An Assessment on Cardiovascular Disease Prediction and Diagnosis using Machine Learning Algorithms

Reddy Anuradha
Assistant Professor,
Department of CSE,
Malla Reddy Institute of Technology & Science,
Hyderabad, Telangana
anuradhareddy.anu@gmail.com

***Abstract*- We live in a postmodern era, and our everyday lives are undergoing significant changes that have a beneficial and negative impact on our health. As a result of these developments, the prevalence of numerous diseases has skyrocketed. The diagnosis of cardiovascular disease is the most challenging task in medicine. Cardiovascular disease diagnosis is complex because it relies on the grouping of enormous amounts of clinical and pathological data. As a result of this issue, there has been a substantial surge in interest among researchers and clinical experts in the efficient and precise prediction of cardiac disease. When it comes to heart disease, getting a proper diagnosis at an early stage is crucial because time is a crucial issue. Heart disease is the leading cause of mortality worldwide, and predicting heart disease at an early stage is crucial. In recent years, machine learning has emerged as one of the most progressive, dependable, and supporting tools in the medical arena, providing the most help for disease prediction with proper training and testing. This study work aims to present a survey of knowledge discovery strategies in databases employing data mining techniques that are already in use in medical research, specifically in Cardiovascular Disease Prediction.**

Index terms: - cardiovascular, machine learning, heart disease, prediction, classification

I. INTRODUCTION

According to WHO (World Health Organization) data, cardiovascular illnesses (CVDs) are the leading cause of death worldwide, with more people dying each year from CVDs than from any other cause. CVDs claimed the lives of 17.9 million people worldwide in 2016, accounting for 31% of all deaths. Heart attacks and strokes account for 85 percent of these deaths. Low- and middle-income nations account for about three-quarters of CVD mortality. In 2015, 82 percent of the 17 million early deaths (before the age of 70) owing to non-communicable illnesses occurred in low- and middle-income countries, with CVDs accounting for 37 percent. The majority of cardiovascular illnesses can be avoided by addressing behavioral risk factors like cigarette use and bad eating habits. Using population-wide initiatives, food and obesity, physical inactivity, and problematic alcohol use can be addressed [1]. The heart is a vital organ in the human body. The effective functioning of the heart is essential to life. If the heart isn't performing properly, it will have an impact on other sections of our body, such as the kidneys and the brain. The functioning of the heart is used to predict cardiac disease. The following are a few of the factors that are used to predict heart disease. • High blood pressure • Cholesterol • Lack of physical activity • Smoking • Obesity • Heart disease in the family

Because heart disease is the leading cause of death in humans, it is necessary to make forecasts in order to lower the risk of heart disease. Generally, doctors will diagnosis heart disease based on the symptoms and physical examination of the patient body. In the healthcare business, predicting heart disease is a difficult challenge. Patients' data, disease diagnoses, computerized patient records, and medical gadgets all make up a large part of the healthcare industry today [2]. It is a crucial resource that must be analyzed during knowledge extraction and will aid decision-making.

Congenital, coronary, and rheumatic heart illnesses are all covered under the umbrella term "heart disease." Coronary heart disease is the most frequent of these illnesses, with over 360,000 Americans dying from heart attacks in 2015. A heart attack is expected to occur every 40 seconds in the United States, according to the Centers for Disease Control and Prevention [18]. As a result, annual spending on heart disease in the United States has climbed to more than $200 billion. Furthermore, according to the American Heart Association, health-care costs related to heart disease are predicted to treble by 2030.

