e-ISBN: 978-93-6252-521-5
IIP Series

CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING TECHNOLOGY

Abstract

Cardiovascular illnesses, often known as heart diseases or CVIs, have been the leading cause of death worldwide in the past few decades that are now a particularly serious illness, not just in India but throughout the entire globe. Therefore, a system that is dependable, accurate, and workable is required to identify these illnesses early enough for proper therapy. Large-scale and sophisticated data processing has been automated by using machine learning techniques and algorithms to a variety of wellness databases. Recently, a number of researchers have started employing different approaches to machine learning to assist the medical community and experts in the detection of heart-related illnesses. In our bodies, the cardiovascular system is the second most important function after the nervous system, which is given greater attention. It circulates blood, supplying it to every organ in the human anatomy. In the medical world, predicting the onset of cardiac disorders is an important task. Data analytics helps the medical center anticipate different ailments and is valuable for making predictions based on additional information. Every month, an enormous variety of patient-related data is kept up to date. Chronic infections can be predicted with the assist of the information that has been stored. Cardiovascular disease can be predicted using a few data mining as well as machine learning techniques, including support vector machines (SVM), Random Forest (RF), and Artificial Neural Network (ANN).Cardiovascular disease diagnosis and prognosis have become difficult issues for medical professionals and institutions, both domestically and internationally. The enormous number of fatalities with heart disease must be decreased, hence a rapid and

Authors

Dr. C. Daniel Nesa Kumar
Assistant Professor
Department of BCA
Sri Ramakrishna College of Arts and Science
Coimbatore.

Dr. J. Jeyaboopathi Raja
Assistant Professor
Department of Computer Science
Sri Ramakrishna College of Arts and Science
Coimbatore.

Dr. M. Manjutha
Assistant Professor
Department of Information Technology
PSGR Krishnammal College for Women
Coimbatore.

Mr. T. Pradeep
Assistant Professor
Department of BCA
Sri Ramakrishna College of Arts and Science
Coimbatore.

effective diagnosis method must be found. In this field, data mining methods and automated learning processes are crucial. Scientists are working more quickly to create systems that use machine learning algorithms to assist physicians in identifying and predicting cardiac problems. The study the endeavor's primary goal is to use algorithms powered by machine learning to anticipate an individual's cardiac condition.

Keywords: Heart Disease, Machine Learning, Classification Algorithms, Naïve Bayes, SVM.

I. INTRODUCTION

More than thirty percent of all fatalities are caused by coronary artery disease (CAD), a kind of coronary artery disease that is still the biggest cause deaths globe. By the year 2030 it is projected that there would be twenty-two million fatalities worldwide if everything is done. Heart attacks and strokes can be caused by plaques on the artery walls that impede blood flow. Numerous risk factors, including a poor diet, inactivity, and heavy consumption of cigarettes and alcohol, contribute to heart disease [1, 2]. A balanced way of life that includes eating less salt, eating more vegetables and fruits, getting regular exercise, giving up alcohol and tobacco, and cutting back on smoking and other unhealthy habits all help to lower the likelihood of cardiovascular disease [3]. Utilizing information about patients gathered from many hospitals and health care facilities is the way to solve these issues. The decision support system is utilized to obtain the outcomes and obtain an additional view from a medical professional with knowledge. This diagnostic strategy saves both cash and time by eliminating needless test conductions [4, 5]. Additional information is generated by such systems since a management system for hospitals is currently being utilized for organizing patients or medical data. The DSS, which makes use of the NB (Naïve Bayes) algorithm, was created to predict cardiac disease. Significant characteristics related to heart disease are retrieved from an ancient database by use of a program and user input via a website [6, 7].

Cardiovascular disease is a condition in which many neurohormonal regulating systems are activated in the early stages. These compensatory mechanisms can quickly result in the consequences of a high-fat diet (Heart Disease), including increased ventricular dysfunction, exertional dyspnea, peripheral swelling, respiratory transformation, and persistent modifications in input and following load. The patient is offered additional alternatives for HFD treatment, such as dietary modifications and implanted or medicated devices like pacemakers or defibrillators. As the main driver of health care expenses in this group of people is treatment for abrupt HFD compensating, the primary concern is making sure that follow-up is being done. Research and data analysis indicate that cardiac conditions are the biggest problem that patients, especially those on HFD, suffer [8, 9]. The first phase in getting medical attention for a number of disorders is early identification and recognition of heart disease.

