Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549490

# EFFICIENT PREDICTION OF DIABETIC CORONARY HEART DISEASE USING ARTIFICIAL NEURAL NETWORK

**1]S.Madhumalar,[2]Dr.S.Sivakumar**

[1]AssistantProfessorofComputerScience,CPACollege,Bodinayakanur

[2]Principal,Associateprofessor,CPACollege,Bodinayakanur

## Abstract

Diabetic Coronary heart disease is caused because of the plaque which block the flow of the blood inside the artery. The identification of this disease at the earlier stage could reverse the same and save valuable human life. To help in predicting it significantly Machine learning algorithm and Artificial Neural network (ANN) can play a major role. In this research work the machine learning algorithms ANN, Support Vector Machine (SVM) and Decision Tree (DT) are applied on the data consisting the various parameter causing the Diabetic Coronary heart disease. Two different datasets namely National Health and Nutrition Examination Survey (NHANES) and South Africa heart disease data are used for evaluation. A comparative analysis on all the three algorithms were carried out and in both the cases the highest accuracy in predicting the disease was produced by ANN.

**Keywords**: Diabetic Coronary heart disease, Artificial Neural Network, Support Vector Machine, Decision Tree, Machine Learning

## 1. Introduction

In the modern society the modality rate due to cause of diabetic coronary heart disease is increasing date by date and people who are in the age group of 20 to 30 are also affected by this disease-causing death in most of the cases. Among the total deaths, one-third occurs with persons below the age of 70 [1]. One of the leading causes of the modality is coronary heart disease and close to nineteen million people lose their precious life in the entire world. Some of the cause of this disease are too much tension, bad habits such as smoking, insufficient physical activity, hypercholesterolemia etc. Sex, smoking, age, family history, poor diet, cholesterol, physical inactivity, high blood pressure, overweightness are the key factors of invoke the diabetic heart disease [2]. Coronary heart disease is one of the vital diseases, causes of death around the world. Predicting the diabetic coronary heart disease based on the symptoms is one of the challenging tasks in the field of medical data analysis. Machine learning (ML) algorithms is useful in predicting in terms of decisions making on the basis of the data available in medical field around the world [3]. Machine learning algorithm can be divided into three categories. Supervised machine learning; task drive, labelled data (Classification/regression) Unsupervised machine learning; data-driven, unlabelled data(clustering); Reinforcement Learning; learning from mistakes [4]. In this research work NHANES data is used for the analysis. The machine learning techniques such as SVM, DT are used for the analysis and well-known machine learning model ANN is used. The ANN model is first trained with the known set of data and then the model is tested for the new data [5]. A comparative analysis is made comparing ANN with the other machine learning algorithms SVM and Decision tree. ANN has the advantage of better adaptability compared to the machine learning algorithms. The machine learning algorithm and the ANN model after analysing the data gives the prediction which helps the medical practitioner to confirm the disease and thereby can take necessary diagnosis to help the patient to recover from the disease. The first steps in the process is to identify the parameter which are correlated to the cause of the disease. After the details analysis the major parameter correlated for the cause of the disease are age, physical level of workout, smoking habit, the gender of patient, cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, very low-density lipoprotein (VLDL) cholesterol, fasting plasma glucose, arterial hypertension, diabetes mellitus, body mass index BMI, and details of family members affected with Coronary heart disease. As the Diabetic Coronary heart disease is a major concern for the society careful selection of the parameter correlating to the disease plays the major role. Any negligence of the parameter which is highly correlated to the disease will make the prediction model created less reliable [6].

In this paper supervised machine learning algorithms are used to predict the diabetic coronary heart disease based on the symptoms. There are three algorithms are used to predict the disease and compare the accuracy of the result.

The flow of the paper is next session discuss about the related work to the Diabetic Coronary heart disease and then next session give the details of the data and the necessary pre-processing steps to be taken before applying to the algorithm. Then the next session gives detailed explanation of the Neural Network models used followed by the machine learning algorithm used. Finally, the result and the conclusion of the work is presented

## 2. Related work

The literature work related to predicting Diabetic Coronary heart disease involving Machine Learning Algorithm and ANN related is presented here. Atkov, Oleg Yu, et al., used ANN model to predict the coronary heart disease. The dataset includes both clinical and functional dataset. Multiple ANN models are created by varying the input parameters and the maximum accuracy obtained by them is 94% [7]. Dutta, Aniruddha, et al. proposed the neural network model with convolutional later for the prediction of coronary heart disease using clinical data [8]. The advantage of this model lies in the design of two convolutional layers which is used to over the imbalance in the data.