The following are some of the most common heart diseases:

- Angina
- Acute coronary syndrome
- Arrhythmia
- Cardiomyopathy
- Congenital heart disease
- Coronary artery disease
- Rheumatic heart diseases

Prediction is an excellent methodology in healthcare settings where doctors lack more information and experience, as well as where there are no specialists. For example, such clinicians may make their own decisions, which may lead to bad outcomes and the death of patients. Prediction of heart disease is utilized for automatic disease diagnosis and to provide enough services in healthcare centers to save people's

lives. The use of a prediction technique assists stakeholders, particularly experts, in making accurate decisions on how to treat patients. Predicting cardiovascular disease is a difficult undertaking, and achieving an automated diagnosis of illness is even more difficult. Because healthcare facilities keep vast amounts of data that are extremely complicated and difficult to analyze [3]. Even if it is a difficult undertaking, employing cardiac disease prediction in medical centers plays an important role in saving people's lives and allowing stakeholders to make active and accurate decisions[23]. Medical data mining has been useful in uncovering hidden patterns in large data sets in the medical field [4]. Clinical diagnosis can be made using these patterns. The accessible raw medical data, on the other hand, is widely dispersed, varied, and large. These data must be collected in a systematic manner before being integrated into a hospital information system [5]. Data mining is a user-friendly way to discovering new and hidden patterns in data [13, 22].

II. Review of the Literature

Data on numerous health-related concerns is collected by medical institutions all over the world. These data can be used to gain meaningful insights utilizing a variety of machine learning techniques. However, the amount of data collected is enormous, and it is frequently noisy. Machine learning approaches can quickly investigate these datasets, which are too large for human minds to understand. As a result, machine learning algorithms have recently proven to be quite beneficial in precisely predicting the presence or absence of heart-related disorders. To automate the examination of big and complicated data, machine learning methods and techniques have been used to a variety of medical datasets [19]. Several machine learning algorithms have recently been used by many researchers to aid the health care industry and experts in the detection of heart-related disorders [6, 20].

[7] Satish Chandra Reddy, N. Satish Chandra Reddy, N. Satish Chandra Reddy, N. Satish Chandra Reddy, N. This research focuses on the categorization and feature selection needed for employing various machine learning methods to forecast heart disease. KNN, SVM, Random Forest, Nave Bayes, and Neural Network are among the algorithms used in the paper. Over the total dataset, the methods utilized in the research produce a better outcome, with an average accuracy of 85.92-89.41%.

Marjia et al.[8] propose employing WEKA software to predict cardiac disease using K Star, J48, SMO, and Bayes Net and Multilayer Perceptron. Based on the performance of several factors, SMO (89 percent accuracy) and Bayes Net (87 percent accuracy) outperform KStar, Multilayer Perceptron, and J48 approaches utilizing k-fold cross validation. The accuracy levels reached by those algorithms are still insufficient. As a result, if the accuracy of the diagnosis is increased, a better judgement can be made.

The Azam and his associates .[9] The research describes automatic diagnosis of coronary artery disease (CAD) patients using optimized SVM, in which SVM parameters are optimized to improve prediction accuracy, yielding a 99.2% accuracy using k-fold cross-validation. The paper aids in the early detection of disease and the reduction of costs. The accuracy attained is good for predicting whether or not a person has heart disease.

Cemil et al. [10] propose using a knowledge discovery process to predict stroke patients using Artificial Neural Networks (ANN) and Support Vector Machines (SVM), which have accuracy of 81.82 percent and 80.38 percent for ANN and SVM, respectively, in the training data set and 85.9% and 84.26 percent for Artificial Neural Network (ANN) and Support Vector Machine (SVM) in the test dataset. For the suggested work, ANN produces better accurate results than Support Vector Machine (SVM). The paper's accuracy is insufficient to accurately predict stroke patients.

Sanavar et al. [11] are a group of researchers who came up with a novel way to solve a problem. The following is a summary of a survey article on the prediction of heart disease. It explains the various methodologies as well as how the proposed approaches are put into action. It also gives an overview of heart disease, the function of data mining in healthcare, and how to apply or use data mining in a healthcare setting.

