IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12507

# Prediction Of Cardiovascular Disease and Their Causes Using Machine Learning Techniques

**Ms. V. Lavanya[1], M. Mathew[2], F. Patrick[3], J. Mukesh kumar[4]**

Assistant Professor, Department of Computer Science and Engineering, DMI College of Engineering, Chennai, India[1]

B.E, Department of Computer Science and Engineering, DMI College of Engineering, Chennai, India[2-4]

**Abstract:** Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias), and heart defects you're born with (congenital heart defects), among others. According to World Health Organization (WHO), cardiovascular disease (CVD) is one of the most lethal diseases that leads to the most number of deaths worldwide. Cardiovascular disease prediction aids practitioners in making more accurate health decisions for their patients. Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Machine learning (ML) is a plausible option for reducing and understanding heart symptoms of disease using the device's vital parameters like body temperature, heart rate, and blood pressure. This project proposes a Random Forest technique as the backbone of computer-aided diagnostic tools for more accurately forecasting heart disease risk levels and sending alert messages to the doctor and the guardian with the location details of the patient. Random Forest modeling is a promising classification approach for predicting medication adherence in CVD patients. This predictive model helps stratify the patients so that evidence-based decisions can be made and patients managed appropriately. The chi-square statistical test is performed to select specific attributes from the Cleveland heart disease (HD) dataset. The data visualization has been generated to illustrate the relationship between the features. According to the findings of the experiments, the random forest algorithm achieves 88.5% accuracy during validation for 303 data instances with 13 selected features of the Cleveland HD dataset.

**Keywords:** CVD (cardiovascular disease), Random Forest algorithm, Machine learning, WHO (world health organization).

## I. INTRODUCTION

A. Cardiovascular System

The cardiovascular system is sometimes called the blood vascular, or simply the circulatory, system. It consists of the heart, which is a muscular pumping device, and a closed system of vessels called arteries, veins, and capillaries. As the name implies, the blood contained in the circulatory system is pumped by the heart around a closed circle or circuit of vessels as it passes again and again through the various "circulations" of the body. The adult survival of the developing embryo depends on the circulation of blood to maintain homeostasis and a favorable cellular environment. In response to this need, the cardiovascular system makes its appearance early in development and reaches a functional state long before any other major organ system. Incredible as it seems, the primitive heart begins to beat regularly early in the fourth week following fertilization.

The vital role of the cardiovascular system in maintaining homeostasis depends on the continuous and controlled movement of blood through the thousands of miles of capillaries that permeate every tissue and reach every cell in the body. It is in the microscopic capillaries that blood performs its ultimate transport function. Nutrients and other essential materials pass from capillary blood into fluids surrounding the cells as waste products are removed.Numerous control mechanisms help to regulate and integrate the diverse functions and parts of the cardiovascular system to supply blood to specific body areas according to need. These mechanisms ensure a constant internal environment surrounding each body cell regardless of differing demands for nutrients or the production of waste products.

B. Heart

The heart is a muscular pump that provides the force necessary to circulate blood to all the tissues in the body. Its function is vital because, to survive, the tissues need a continuous supply of oxygen and nutrients, and metabolic waste products have to be removed. Deprived of these necessities, cells soon undergo irreversible changes that lead to death. While blood is the transport medium, the heart is the organ that keeps the blood moving through the vessels. The normal adult heart pumps about 5 liters of blood every minute throughout life. If it loses its pumping effectiveness for even a few minutes, the individual's life is jeopardized.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12507