Babaoglu, Ismail, et al. compared the support vector machine with multilevel neuron perceptron model in predicting the coronary artery disease [9]. The data used for this analysis is got by performing the stress test. K cross validation is used to increase the accuracy of the test result. The results show that both the model achieves similar results.

Ivanov, Ivelin Georgiev, et al. proposed the modification in the support vector machine for predicting the Coronary heart disease by making the input n folds there by one fold is used for testing and remaining folds are used for training the model [10]. Dinh, An, et al. used various machine learning models such as logistic regression, SVM, Random forest, XGBoost and Ensemble on data for various timeline and for various category of diabetic and it is found that Random forest, XGBoost and Ensemble performs well [11]. The experiments were carried out for clinical and non-clinical data separately. Again the machine learning algorithms are implemented for the models for the data containing parameter for cardiovascular disease.

Khdair, Hisham, et., analysed different machine learning algorithm such as SVM, K-Nearest Neighbour, Neural Networks and a specific technique namely SMOTE is used to overcome the imbalanced classification problem [12]. The result shows that SVM algorithm outperformance the other algorithm. Fan, Rui, et al. used the neural network model in predicting type 2 diabetes and to improve the accuracy fivefold cross validation is used [13].

Qu, Yimin, et al. used retinal image of the patient to analysis the existence of the coronary heart disease [14]. The retinal characteristic was estimated using the ARIA algorithm and followed by that machine learning algorithm is used with ten-fold cross validation method and an accuracy of 85 percent could be achieved. Helman, Stephanie M., et al. surveyed all the recent article related to the Coronary heart disease and it is found that neural network and the SVM algorithm are mostly used which achieves the accuracy level of 80 percent [15]. Krishnani, Divya, et al., used random forest, k nearest neighbours and decision tree for predicting Coronary heart disease and found that by cleaning the data random forest could provide height accuracy [16].

## 3. Materials and Methods

### 3.1 Dataset

The data is collected from National Health and Nutrition Examination Survey (NHANES). The data is segregated into two categories namely the clinical data which got through the lab test and the second category is non-clinical data such as age, level if physical activity etc. Apart from this data the data from our survey and some clinical data also added up to the existing NHANES data. Thus, the overall data collected has 900 data of the people with Diabetic Coronary heart disease and 1400 data of the people without Diabetic Coronary heart disease. Thus, the final data represents the collection of statics and clinical data. The table1 and table 2 below lists the details of the various variable considered for the analysis.

Table 1: Clinical Parameters Considered for the Analysis

| S.no | Parameter | Description |
|---|---|---|
| 1 | Total Cholesterol | It the sum of the good and the bad Cholesterol in the blood. |
| 2 | HDL | High-density lipoprotein is the good Cholesterol |
| 3 | LDL | Low- density lipoprotein is the bad Cholesterol |
| 4 | VLDL | Low- density lipoprotein is also the bad Cholesterol |
| 5 | Arterial hypertension | Indicates high blood pressure in the arteries |
| 6 | Fasting plasma glucose | The expected range is between 70-100 mg/dL |
| 7 | Diabetes mellitus | Parameter indicating usage of blood sugar by the body |

Table 2: Non Clinical Parameters Considered for the Analysis

| S.no | Parameter | Description |
|---|---|---|

| | | |
|---|---|---|
| 1 | Age | The age of the people |
| 2 | Gender | Gender of the people. Considered as some study shows male are more vulnerable to this disease |
| 3 | BMI(Body Mass Index) | Parameter indicating a person is normal, overweight, underweight or obese |
| 4 | Physical Activity | The activity are classified as less, moderate and vigorous |
| 5 | Smoking Habit | Smoking habit of the people |
| 6 | Genetic | To know about any family members are affected with this disease |

As indicated Table 2 consists of the clinical and Table 3 consists of the non-clinical data. It is always a better practise to clean the data before feeding to the machine learning model as any un-cleaned data will make the model created less accurate. The null values in the data if any has to be removed especially dealing with medical data so as to make the model create to work accurately.