These days, the HFD is being recognized as a potential problem for conditions like heart disease, sleeplessness, and pressure. In order to diagnose HFD on an ECG, one must notice differences in the beating heart frequency between one PQRST wave and the subsequent wave. MCG (Magenetocardiography) is a new and prospective non-invasive diagnostic method for early identification of IHD. Compared to an ECG, an MCG is less affected by electrode-skin touch disruption but it is more susceptible to transverse causes and vortex currents that penetrate injured heart tissue MCG translation has restricted use in clinics, requires a lot of time to complete, and is heavily reliant on translating knowledge though having a good signal fidelity. Therefore, an autonomous system that can identify and pinpoint infarction early on could prove beneficial to physicians [10].

For an overall decrease in the death rate, early detection of cardiac conditions with improved diagnostics and patients at greater risk using a prediction model can be advised. This also improves making choices for subsequent preventative and curative measures. A model of prediction is used in CDSS to assist doctors in determining the patient's risk of heart

disease and to prescribe suitable medications to manage subsequent risk. Furthermore, a large body of research has demonstrated that the application of CDSS can enhance clinical choice-making, preventative care, and judgement quality, in that order [9, 11]. In several nations, coronary artery disease (CAD), also referred to as ischemic heart disease (IHD), is the primary cause fatalities for persons over 35. It rose to the top reason for fatality worldwide during the identical period. IHD happens when stenosis of the coronary arteries reduces the amount of blood that reaches the heart. Serious outcomes from myocardial injury can include ventricular fibrillation as well as a myocardial infarction, which can cause abrupt death of the heart.

For the last ten years, heart attacks have been the leading cause fatalities globally. The World Health Organization (WHO) estimates that cardiovascular illnesses account for about 17.9 million fatalities annually, with coronary artery disease and cerebral stroke accounting for 80% of these deaths [12]. Heart disease is caused by numerous behavioural variables, including genetic predisposition and behaviours from both the personal and professional spheres. Heart disease is frequently determined by a combination of physiological variables including obesity, hypertension, high blood cholesterol, and pre-existing heart diseases, as well as risk factors like smoking, excessive beverage and caffeine usage, stress, and lacking in exercise. In order to take preventative action to minimize the consequences that result from heart disease, it is essential to diagnose the condition as soon as possible. Nowadays top healthcare challenges are providing high-quality care and making precise and on-time predictions. Technology can help alleviate the latter issue with the aid of data mining and machine learning. Data mining is the method of removing useful information from a big collection of raw data. It entails using a variety of applications to analyse patterns in massive data sets. It also includes computing power in conjunction with efficient information storage and gathering. A subset of data mining techniques called machine learning (ML) effectively handles gigantic, properly organized datasets. Machine learning has applications in medicine that include identifying illnesses, recognition, and predictions. To determine which algorithm used for machine learning is particularly precise, a number of them are examined, including Random Forest, KNearest Neighbor, Decision Tree, Naïve Bayes, Support Vector Machine, Logistic Regression, and the ensemble approach of XG Boost. The discussion of the current classification methods are done in this study. The scopes of upcoming studies and potential avenues for advancement are also mentioned in the report.

II. LITERATURE ANALYSIS AND RELATED WORK

Using skin sensors, the ECG records the electrical signals of the cardiac muscle in a variety of wave shapes. It is a noninvasive method of identifying heart disease that takes cardiac wellness, blood pressure, and pulse into account. The human physique's cell count is not in close association with the environment. Additionally, they rely on the circulatory system to provide them with transportation. Two types of fluids circulate through the circulatory system of the body. The first kind of liquids is plasma. The vessels in the heart are formed here by the circulatory pathway. The next kind of liquids is lymphatic. The lymph nodes plus lymphatic veins make up the lymphatic system's structural elements. The circulatory system and the system of lymphocytes can combine to produce cardiovascular disease [12]. A heart cycle is a sequence of actions in the heart's rhythm. Both atria normally contract during a heart cycle, with each ventricular contracting synchronistically just a portion of an instant apart. Heart muscle cells are used to create and connect the heart, so

when one contracts, it excites neighboring cells. Aerobics respiration is facilitated by the muscles' relaxation throughout cardiac beats.

Gomathi et al. employed decision tree and Naive Bayes information mining approaches to forecast various illness kinds. Their primary focus was on cancer of the breast, diabetes, and coronary artery disease prognosis. The erroneous measures were used to generate outcomes. The Naive Bayes classifier approach was proposed by Miranda et al. for predicting heart disease. The writers have taken into account a few significant risk variables to determine heart disease.

Avinash Golande and colleagues investigate several machine learning techniques that are useful in the categorization of cardiac disease. The classification algorithms KNN, K-Means, and Decision Tree were investigated, and their respective accuracy was examined. This study suggests that decision trees yielded the best results, and it also suggests that they can be made more effective by combining several strategies and fine-tuning their parameters.