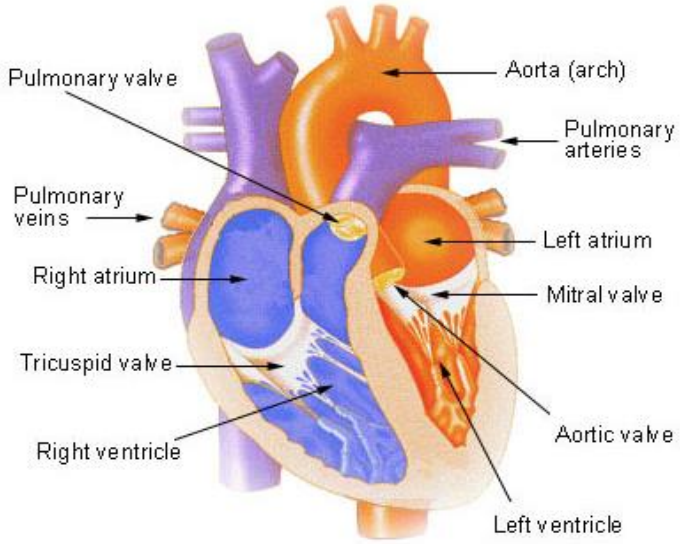



Figure 1: Internal view of the Heart

C. Problem Identified

The major challenge in heart disease is its detection. There are instruments available that can predict heart disease but either it is expensive or is not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time, and expertise. Since we have a good amount of data in today's world, we can use various machine-learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data. We propose the machine learning algorithm Support Vector Machine (SVM) for CHD interpretation. As is known to all, SVM can learn a hierarchical feature representation from raw data automatically, so it does need any handcrafted features by experts.

D. Machine Learning

Machine learning is a branch of AI. Other tools for reaching AI include rule-based engines, evolutionary algorithms, and Bayesian statistics. While many early AI programs, like IBM's Deep Blue, which defeated Garry Kasparov in chess in 1997, were rule-based and dependent on human programming, machine learning is a tool through which computers can teach themselves, and set their own rules. In 2016, Google's DeepMind beat the world champion in Go by using machine learning–training itself on a large data set of expert moves.

E. Machine Learning Work

A machine learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it predicts based on the model.

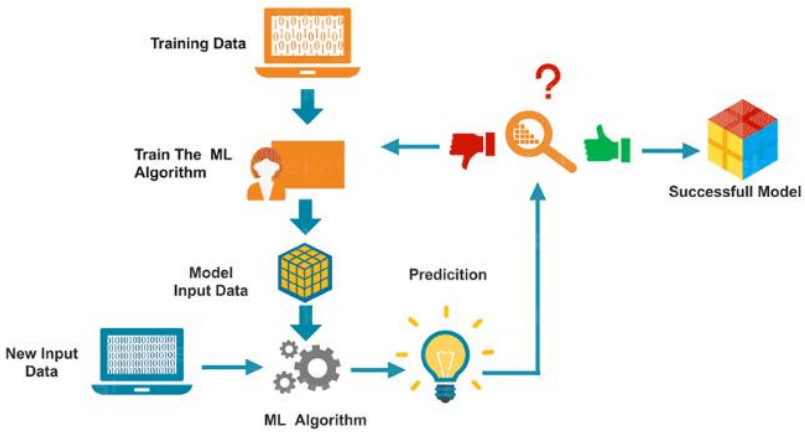



Figure 2: Machine learning algorithm

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.12507

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set. This is just a very high-level example as there are many factors and other steps involved.

F. Types of Machine Learning
Machine learning is sub-categorized into three types:

- Supervised Learning – Train Me!
- Unsupervised Learning – I am self-sufficient in learning
- Reinforcement Learning – My Live My Rules! (Hit & Trial)

Supervised Learning: More Control, Less Bias-Supervised machine learning algorithms apply what has been learned in the past to new data using labeled examples to predict future events. By analyzing a known training dataset, the learning algorithm produces an inferred function to predict output values. The system can provide targets for any new input after sufficient training. It can also compare its output with the correct, intended output to find errors and modify the model accordingly.

Unsupervised Learning: Speed and Scale-Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. At no point does the system know the correct output with certainty. Instead, it draws inferences from datasets as to what the output should be. Reinforcement Learning: Rewards Outcomes-Reinforcement machine learning algorithms are a learning method that interacts with its environment by producing actions and discovering errors or rewards. The most relevant characteristics of reinforcement learning are trial and error search and delayed reward. This method allows machines and software agents to automatically determine the ideal behavior within a specific context to maximize its performance. Simple reward feedback — known as the reinforcement signal — is required for the agent to learn which action is best.