### 3.2 ANN Model

ANN model can be used for the classifying the disease into Diabetic Coronary heart disease or not Diabetic Coronary heart disease. In the absence of any strong mathematical function the best way to predict the result by training the data is the ANN model. The significance of ANN is in the correct classification for the unseen data thus making this model the most chosen for complex classification of the data. Thus many ANN model can be created for different dataset also for predicting various disease of interest. The model will perform better if the dataset has the parameter which has high level of correlation to the desired disease prediction. In this article the ANN model created is separately used for clinical and on clinical data and the accuracy is noted for each cases. The following figure 1 shows the general structure of the ANN model.

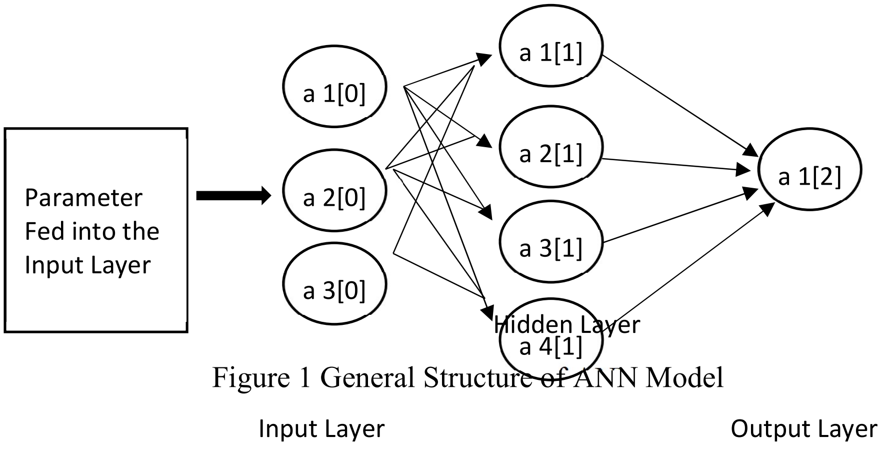

Figure 1 General Structure of ANN Model

## 3.3 Construction of ANN Model

Artificial Neural Network model has 3 layers namely input, hidden and the output layer. The neuron available at each later are interconnected from the input layer to the output layer. In the model created for predicting the Diabetic Coronary heart disease three hidden layers are created. Activation function plays the major role in the ANN model. The activation function used for the hidden layer is sigmoid function and the activation function used for the output layer is soft max function as it works better for the classification problem. The data are split into training and test data and the model is first trained with the know set of data and thus it helps the model to learn and the same model is then tested for its accuracy with the test data. In the forward propagation each and every neuron collects significant amount of information and the weight are adjusted during the backward propagation to as to make the system predict accurate results. Root mean square propagation (RMSProp) is used in the backward propagation so as to ensure minimum cost. The algorithm for the ANN model is mentioned below.

**Algorithm for predict diabetic coronary heart disease using ANN**

1. Begin
2. The pre-processing of diabetic coronary heart disease dataset using Exploratory Data Analysis (EDA) methods. Handles the missing values and repeated data also.
3. Feature selection techniques using standard state of art and proposed DCHD using ANN algorithm
4. The dataset can be divided into train and test. Train the classifiers using train Dataset
5. Validate using testing dataset
6. Compute the performance evaluation metrics
7. End

In the ANN model the parameter which have high correlation towards the Diabetic Coronary heart disease are considered so as to improve in achieving higher accuracy of the model. The neuron used at various stages in the ANN model are used to collect the finest of the details which helps in providing a good relation between the input and the output of the model. The equation 1 represents the output of each neuron and equation 2 and 3 represents the softmax and the relu activation function.

$$N_0 = f\,(\sum w\, n_i + \text{bias}) \qquad (1)$$

Where $N_0$ represents the output from neuron and w represents the weight which will be adjusted during backward propagation in the training process to improve the model efficiency, $n_i$ represents the input to each neuron in the mode.