Megha Shahi and her colleagues .[12] The goal of this study is to develop a system for predicting heart disease utilizing data mining techniques and WEKA software for automatic illness detection and to provide service quality in healthcare centers. SVM, Nave Bayes, Association rule, KNN, ANN, and Decision Tree were among the algorithms utilized in the study. According to the article, SVM has an effective and efficient accuracy of roughly 85% when compared to other data mining methods.

III. Basis of Machine Learning Algorithms

A. Support Vector Machine

Support vector classifiers are classified as linear, nonlinear, radial basis function (RBF), sigmoid, or polynomial based on their kernel functions [14]. The support vector or data points are separated by the hyperplane or support vector machine[15]. Binary SVMs are classifiers that distinguish between data points belonging to two categories. An n-dimensional vector represents each data object (or data point) [16]. Each of these data points can only be classified into one of two groups. A hyper plane is used to separate them by a linear classifier. There are numerous hyper planes that can be used to separate data samples. SVM chooses the hyper plane with the biggest margin to obtain the greatest separation between the two classes. The margin is the sum of both categories' shortest distances from the separating hyper plane to the closest data point. A hyper plane like this is more likely to generalize, which means it will correctly identify unseen or testing data points. To handle nonlinear classification problems, SVM performs the mapping from input space to feature space.

B. Random Forest

Random forest creates a large number of decision trees during training and outputs the mean forecast for regression and the mode prediction for classification. The decision trees examine the many patterns in the data. The classification forecast is based on the majority vote.

C. K-Nearest Neighbor's

It's a simple classifier that can't handle sounds, it's straightforward to design and comprehend, it takes a little amount of time to train, and it uses the entire training set for prediction. Heart disease has been predicted using the K-Nearest Neighbors (K-NN) method with a weighting value. For the heart, comparable algorithms (KNN) were combined with feature selection techniques like as particle swarm optimization (PSO). When it comes to disease prediction, it's significantly more accurate.

D. Decision Tree

For inductive inference, decision tree learning is one of the most commonly used and useful machine learning approaches. It's a method for approximating discrete valued functions that can learn disjunctive formulations and is robust to noisy data. A decision tree is used to represent this learning function. Classification trees are tree models with a finite set of values for the goal variable. Leaves indicate classification outcomes, and branches represent feature combinations that point to those classification results in these tree topologies.

E. Naive Bayes

The Bayes theorem is used to create a classification algorithm called Naive Bayes. The Naive Bayes model is simple to construct and is especially good for huge data sets. Naive Bayes is renowned to outperform even the most advanced classification systems due to its simplicity. It is assumed that the existence or absence of a specific class feature has no bearing on the presence or absence of any other feature. It's based on the concept of conditional probabilities. The Naive Bayes classifier has the advantage of using only a little quantity of training data to estimate the parameters (variable means and variances) required for classification. Because independent variables are assumed, all that needs to be established are the variances of the variables for each class. It can be used to solve problems involving binary and multiclass categorization.

IV. CONSIDERATIONS

The assessment field has been inundated by statistical models for estimating that are incapable of producing good performance results. Statistical models fail to store categorical data, handle missing values, or deal with massive data points. All of these factors contribute to the importance of MLT [17]. Many applications, such as image detection, data mining, natural language processing, and illness diagnostics, rely on machine learning. ML has potential solutions in all of these areas. This study presents a review of various machine learning algorithms for diagnosing diseases such as heart disease, diabetes, liver disease, dengue fever, and hepatitis. It has been discovered that for the SVM improves the accuracy of heart disease identification. The benefits and drawbacks of various algorithms are highlighted in this survey. Machine learning techniques for disease prediction improvement graphs. Based on the results of the investigation, it is obvious that these algorithms improve the accuracy of different diseases, allowing for better decision-making.