## II. LITERATURE SURVEY

**1. Suresh Kumar S, Pavithra S, Jemima Preethi M., et al. (2020)** Cardiovascular disease is a leading cause of death worldwide, and early prediction and prevention of the disease can reduce mortality rates. Machine learning algorithms can be used for prediction, but the accuracy of different algorithms needs to be compared to identify the best-performing algorithm. The study used a dataset of 303 patients, with 14 features including age, sex, blood pressure, and cholesterol levels. Four machine learning algorithms - K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine - were trained and tested using 10-fold cross-validation. The performance was evaluated using accuracy, precision, recall, and F1-score metrics. The Random Forest algorithm outperformed the other three algorithms, achieving an accuracy of 90.38%, precision of 91.26%, recall of 90.55%, and F1-score of 90.9%. This was been developed by Suresh Kumar in 2020.

**2. Jyoti Yadav, Neha Bhargava, Gaurav Kumar, Alok Kumar Singh. et al. (2021)** Cardiovascular disease is the leading cause of death globally, and early prediction can help in early intervention and better management. The authors conducted a systematic literature review of articles published between 2017 and 2020, using the PubMed, ScienceDirect, and IEEE Xplore databases. The selected articles were analyzed based on the machine learning algorithms used, dataset characteristics, performance metrics, and limitations The review found that machine learning techniques, including decision trees, random forests, support vector machines, and artificial neural networks, have been widely used for cardiovascular disease prediction. The study also identified several challenges, including data imbalance, overfitting, and lack of interpretability of models.

**3.Durga Prasad Sharma, Dhirendra Pratap Singh, Alok Kumar Yadav, and Vaibhav Pandey. et al. (2020)** Cardiovascular disease (CVD) is the leading cause of death globally, and its prediction is challenging due to the involvement of multiple risk factors. The study used a dataset of 12,201 patients from the National Health and Nutrition Examination Survey (NHANES) 2011-2012. The dataset was preprocessed and feature-selected, and then five machine learning algorithms, including logistic regression, decision tree, random forest, support vector machine, and K-nearest neighbor were trained and tested for predicting CVD risk. The study used only one dataset, which may limit the generalizability of the results. The study did not include some potential risk factors for CVD, such as physical activity and family history of CVD.

**4.Ahmed Al-Mallah, Mouaz Al-Mallah, Yasar Albakri, and Fatima AI-Anazi. et al. (2019)** Cardiovascular disease (CVD) is the leading cause of death worldwide. Early detection and prediction of CVD can improve patient outcomes and reduce healthcare costs. To develop a machine learning model for predicting CVD using clinical data. The authors used a retrospective cohort study design and developed a machine learning model using three different algorithms: logistic

DOI: 10.17148/IJARCCE.2023.12507

regression, decision tree, and neural network. They trained and tested the model using a dataset of 13,611 patients with 54 clinical features. The authors found that the neural network algorithm outperformed the logistic regression and decision tree algorithms, with an accuracy of 87.3%. The most important features for predicting CVD were age, systolic blood pressure, and body mass index.

**5. Muhammad Attique Khan, Muhammad Usman Ghani Khan, Naveed Iqbal Qureshi. et al. (2021)** CVD is a major public health concern, and early identification and management of risk factors can improve patient outcomes. Traditional risk prediction models have limitations in accuracy and applicability to different populations. To develop a machine learning model for predicting CVD risk using big data analytics. The authors used a retrospective cohort study design and developed a machine-learning model using the Random Forest algorithm. They trained and tested the model using a dataset of 11,043 patients with 45 clinical features and demographic data. They also used feature selection and data pre-processing techniques to improve the model's performance. The authors found that the Random Forest model had an accuracy of 83.56% in predicting CVD risk. The most important features for predicting CVD risk were age, blood pressure, and cholesterol levels. The model was also able to identify subgroups of patients with a higher risk of CVD.

## III. PROPOSED SYSTEM

The proposed methodology for web-based cardiovascular disease (CVD) prediction and its causes using Random Forest Algorithm. It involves collecting and pre-processing health and sensor data, training a Random Forest model for CVD prediction, developing an alert system, generating personalized recommendations, evaluating the system's performance, and deploying it on a web-based platform. Careful attention to data quality, model validation, and privacy and security concerns is essential for the success of the proposed system.