The softmax activation function is used at the output layer which predicts the classification based on the probability value. The softmax function s(x) is mentioned below

$$S(x) = 1 / (1+ e^{x}) \qquad (2)$$

Any value which has a higher probability value will be concluded into the corresponding classification.

Relu activation function has got the name from the rectifier used in the electronic devices and its significance is it will activate only certain neuron in the ANN model. It will produce output only for the input which is positive and rest all values will be considered zero.

$$R(x) = \max(0, x) \quad (3)$$

### 3.4 Machine Learning Algorithm

There is various machine learning model available and the algorithm which are well suited for the classification problems are Support vector machine and Decision tree. Since the objective is to predict the whether the patient is having Diabetic Coronary heart disease or not these machine learning algorithms are considered. The advantage of these machine learning algorithms are they are not sensitive to the outliers and the structure of the data. Thus these algorithm works well even for the un-cleaned data.

### 3.4.1 Support Vector Machine

Support Vector Machine is a machine learning algorithm which finds its application in classification as well as in the regression problems. More dominantly it is used in the classification rather than the regression problem. The SVM analysis the data and then creates the decision boundary in the n dimensional space. Any new data point will for in to a certain separated by the boundary and thus is much useful for the classification problem. The boundary created for classification of the data is called as the hyperplane. The data points which are close to the boundary are called as the support vectors thus this algorithm is called as support vector machine. SVM creates better decision boundary for both the linear and the nonlinear data. The figure 2 below indicates the SVM.

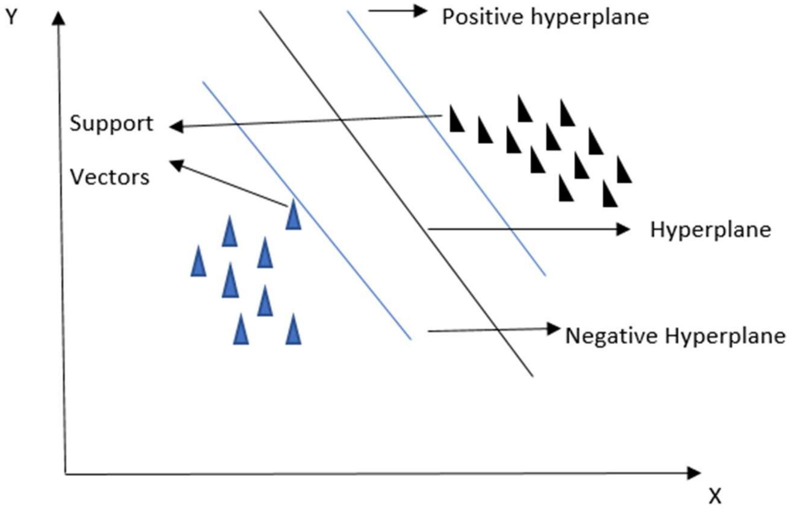

Figure 2 Support Vector Machine

As represented in the figure 2 the maximum margin available is the difference between the positive and the negative hyperplane. Any point below the support can be categorised to a particular classification.

### 3.4.2 Decision Tree

A decision tree is an important machine learning algorithm used for the classification application. The structure of the decision tree will be similar to the flowchart. Each node classifies the data into two parts based on certain criteria. Thus based on the outcome each leaf label will be classified and will be associated with certain label.

### 3.4.3 Construction of the Decision Tree

Each and every layer of the tree the data will be subdivided based on the certain criteria and thereby will be labelled appropriately. This process will be applied further in a recursive way such that all the data are classified into its corresponding category. The advantage of the model is that it does not require any prerequisite knowledge about the data. This algorithm also provides high accuracy in predicting the results even for a higher dimension of the data. The figure 3 below shows the representation of the decision tree algorithm.