An artificial intelligence (AI) model created by Fahd Saleh Alotaibi compares five distinct methods. When contrasted to Matlab and Weka, the precision of the Rapid Miner tool was greater. This study examined the classification algorithms' accuracy using Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM. The most accurate algorithm was the decision tree algorithm.

Several categorization algorithms were employed in a study conducted by Theresa Princy, R. et al. to anticipate cardiac disease. Naive Bayes, KNN (K-Nearest Neighbor), Decision trees, and neural networks among the classification methods employed. The efficacy of the classifications was examined for various numbers of characteristics.

Purushottam, et al.'s research developed decision tree and hill climbing algorithms, which are utilized in the System for Effective Heart Disease Prediction. SVM and KNN algorithms produce results using split illnesses which, according to the dependent variables, can be either vertical or horizontal. Nonetheless, a decision tree is a structure that is built on the choices made in each tree and has the appearance of a tree having leaves, branches, and roots. The decision tree also provides an explanation for the values of the characteristics in the dataset. They also employed the Cleveland dataset. The information set is split into testing and training stages using certain methods. The ML algorithm, which uses these ML approaches' accuracy, is utilized for segmentation. The heart disease dataset is considered a strong fit because it is also influenced, irregular, and intricate, and it can manage this type of data.

For information training and evaluation, Aditi Gavhane et al.'s suggested study "Prediction of Heart Disease Using Machine Learning" employs a perceptron neural network with multiple layers technique. In this technique, amongst both layers of input and output, here will likely be a single input layer, one output layer, and perhaps more hidden layers. Hidden layers link every single input node to the result layer. This link has arbitrary weights allocated to it. Bias is an additional input, and its amount of weight is determined by the requirements of the terminals' link.

Golande, Avinash, and others suggested In "Heart Disease Prediction Using Efficient Data Mining and Machine Learning Approaches," a number of data mining approaches are employed to assist medical professionals in differentiating across various forms of cardiovascular disease. Popular techniques include Naive Bayes, Decision Trees, and K-nearest Neighbors. Other unique characterization-based processes that are applied involve packing the computation, element depth, sequential inconsequential streamlining, neural systems, straight Kernel selfarranging guiding, and SVM.

III. APPROACH METHODOLOGY

The initial stage of how the system works is the gathering of information and the identification of the most important characteristics. After that, the pertinent data has been processed into the required format. The data is then divided into data for testing and training. The model is trained using the algorithms and the initial training data. The accuracy of the framework is ascertained by subjecting it to evaluation using test data.

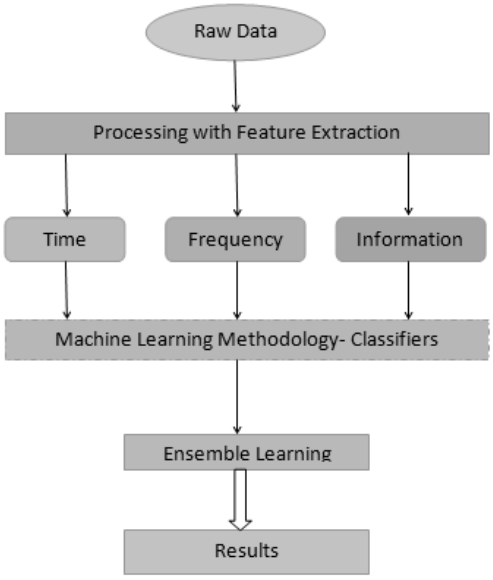

Figure 1: Workflow of Cardiovascular Disease Detection

Classification Algorithms

Using already-existing data, classifications is a supervised learning method that forecasts the result. This research suggests a classification algorithm-based method for the identification of coronary artery disease. Single models are trained using the data used for training, which has been split into a training set and a test set in a proportion of 80 to 20. Using the test information set, the classifiers' efficacy is evaluated. The section that follows provides an explanation of how each classifier operates.

1. *Logistic Regression*

A supervised learning classified approach called logistic regression (LR) is used to estimate the likelihood of an objective variable. Due to the dichotomy structure of the reliant or goal factor, there are only two potential classes: 0 for loss and 1 for victory.

2. *Naïve Bayes*

The supervised algorithm used is the Naïve Bayes classifier. It is a straightforward classification method based on the Bayes theorem. Neutrality among qualities is assumed. A statistical idea called the Bayes theorem is utilized to determine the odds. There is never a connection nor a relationship between the variables. To optimize the likelihood, each attribute adds to its separately.