Based on the foregoing research, it can be stated that machine learning algorithms have a lot of potential in predicting cardiovascular or heart-related disorders. Each of the algorithms listed above performed exceptionally well in some circumstances but poorly in others, possibly due to overfitting. Because they incorporate numerous algorithms, such as multiple Decision Trees in the case of Random Forest, Random Forest and Ensemble models have done exceptionally well. Machine learning algorithms and methodologies have proven to be quite accurate in predicting cardiac problems.

Using a pooled dataset, the cardiac disease prediction was tested using the classification and feature selection algorithms developed in the CARET package of the R tool. The database, preprocessing, analytical tools, and procedures all influence the model's accuracy. When compared to using all of the dataset's features, it's critical to choose the bare minimum and prominent properties to optimize speed. Random forest has the most accuracy in a three-percentage split (without and with feature selection). The random forest may be utilized as an excellent classification method for accurately predicting heart disease with an accuracy of 90–95 percent, according to this study. The fact that the accuracy differences between the dataset and selected characteristics (8 and 6) are less variable suggests that these features may be beneficial for heart disease prediction [5].

Heart disease diagnosis is complex because it relies on the grouping of enormous amounts of clinical and pathological data [25]. This work proposes utilizing a genetic approach to investigate several heart disease prediction models and identify relevant heart disease features. Different prediction models were investigated in this study, and trials were undertaken to determine the best classifier for predicting heart disease. For the prediction of patients with cardiac illnesses, four classifiers were used: Random Forest, Nave Bayes, Decision Tree, and Support Vector Machine. The performance of the Naive Bayes classifier is more accurate in most circumstances, according to observation. The genetic algorithm feature selection technique suggests the most significant features for heart illnesses, according to another finding from this study. The results also suggest that combining a genetic algorithm with prediction models increases the models' performance.

V. ILLUSTRATION OF THE PROPOSED SYSTEM

Massive amounts of data generated by a variety of sources have become critical for gathering, storing, searching, and sharing, but they are difficult to comprehend and analyze. Because of the large amount of data and the rising expense of diagnosis, researchers have been looking for ways to improve the model's accuracy and provide better disease prediction results.

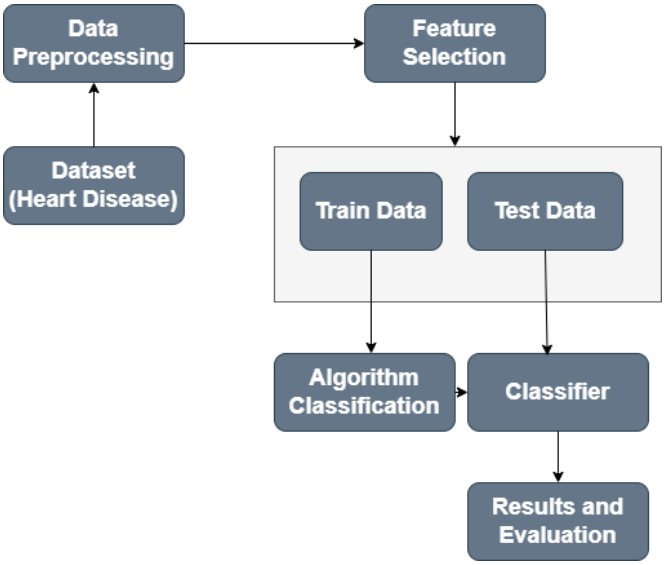

Fig 1. A model for predicting CVD has been proposed

A. Data Preprocessing

Data preprocessing is a data mining approach that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific behaviours or trends, as well as including numerous inaccuracies. Preprocessing data is a tried and true means of resolving such problems. Data preprocessing is the process of preparing raw data for subsequent processing.

Missing values are replaced with mode values based on the particular data set during the data preparation stage. Source of data Second, outliers in the dataset (i.e., heart disease patients with high values of corresponding variables) are not deleted.