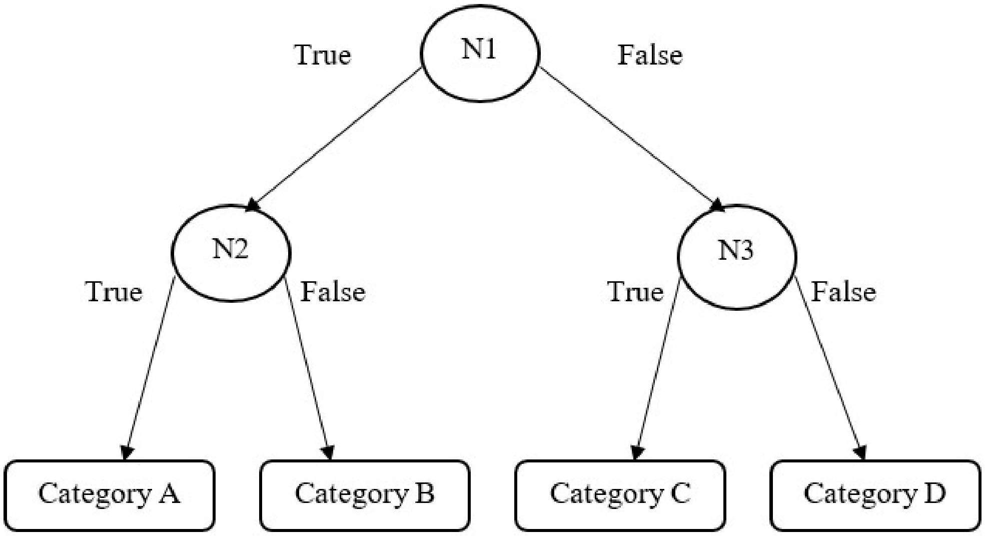

Figure 3 Representation of Decision Tree Algorithm

As represented in the above diagram each node checks for the given condition and based on the decision made a particular path is followed and this process continues till a point where no further classification is possible.

## 4. Result and Discussion

The entire dataset was first divided into training and test data. The training set had 80% of the data, whereas the testing set contained 20% of the data. Then the ANN model will be first trained with the known set of data and once the model learn then it will be tested with the test data. The accuracy can be increased if random sample of data are used for training and testing and for this purpose k-fold cross validation method is used. In this method the data will be divided into k subsets of equal size. Thus the training and testing are performed k times. In each of these cases one fold will be used for the testing and the remaining fold will be used for training the model. Since it is repeated K times the average of these will give the test accuracy of the model.

As mentioned earlier the data collected for clinical and nonclinical data are individually trained using the machine learning algorithm and the ANN model. The following important parameter are notes such as accuracy, precision, recall and F1 score for both the cases. The above parameters are derived from true positive, true negative, false positive and false negative of the confusion matrix. Here true positive is the case where the classified one and the predicted one are the same and true negative is the case where the again the predicting for non-belonging to the class is identified properly. The parameters are calculated by Accuracy = TP+TN/TP+FP+FN+TN, Precision = TP/TP+FP, Recall = TP/TP+FN and F1 score= 2*(Recall * Precision) / (Recall + Precision).

The table 3 below shows the details of the result for diabetes classification and Coronary heart disease classification with non- clinical data

Table 3 Table of Results for Non-Clinical Data

| S.no | Category | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 1 | diabetes classification | SVM | 0.834 | 0.76 | 0.76 | 0.76 |
| 2 | diabetes classification | Decision Tress | 0.786 | 0.74 | 0.74 | 0.74 |
| 3 | diabetes classification | ANN | 0.914 | 0.84 | 0.84 | 0.84 |
| 4 | Coronary heart disease classification | SVM | 0.843 | 0.81 | 0.81 | 0.81 |
| 5 | Coronary heart disease classification | Decision Tress | 0.791 | 0.74 | 0.74 | 0.74 |
| 6 | Coronary heart disease classification | ANN | 0.932 | 0.89 | 0.89 | 0.89 |

The table 4 below shows the details of the score for diabetes classification and Coronary heart disease classification with clinical data

Table 4. Table of Results for Clinical Data

| S.no | Category | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 1 | diabetes classification | SVM | 0.861 | 0.81 | 0.81 | 0.81 |
| 2 | diabetes classification | Decision Tress | 0.714 | 0.70 | 0.70 | 0.70 |
| 3 | diabetes classification | ANN | 0.912 | 0.89 | 0.89 | 0.89 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | Coronary heart disease classification | SVM | 0.892 | 0.81 | 0.81 | 0.81 |
| 5 | Coronary heart disease classification | Decision Tress | 0.784 | 0.72 | 0.72 | 0.72 |
| 6 | Coronary heart disease classification | ANN | 0.921 | 0.90 | 0.90 | 0.90 |