The phrase "naïve bayes classifier" refers to a basic probabilistic classifier that applies the Bayes Theorem under robust independent conditions. It is predicated on the idea that the existence or nonexistence of a certain class characteristic is independent of the existence or nonexistence of any other feature [11]. The foundation of the Naive Bayes method is conditional probability. It makes use of the Bayes theorem, a method that counts the frequency of values and value permutations in the past information to arrive at the odds. The Bayes Theorem calculates the likelihood of an event based on the likelihood of an earlier event. The Bayes theorem can be expressed as follows if A is the previous event and B is the conditional event.

Prob (B given A) = Prob(A and B)/Prob(A)

The procedure calculates the total number of situations where the two events occur together and divides the result by the quantity of situations in which A happens alone to determine the likelihood of B with A. The Naive Bayes classifier has the benefit of only needing a little quantity of training data to estimate the parameters (variable means and variances) needed for classification. Only the variances of the variables for each class—rather than the complete set—need to be calculated because independent factors are assumed. The two binary and multi-class classification issues can be solved using it.

3. *Support Vector Machine*

In multilayer distance, a model created using SVM is essentially an illustration of various classes on a hyperplane. The SVM will generate the hyperplane iteratively in order to decrease the mistake. SVM seeks to determine a maximum marginal hyperplane (MMH) by classifying the data points.

Developed by Corinna Cortes and Vladimir Vapnik, the Support Vector Machine (SVM) concept is utilized in the fields of statistics and computer science to refer to a family of similar supervised learning techniques for regression analysis and classification that examine data and identify patterns. SVM has demonstrated excellent efficiency in a variety of usage domains. It builds a hyperplane or group of hyperplanes in an infinite or high-dimensional space that can be applied to regression, classification, and other tasks. In [14] SVMs are incredibly helpful for classifying data. SVMs use an ideal hyperplane to divide d-dimensional data into two classes with a maximum interclass margin in order to recognize the results. SVMs cast data into a higher dimensional space where it is separable by using known kernel operations. 15 and 16 In machine learning, classifying data is a typical problem. Assume that each of the two categories represented by the given data points must be chosen in order to determine which class a new data point will belong to. A data point is considered a p-dimensional vector (a list of p numbers) in the context of support vector machines. Our

goal is to determine if we can divide such points using a (p − 1)-dimensional hyperbolic plane. We refer to this as ordinal classifiers. 17]. Plotting the training variables in a space with high dimensions and labeling each vector according to its class, the SVM is a learning machine. 18] SVM depends on the risk-minimization concept, which attempts to reduce the rate of mistakes. [20], [19] Supervised learning is the method used by SVM to categorize data. In other words, SVM creates a model based on training data that is used to forecast test data's target values. To conduct classification, SVM needs to solve a particular optimization problem given a tagged training set (xi,yi) [22].

4. *K-Nearest Neighbour*

One technique for supervised classification is the K-Nearest Neighbor algorithm. Objects are categorized based on their closest neighbor. This kind of training is instance-based. The Euclidean distance is used to calculate an attribute's distance from its neighbors. It employs a set of designated points and applies them to the marking of an additional point. Based on how closely they resemble one another, the data are grouped. The K-NN algorithm can be easily implemented without requiring the creation of a model or additional presumptions. This approach is flexible and can be applied to search, regression, and classification tasks. K-NN is the simplest method; however its reliability is impacted by characteristics that are unnecessary and noise.

5. *Decision Tree Algorithm*

Decision trees are a type of supervised learning approach that are commonly used for solving classification problems, while they can also be used to address regression problems. It is a tree-structured classifier, with each leaf node representing the categorization outcome, its internal nodes representing the features of a dataset, and branches representing the process of making a choice. A decision tree consists of two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make actions and have many branches, in contrast to Leaf nodes, which are the outcomes from choices and have no more branching. The test is run or decisions are made based on the highlights of the provided dataset. It is an animated representation of every possible reaction, given the situation, to a choice or problem. It starts with the node that represents the root and expands on consecutive branch to produce an arrangement like a tree, just like a real tree. For this reason, it is called a decision tree. A tree is built using the CART algorithm, which stands for Classification and Regression Tree algorithm.

The biggest indicators are used by this technique to divide the data into multiple comparable groupings. After calculating each attribute's entropy, the data are split into models with the most knowledge gained or the lowest entropy:

$$E(S) = \sum_{i=1}^{c} p_i \, log_2 \, p_i$$

$$IG\ (Y, X) = E(Y) - E(Y|X)$$

The outcomes are simpler to read and understand. Because it analyzes the dataset in a tree-like graph, this method performs more accurately compared to alternative techniques. To make decisions, just one characteristic is examined at a time, and the data could be over classified.