A. Dataset Introduction

In this project, we collected a heart disease dataset known as the Cleveland heart disease database from an online machine learning and data mining repository of the University of California, Irvine (UCI). it covers the role of various subsystems/modules/classes along with implementation details listing the code for the major functionalities.

B. Data set Preprocessing

Every dataset consists of various types of anomalies such as missing values, redundancy, or any other problem for removing this problem there is a need for a certain step called processing data. The pre-processing step is needed to overcome such a problem. There are three pre-processing steps:

Formatting**:** The data set used for implementation is taken from the UCI repository, it may contain certain attributes whose names are not clear in the (dataset name) and also contain certain unrelated attribute which is not useful for the greater performance of the proposed work. An attribute name "Thal" has been removed from the dataset by using the following command in R, Dataset $Thal<-Null

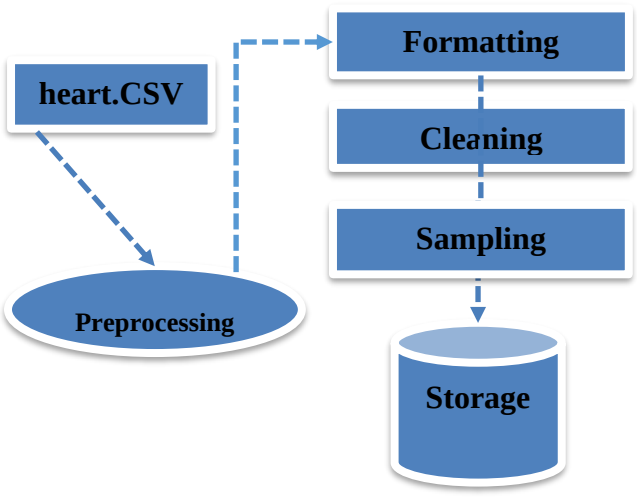



Figure 3: Data Processing

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12507

Cleaning: This part of pre-processing belongs to removing or fixing missing out the entry in the data frame. A row containing these incomplete columned to be removed also for removing certain redundant entries in the data frame this step is recommended

Sampling: Sampling is also done on the dataset to enhance the performance of the algorithm on the sample data set may lead the algorithm to take longer time.

C. Feature Selection

In feature selection, irrelevant features are eliminated and the most important or relevant features are applied to the network. Thus, if we supply all features to Random Forest, some features may be noisy and if they are learned in the training process, they may degrade the generalization of the network although the network will show good performance on the training data. That is why a large number of features are also considered one of the main causes of overfitting. Thus, searching out an optimal subset of features by eliminating noisy features can help Random Forest to show good performance on both training and testing data. In this module, we use the $X^2$ statistical model to eliminate irrelevant features. In the feature's elimination process, we compute $X^2$ statistics between each non-negative feature $F_i$ and class i.e., y. The $X^2$ model performs an $X^2$ test that measures dependence between the features and class. Hence, the model is capable of eliminating those features which are more likely to be independent of class. Because these features can be regarded as irrelevant for classification.

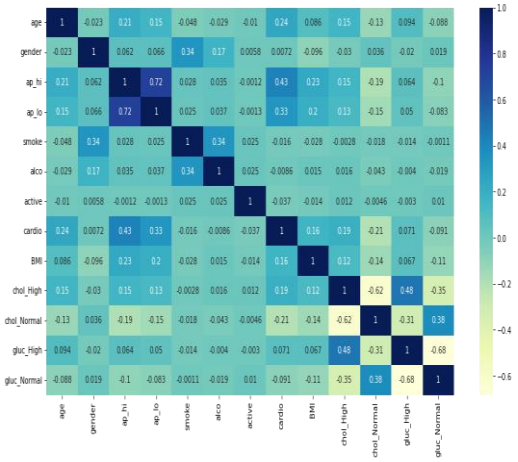



Figure 4: Heatmap

D. Random Forest CVD Classification

Random Forest is an ensemble learning algorithm that is widely used for classification and regression tasks. It is a combination of multiple decision trees, where each tree is built using a subset of the available data and a random selection of features. In the case of Cardiovascular Disease (CVD) prediction, the Random Forest algorithm can be used to build a classification model that predicts the likelihood of a person having CVD based on their age, gender, blood pressure, cholesterol levels, smoking habits, diabetes status, and family history of CVDs.