From the above two tables it is quite evident that the ANN model outperforms the other machine learning algorithm such as Decision tree and SVM. The machine learning technique in future will be a part of the entire diagnosis system for all types of disease and will assist the medical practitioner in taking the better decision [17-18]. Uploading the data in the cloud and applying the various machine learning algorithm to predict the algorithm giving better accuracy will be a normal practice in the near feature [19]. The figure 4-7 represents the various score for diabetes and Coronary heart disease classification for non-clinical and clinical data respectively. The test data are taken for the analysis and the results of ANN algorithm produce the better accuracy among the other two algorithms for diabetic coronary heart disease for both clinical and non-clinical data. Because ANN is capable of learning and modelling complicated correlations between features or inputs. This will assist the model in appropriately classifying the unseen data. The information in ANN is stored throughout the network, therefore if certain bits of information are lost, it will not influence the network's functioning or output, hence ANN has a higher accuracy rate than other approaches.

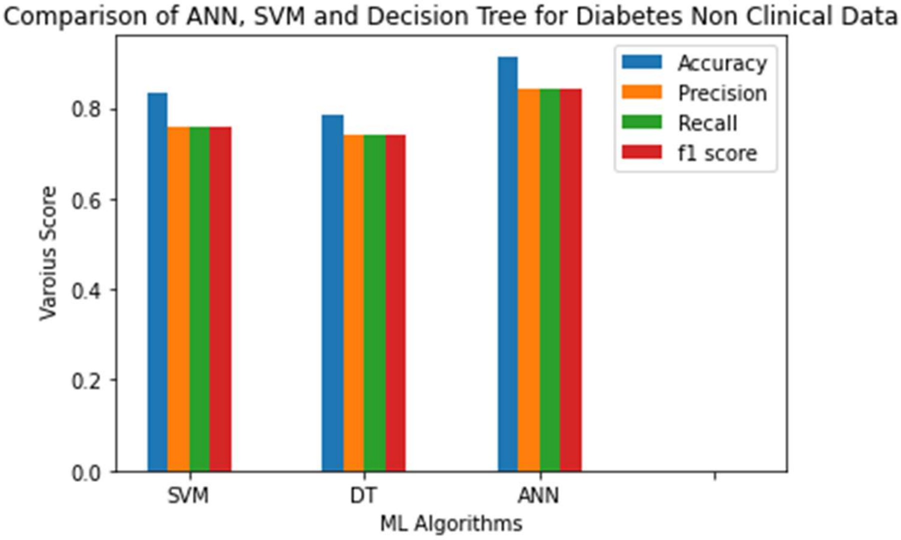

Figure 4: Comparisons of ANN, SVM and Decision Tree for Diabetes – Non Clinical data

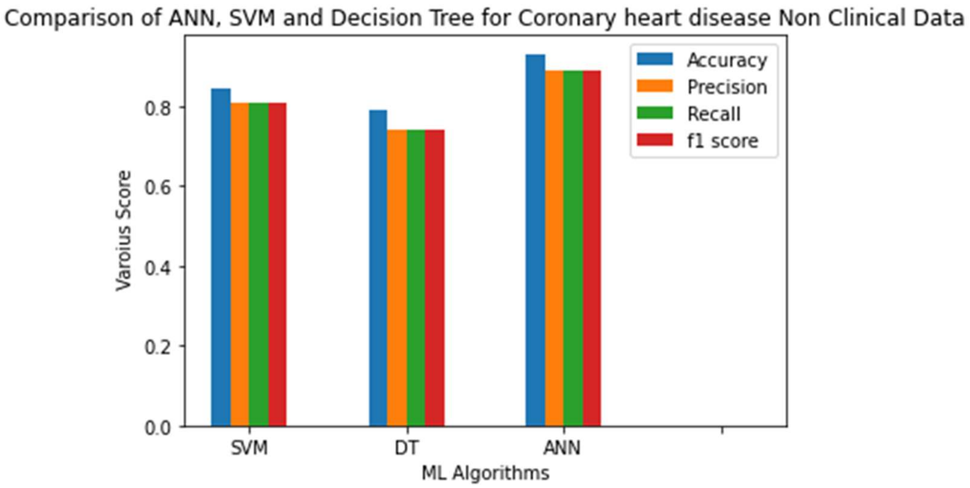

Figure 5: Comparisons of ANN, SVM and Decision Tree for Coronary Heart disease Non Clinical data