6. *Random Forest*

An computational method for supervised classification is the random forest algorithm. A forest is created by several trees in this process. In a random forest, every single tree expresses class anticipation, and the class that receives the greatest number of votes becomes the model's final forecast. The accuracy of the random forests classification increases with tree count. It can handle values that are missing and perform well in classification tasks, but it is also utilized for regression tasks. In addition, because it needs bigger data sets along with more trees, it takes longer to get recommendations and yields inexplicable outcomes.

7. XGBoost

An optimized distributed gradient model with a focus on portability, flexibility, and high efficiency is called XG Boost. It is an aggregation machine learning technique with a decision tree foundation that makes use of the gradient boosting framework. Through parallel processing, tree pruning, handling missing variables, and periodicity to prevent excessive fitting or bias, it offers an efficient gradient boosting approach.

IV. CONCLUSION

Machine learning is crucial to the study of disease prediction. This work forecasts cardiac illness using a variety of machine learning techniques. We have examined the classifier techniques, including Random Forest, XGBoost, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, and Logistic Regression. With the use of machine learning techniques, early detection of the issue could preserve lives and prevent coronary artery disease by taking preventative measures. This paper's primary goal is to give readers with an understanding of prediction models for the heart failure under consideration, this will be useful when applying DT and GNB in medical fields. We come to the conclusion that machine learning techniques work effectively for coronary artery disease.

REFERENCES

[1] Tao R., Zhang S., Huang X., et al. Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. IEEE Transactions on Biomedical Engineering . 2019;66(6):1658–1667. doi: 10.1109/tbme.2018.2877649.

[2] Mohan S., Thirumalai C., Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access . 2019;7 doi: 10.1109/access.2019.2923707.81542

[3] Spencer R., Thabtah F., Abdelhamid N., Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digital Health . 2020;6 doi: 10.1177/2055207620914777.2055207620914777

[4] Fitriyani N. L., Syafrudin M., Alfian G., Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access . 2020.

[5] Mienye I. D., Sun Y. Improved heart disease prediction using particle swarm optimization based stacked sparse autoencoder. Electronics . 2021.

[6] Javeed A., Zhou S., Yongjian L., Qasim I., Noor A., Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access . 2019.

[7] Wankhede J., Kumar M., Sambandam P. Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO. IET Systems Biology . 2020;14(6):380–390.

[8] Wang J., Liu C., Li L., et al. A stacking-based model for non-invasive detection of coronary heart disease. IEEE Access . 2020;

[9] Zhenya Q., Zhang Z. A hybrid cost-sensitive ensemble for heart disease prediction. BMC Medical Informatics and Decision Making . 2021.
[10] Khan M. A., Algarni F. A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access . 2020.
[11] Alarsan F. I., Younes M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. Journal of Big Data . 2019.
[12] Ali L., Rahman A., Khan A., Zhou M., Javeed A., Khan J. A. An automated diagnostic system for heart disease prediction based on ${\chi^{2}}$ statistical model and optimally configured deep neural network. IEEE Access . 2019.
[13] Kendale S, Kulkarni P, Rosenberg A D and Wang J 2018 Supervise machine learning predictive analytics for prediction of postinduction hypotension Anaesthesiology 129 675-88.
[14] Ajit Solanki, Mehul P. Barot 2019 Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-2S2.
[15] Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD 1997, 3, 34–39.
[16] Maas, A.H.; Appelman, Y.E. Gender differences in coronary heart disease. Neth. Heart J. 2010, 18, 598–602.
[17] Bhunia, P.K.; Debnath, A.; Mondal, P.; D E, M.; Ganguly, K.; Rakshit, P. Heart Disease Prediction using Machine Learning. Int. J. Eng. Res. Technol. 2021, 9.
[18] Mohanty, M.D.; Mohanty, M.N. Verbal sentiment analysis and detection using recurrent neural network. In Advanced Data Mining Tools and Methods for Social Computing; Academic Press: Cambridge, MA, USA, 2022; pp. 85–106.
[19] Menzies, T.; Kocagüneli, E.; Minku, L.; Peters, F.; Turhan, B. Using Goals in Model-Based Reasoning. In Sharing Data and Models in Software Engineering; Morgan Kaufmann: San Francisco, CA, USA, 2015; pp. 321–353.
[20] Fayez, M.; Kurnaz, S. Novel method for diagnosis diseases using advanced high-performance machine learning system. Appl. Nanosci. 2021.
[21] Hassan, C.A.U.; Iqbal, J.; Irfan, R.; Hussain, S.; Algarni, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors 2022, 22, 7227.
[22] Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. Sustainability 2022, 14, 14208.

