm[17]	93.75%	Prediction of Coronary Artery Disease	To select the best set of features, the complexity of the task is high.
Multi-layer perceptron[19]	88.4 %.	Predicting the CAD patients	Important features are missing in the feature selection.
Statistical method[23]	62.7%	Identifying patients with a dominant left coronary artery	For testing, a smaller number of patients were considered.

Table 3: Coronary artery disease prediction in patients with diabetes and unusual cholesterol levels

Age	Gender		Family history	Diabetes		Cholesterol	CAD Risk
	Male	Female		Type 1	Type 2		
20 - 30	2	6	Present	1	2	126.06-176.99mg/dL	8
31 - 45	28	26	Present	46	8	99.00 - 259.01mg/dL	54
46 - 85	991	455	Present	292	1154	86.00 - 404.87mg/dL	1446

Table 4: LDH, hypertension, and bilirubin levels in coronary artery disease patients

Age	Gender		LDH	Bilirubin levels	Hypertension	CAD Risk
	Male	Female				
20 - 30	2	6	35-577 U/L	0 - 124.2 mg/dL	Systolic over 180 mm Hg	8
31 - 45	28	26	32-159 U/L	0-61.6 mg/dL	Systolic over 180 mm Hg	54
46 - 85	991	455	4-1092 U/L	0 - 121.41 mg/dL	Systolic over 180mm Hg	1446

Table 5: A Comparison of Different Prediction Methods with the Proposed Method for Coronary Artery Disease in terms of accuracy

Author/System	Dataset	Method	FS	ACC
Alizadehsani <i>et al</i> [3]	Z-Alizadehsani	Sequential minimal optimization	Yes	92.09%
Shouman <i>et al.</i> [29]	Cleveland	Decision tree	No	84.10%
Resul Das <i>et al</i> [30]	Cleveland	Artificial neural networks	No	89.01%
Kemal Polat <i>et al.</i> [21]	Cleveland	K-Nearest Neighbor	No	87.00%
Randa El-Biary <i>et al.</i> [13]	Cleveland	C4.5 Decision Tree	Yes	78.54%
Luxmi Verma <i>et al.</i> [19]	Cleveland	Multinomial & Logistic Regression	Yes	90.28%
My Chau Tu <i>et al.</i> [31]	UCI Repository	Bagging with Decision Tree	Yes	81.41%
Rajalaxmi <i>et al.</i> [32]	Cleveland	BABC+Naïve Bayes	Yes	86.04%
Armin Attar[33]	Medical centre	Statistical method	No	62.70%
Akanksha Pathak[34]	Own dataset	Multiple kernel learning	No	91.19%
Ankush D. Jamthikar[35]	Medical lab	Ensemble ML	No	86.10%
Proposed method	Cleveland	IANN +MHA	Yes(18 features)	97.50%