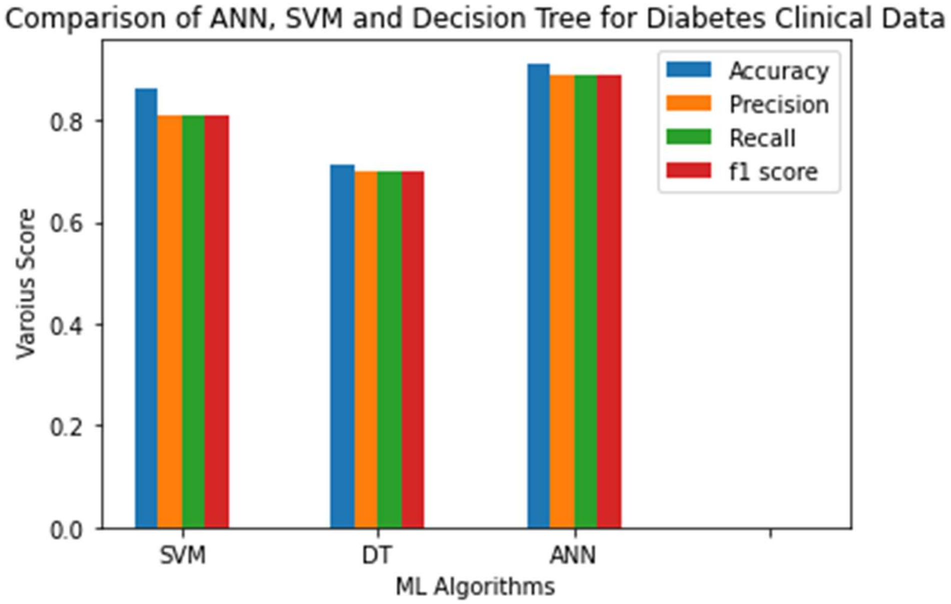

Figure 6: Comparisons of ANN, SVM and Decision Tree for Diabetes Clinical data

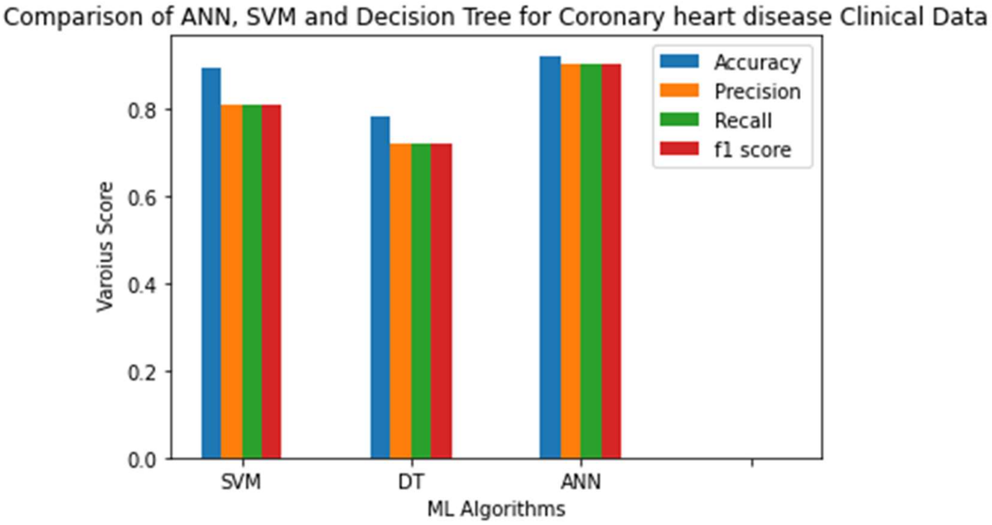

Figure 7: Comparisons of ANN, SVM and Decision Tree for Coronary Heart Disease Clinical data