B. Feature Selection

Feature selection is a key stage in data preprocessing since it allows for the removal of unneeded features and the improvement of performance in order to develop a better classification model [21]. The feature selection is carried out on the dataset in order to pick a subset of relevant characteristics for model construction, with the goal of improving model accuracy. In data mining, feature selection is an effective data preparation strategy for reducing data dimensionality. It is critical in medical diagnosis to identify the most significant disease-related risk factors [24]. Relevant feature identification aids in the removal of redundant and unneeded attributes from the illness dataset, resulting in faster and more accurate findings [26].

C. Classification

The dataset with its attributes is segregated into training and testing data after data normalization to develop a classification model. Classification and prediction is a data mining process in which training data is used to construct a model, which is then applied to testing data to obtain prediction results. On illness datasets, various classification methods such as K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Random Forest, and Nave Bayes have been used to diagnose disease. The development of a revolutionary categorization system that can speed up and simplify the process of disease diagnosis is critical. Accuracy, sensitivity/recall, and specificity are used to measure a model's performance on test data. True positives (risk class) and true negatives (normal class) are measured by sensitivity and specificity, respectively. As a result, sensitivity and specificity values are used to assess the classifiers' prediction ability.

VI. CONCLUSION

According to the knowledge of reviewed literature, heart disease is one of the leading causes of death worldwide. Cardiovascular disease refers to a set of disorders that affect the heart and blood arteries. Heart disease diagnosis is very difficult since it is based on the grouping of enormous amounts of clinical and pathological data. The major goal of this work is to identify various cardiovascular disease (CVD) prediction models and choose important disease features using a machine learning algorithm. Risk variables such as hypertension and family history must be considered as predictors in the proposed study, and selected features must be used for accurate heart disease prediction. Various classification techniques are used to predict the presence and absence of cardiac disease. The accuracy, sensitivity, and specificity of the prediction models are used to evaluate their performance.

REFERENCES

[1] Krishna, N. M., Sekaran, K., Vamsi, A. V. N., Ghantasala, G. P., Chandana, P., Kadry, S., ... & Damaševičius, R. (2019). An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using EEG signals. IEEE Access, 7, 77905-77914

[2] Patan, R., Ghantasala, G. P., Sekaran, R., Gupta, D., & Ramachandran, M. (2020). Smart healthcare and quality of service in IoT using grey filter convolutional based cyber physical system. Sustainable Cities and Society, 59, 102141

[3] Chandana, P., Ghantasala, G. P., Jeny, J. R. V., Sekaran, K., Deepika, N., Nam, Y., & Kadry, S. (2020). An effective identification of crop diseases using faster region based convolutional neural network and expert systems. International Journal of Electrical and Computer Engineering (IJECE), 10(6), 6531-6540.

[4] Reddy, A. R., Ghantasala, G. P., Patan, R., Manikandan, R., & Kallam, S. Smart Assistance of Elderly Individuals in Emergency Situations at Home. Internet of Medical Things: Remote Healthcare Systems and Applications, 95

[5] MANDAL, K., GHANTASALA, G. P., KHAN, F., SATHIYARAJ, R., & BALAMURUGAN, B. (2020). Futurity of Translation Algorithms for Neural Machine Translation (NMT) and Its Vision. Natural Language Processing in Artificial Intelligence, 53

[6] Mr. Chala Beyene, Pooja Kamat "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques" International Journal of Pure and Applied Mathematics , ijpam Volume 118 No. 8 ,2018, 165-174

[7] P. Suresh and M.D. Ananda Raj "Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm" International Journal of Pure and Applied Mathematics , ijpam, Volume 119, No. 16, 2018, 5323-5336

[8] A. Davari Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," Comput. Methods Programs Biomed., vol. 138, 2017,pp. 117–126

[9] C. Colak, E. Karaman, and M. G. Turtay, "Application of knowledge discovery process on the prediction of stroke," Comput. Methods Programs Biomed., vol. 119, no. 3,2015, pp. 181–185