The Random Forest algorithm works as follows:

Random Sampling: A random sample of the available data is selected to build each decision tree in the forest. This ensures that each tree is built on a different subset of data, reducing overfitting and increasing the diversity of the model.

Random Feature Selection: At each node of the decision tree, a random subset of features is considered to split the data. This also helps to reduce overfitting and increase the diversity of the model.

Voting: Once all the decision trees are built, the prediction for a new data point is made by taking the majority vote of all the decision trees. This voting mechanism ensures that the model is robust to noise and outliers in the data.

The Random Forest algorithm has several advantages over other classification algorithms. It can handle large datasets with high-dimensional features and is less prone to overfitting. It is also easy to interpret the results and identify the most important features that contribute to the prediction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12507

E. Performance Analysis

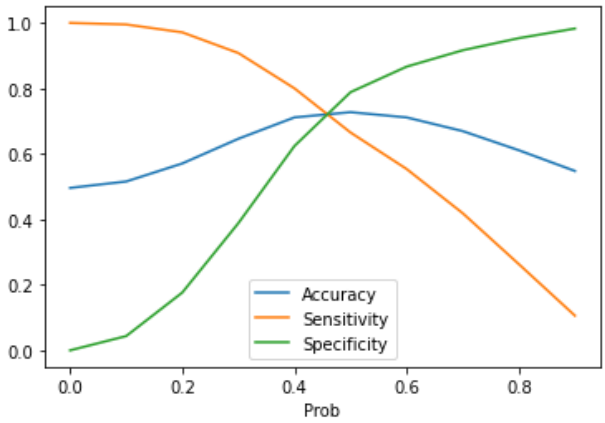
Evaluation Metrics To comprehensively evaluate the classification performance and effectiveness of our proposed method, we applied accuracy, recall, F1 score, precision, specificity, ROC, and AUC evaluation metrics. For the sake of expression of the significance and calculation formula of these evaluation metrics, we introduced the confusion matrix (See Table 5) first. The confusion matrix is a specific matrix used to visually present the performance of the algorithm. The confusion matrix of binary classification consists of two rows and two columns.

Rows represent the true labels of the two classes in the dataset (denoted true). Columns represent the predicted label of the two classes acquired by the model (denoted as type). As shown in Table 5, the confusion matrix of binary classification includes four indicators: TN, FN, FP, and TP. The four indicators are defined as follows. We specified that the label of the positive class is 1 and the label of the negative class is 0.

F. Accuracy

Accuracy refers to the proportion of samples that can be correctly predicted by the model in all samples. The calculation equation of accuracy is as follows. TN, TP, FN, and FP refer to true negative, true positive, false negative, and false positive, respectively.

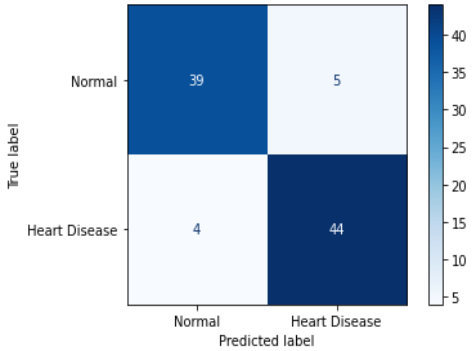
**Accuracy = TN + TP TN + TP + FN + FP (1)**




Accuracy is one of the most frequently used and most important model performance evaluation metrics. However, in the dataset with a class imbalance problem, due to the influence of majority class samples, the accuracy is often difficult to accurately measure the classification ability of the model. Therefore, in the dataset with a class imbalance problem, in addition to accuracy, more evaluation indicators need to be applied.

G. ROC and AUC

The area under the curve (AUC) is the area under the receiver operating characteristic (ROC) curve. The ROC curve is drawn with the false positive rate (FPR) as the x-axis and the true positive rate (TPR) as the y-axis. The ROC curve intuitively reflects the relationship between specificity and recall. The value of AUC is between 0 and 1, when the value of the x-axis (i.e., the false positive rate (FPR) of the model) is closer to 0, and the value of the y-axis (i.e., the true positive rate (TPR) of the model) is closer to 1, the value of AUC is closer to 1. The closer the AUC value is to 1, the higher the prediction performance of the classifier.