## 5. Conclusion

Machine learning can be effectively implemented in healthcare filed. The aim of this paper is to discover a model for efficient prediction of diabetic coronary heart disease among the dataset. The dataset chosen from the online repository. Prediction of Diabetic coronary heart disease using SVM, Decision tree and ANN is carried out and the better accuracy is achieved with ANN model for both the clinical and the non-clinical data. The objective of this research work is to identify the Diabetic coronary heart disease during examination and thereby moving towards proper prediction and thus avoiding the loss of precious human life. In near future all the medical practitioner will be getting the necessary suggestion for all the disease as there are abundant amount of data available with various clinics. Also cloud based data analysis will be carried out using appropriate machine learning algorithm making the path for the future.

## References

1. Dolatabadi, Azam Davari, Siamak Esmael Zadeh Khadem, and Babak Mohammadzadeh Asl. "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM." Computer methods and programs in biomedicine 138 (2017): 117-126.
2. Fernandez, Renny, and Terrance Frederick Fernandez. "4 Forecasting Time Series Data Using ARIMA and Facebook Prophet Models." (2021): 47-60.
3. Sivaramakrishnan, S., et al. "Forecasting Time Series Data Using ARIMA and Facebook Prophet Models." Big data management in Sensing: Applications in AI and IoT (2022): 47.
4. Harika, Vangala Ramanuja, and S. Sivaramakrishnan. "Image Overlays on a video frame Using HOG algorithm." 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE). IEEE, 2020.
5. Sivaramakrishnan, S., and T. Kesavamurthy. "Identifying Cluster Head and Data Transmission Through Them for Efficient Communication in Wireless Sensor

Network." Journal of Computational and Theoretical Nanoscience 14.8 (2017): 4014-4020.

6. Vibha, T. G., and S. Sivaramakrishnan. "A Survey of Deep Learning Region Proposal and Background Recognition Techniques for Moving Object Detection." Computer Networks and Inventive Communication Technologies. Springer, Singapore, 2023. 147-164.
7. Atkov, Oleg Yu, et al. "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters." Journal of cardiology 59.2 (2012): 190-194.
8. Dutta, Aniruddha, et al. "An efficient convolutional neural network for coronary heart disease prediction." Expert Systems with Applications 159 (2020): 113408.
9. Babaoğlu, İsmail, et al. "A comparison of artificial intelligence methods on determining coronary artery disease." International Conference on Advances in Information Technology. Springer, Berlin, Heidelberg, 2010.
10. Ivanov, Ivelin Georgiev. "Improving the accuracy of the machine learning predictive models for analyzing CHD dataset." J. Math. Comput. Sci. 12 (2022): Article-ID.
11. Dinh, An, et al. "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning." BMC medical informatics and decision making 19.1 (2019): 1-15.
12. Khdair, Hisham. "Exploring machine learning techniques for coronary heart disease prediction." International Journal of Advanced Computer Science and Applications 12.5 (2021).
13. Fan, Rui, et al. "AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus." Scientific reports 10.1 (2020): 1-8.
14. Qu, Yimin, et al. "Risk Assessment of CHD Using Retinal Images with Machine Learning Approaches for People with Cardiometabolic Disorders." Journal of Clinical Medicine 11.10 (2022): 2687.
15. Helman, Stephanie M., et al. "The role of machine learning applications in diagnosing and assessing critical and non-critical CHD: a scoping review." Cardiology in the Young (2021): 1-11.
16. Krishnani, Divya, et al. "Prediction of coronary heart disease using supervised machine learning algorithms." TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019.
17. Fernandez, Renny, and Terrance Frederick Fernandez. "15 Lethal Vulnerability of Robotics in Industrial Sectors." (2021): 227-238.
18. Milan, Aiswariya. "Lethal Vulnerability of Robotics in Industrial Sectors." Big data management in Sensing: Applications in AI and IoT (2022): 227.
19. Devi, BS Kiruthika, et al. "AN IMPROVED SECURITY FRAMEWORK IN HEALTH CARE USING HYBRID COMPUTING." Malaysian Journal of Computer Science (2022): 50-61.