[10] M. Gandhi, "Predictions in Heart Disease Using Techniques of Data Mining," Int. Conf. Futur. Trend Comput. Anal. Knowl. Manag., 2015

[11] U. R. Acharya et al., "Application of higher- order spectra for the characterization of Coronary artery disease using electrocardiogram signals," Biomed. Signal Process. Control, vol. 31,2017, pp. 31–43

[12] M. Shahi and R. Kaur Gurm, "Heart disease prediction system using data mining techniques," Orient. J Comput. Sci. Technol., vol. 6, no. 4, 2013,pp. 457–466

[13] G S Pradeep Ghantasala, D. Nageswara Rao, Mandal K (2021) MACHINE LEARNING ALGORITHMS BASED BREAST CANCER PREDICTION MODEL. Journal of Cardiovascular Disease Research, 12 (4), 50-56. doi:10.31838/jcdr.2021.12.04.04

[14] Kumari, N. V., & Ghantasala, G. P. (2020). Support Vector Machine Based Supervised Machine Learning Algorithm for Finding ROC and LDA Region. Journal of Operating Systems Development & Trends, 7(1), 26-33

[15] "A survey on Microcalcification identification and classification using CAD System", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.2, Issue 5, page no.186-190, MAY-2015

[16] G. S. Pradeep Ghantasala, Nalli Vinaya Kumari. Mammographic CADe and CADx for Identifying Microcalcification Using Support Vector Machine. Journal of Communication Engineering & Systems. 2020; 10(2): 9–16p

[17] Ghantasala, G. P., & Kumari, N. V. (2021). Identification of Normal and Abnormal Mammographic Images Using Deep Neural Network. Asian Journal For Convergence In Technology (AJCT), 7(1), 71-74

[18] Ghantasala, G. P., & Kumari, N. V. (2021). Breast Cancer Treatment Using Automated Robot Support Technology For Mri Breast Biopsy. INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS, 1(2), 235-242

[19] Ghantasala, G. P., Reddy, A., Peyyala, S., & Rao, D. N. (2021). Breast Cancer Prediction In Virtue Of Big Data Analytics. INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS, 1(1), 130-136

[20] Ghantasala, G. P., Reddy, A. R., & Arvindhan, M. Prediction of Coronavirus (COVID-19) Disease Health Monitoring with Clinical Support System and Its Objectives. In Machine Learning and Analytics in Healthcare Systems (pp. 237-260). CRC Press

[21] Ghantasala, G. P., Kallam, S., Kumari, N. V., & Patan, R. (2020, March). Texture Recognization and Image Smoothing for Microcalcification and Mass Detection in Abnormal Region. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE

[22] Bhowmik, C., Ghantasala, G. P., & AnuRadha, R. (2021). A Comparison of Various Data Mining Algorithms to Distinguish Mammogram Calcification Using Computer-Aided Testing Tools. In Proceedings of the Second International Conference on Information Management and Machine Intelligence (pp. 537-546). Springer, Singapore

[23] Sreehari, E., & Ghantasala, P. G. (2019). Climate Changes Prediction Using Simple Linear Regression. Journal of Computational and Theoretical Nanoscience, 16(2), 655-658

[24] Kishore, D. R., Syeda, N., Suneetha, D., Kumari, C. S., & Ghantasala, G. P. (2021). Multi Scale Image Fusion through Laplacian Pyramid and Deep Learning on Thermal Images. Annals of the Romanian Society for Cell Biology, 3728-3734

[25] Ghantasala, G. P., Kumari, N. V., & Patan, R. (2021). Cancer prediction and diagnosis hinged on HCML in IOMT environment. In Machine Learning and the Internet of Medical Things in Healthcare (pp. 179-207). Academic Press

[26] Ghantasala, G. P., Tanuja, B., Teja, G. S., & Abhilash, A. S. (2020). Feature Extraction and Evaluation of Colon Cancer using PCA, LDA and Gene Expression. Forest, 10(98), 99