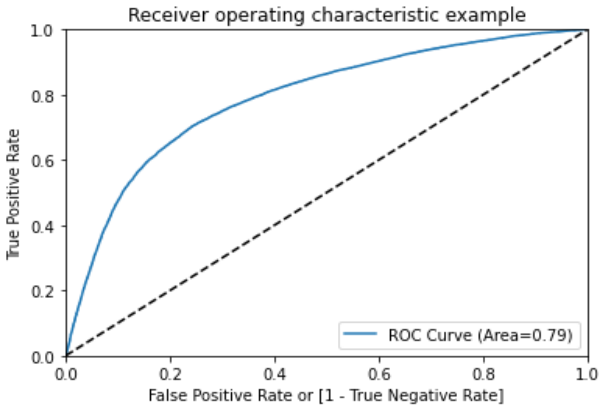
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12507

| Classifiers | Accuracy | Recall | F1 | Precision | Specificity | AUC |
|---|---|---|---|---|---|---|
| Random Forest | 0.97 ± 0.060 | 0.922 ± 0.068 | 0.980 ± 0.038 | 0.963 ± 0.041 | 0.881 ± 0.095 | 0.93 ± 0.05 |




## IV. RESULT & DISCUSSION

The importance of the cardiovascular disease (CVD) prediction and alert system with sensor data using Random Forest lies in its potential to improve the early detection and management of CVDs, which are a leading cause of morbidity and mortality worldwide. CVDs are a major health concern globally, and their prevalence is expected to increase in the coming years due to factors such as aging populations and changes in lifestyle and dietary habits. Early detection and management of CVDs can significantly improve patient outcomes and reduce healthcare costs, but current diagnosis and management processes are often subjective and time-consuming. The proposed system addresses these challenges by leveraging sensor data and machine learning algorithms to predict the likelihood of CVDs and provide personalized recommendations and alerts to patients, guardians, doctors, and ambulance services in case of any emergency. By improving the accuracy and timeliness of CVD diagnosis and management, the system can improve patient outcomes and reduce healthcare costs. The best model is the random forest tree model. The accuracy is 96.7%, with an f1-Score, recall and precision of 97.5%

## V. CONCLUSION

In this project, the machine learning-based support vector machine classification and prediction models were developed and evaluated based on the diagnostic performance of coronary heart disease in patients using sensitivity, specificity, precision, FScore, AUC, DOR, 95% confidence interval for DOR, and K-S test. The developed machine learning classification and prediction models were built with a multilayer perceptron equipped with linear and non-linear transfer functions, regularization and dropout, and a binary sigmoid classification using machine learning technologies to create a strong and enhanced classification and prediction model.

The developed Random Forest-based classification and prediction models were trained and tested using the holdout method and 28 input attributes based on the clinical dataset from patients at the Cleveland Clinic. Based on the testing results, the developed machine learning models achieved diagnostic accuracy for heart disease of 83.67%, a probability of misclassification error of 16.33%, a sensitivity of 93.51%, a specificity of 72.86%, a precision of 79.12%, an F-score of 0.8571, AUC of 0.8922, the K-S test of 66.62%, DOR of 38.65, and 95% confidence interval for the DOR of this test of [38.65, 110.28]. These results exceed those of currently published research.

Therefore, the developed machine learning classification and prediction models can provide highly reliable and accurate diagnoses for coronary heart disease and reduce the number of erroneous diagnoses that potentially harm patients. Thus, the models can be used to aid healthcare professionals and patients throughout the world to advance both public health and global health, especially in developing countries and resource-limited areas where there are fewer cardiac specialists available.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.12507

## REFERENCES

[1] K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017. [2] Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017.

[3] Dr.S.Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.

[4] Ashok Kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016.

[5] Megha Shahi, R. Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 2017, pp.457-466.

[6] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis of the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018. [7] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.

[8] Jayami Patel, Prof. Tejal Upadhay, Dr. Samir Patel, "Heart disease Prediction using Machine Learning and Data mining Technique", March 2017.

[9] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", 2016, pp.962-969.

[10] K.Gomathi, Dr. D.Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December 2016, pp.12-14.