2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-9737-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACCS57279.2023.10112952

An Effective Heart Disease Prediction Method using Extreme Gradient Boosting Algorithm Compared with Convolutional Neural Networks

Cuddapah Anitha
Computer Science and Systems Engineering
Sree Vidyaniketan Engineering College,Mohan Babu University,
Tirupati, Andhra Pradesh
dranithacuddapah17@gmail.com

S Rajkumar
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai
rajkumar.srinivasan@vit.ac.in

Dhanalakshmi R
Department of Computer Science and Engineering,
Saveetha School of Engineering,
SIMATS, Chennai, Tamil Nadu,
India-602105.
dhanalakshmir.sse@saveetha.com

***Abstract*—Heart disease is one of the most serious and common problems to human beings around the face;% to 90% of people are affected by heart problems in which, and Several methods exist to identify heart disease in an earlier stage, but few methods fail to manage the accuracy due to the large volume of data. In addition, the methods require more time to process those features. To overcome these research issues, effective intelligent techniques are proposed to predict heart disease in an earlier stage. Additionally, our model's balanced accuracy (89.5%) outperforms the XGBoost classifiers' accuracy rates. Excellent recall values and area under the curve, together with high specificity and test accuracy, are obtained using the CNN classifier. Designers contrast the XGboost, KNN, SVM, and Naive Bayes Classifiers' classification precision on various healthcare datasets. Finally, this leads to a reduction in the overall error rate and also improves the recognition accuracy by more than 90%.**

Keywords—Heart disease; Feature Selection; convolutional neural network; Classification; Optimization; Neural Network

I. INTRODUCTION

Heart problems have significantly impacted the world. A prevalent form of heart illness is risk factors for heart disease. A buildup of plaque in the coronary artery walls brings it on. The responsibility for supplying coronary arteries receives blood from the heart and other internal organs. Chest pain and discomfort are typical signs of cardiac disease.

Heart attacks might occasionally be the disease's earliest warning indication. Weakness, dizziness, nausea, a cold sweat, arm pain, and shortness of breath come along with this. A family history of the condition, being overweight, not exercising, eating poorly, smoking, and other factors are among the main causes of this illness. Heart illness can result in heart failure and cause the patient to die if it is not adequately treated promptly.

A computational model that resembles the human brain is called a neural network. It is a group of neurons or nodes. These nodes are arranged in layers, with each layer's neurons processing input and sending output to the layer above it.

Different layers might carry out various modifications. Through various hidden layers, data is transferred between the input layer and the output layer (final layer). One of the most popular techniques for CNN [1] has improved the categorization precision of heart disease data sets. An ANN architecture is depicted in Fig. 1.

A neural network is capable of producing effective classification rules. A multilayer perceptron with a specific architecture for recognising two-dimensional data information is the basis of the convolution neural network technique. Always add more layers, such as output, sample, convolution, and layers. The most successful subfield of machine learning, known as deep learning, focuses on learning levels of representations.

Healthcare product development must focus on producing high-quality outcomes at reasonable costs. Healthcare organisations are also looking for clinical testing that may be done inexpensively and without intrusion. By creating a computer-based decision support system for identifying various ailments, businesses may better meet the requirements of millions worldwide [2].

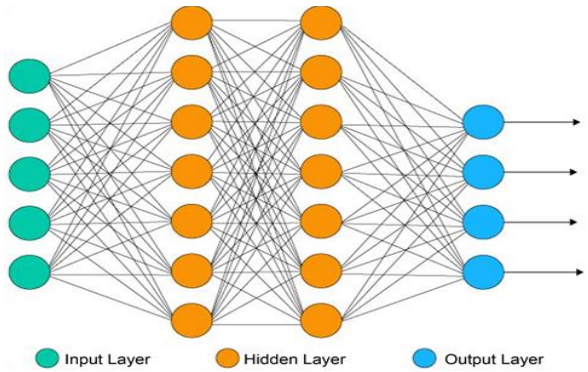

Fig. 1. Diagram of Convolutional Neural Networks (CNN)

979-8-3503-9737-6/23/$31.00 ©2023 IEEE

In this study, we develop an object classifier based on convolutional neural networks to distinguish between normal and pathological heart disease records. KNN, Naive Bayes, SVM, and ANN, four currently used data mining algorithms, are compared using our suggested approach with more accuracy.

II. Related Work

Heart disease prediction is a topic of current research for many scientists.

Awais Mehmood and others [3] Convolutional neural networks, a deep learning algorithm, is used in the proposed CardioHelp approach to forecasting a patient's likelihood of developing cardiovascular disease (CNN). The suggested method focuses on modelling temporal data by using CNN during HF prediction's first phases. The dataset was prepared, the results were examined using cutting-edge techniques, and the findings were encouraging. The proposed method performs better than the current methods in performance evaluation measures. Calculating the proposed method's accuracy yields a figure of 97 percent.

For the purpose of forecasting heart illnesses, Fazl-e-Rabbi et al. employed multiple classifiers. The Cleveland dataset, which included 270 entries and 76 attributes, was taken according to UCI archive. Set of data's 13 attributes were the only ones used in this investigation. Three various classifiers were utilised for SVM, an artificial neural network, and k-nearest neighbour are methods for heart disease prediction. SVM provided 85.18% classification accuracy. When the number of k was increased till k=10, the accuracy value using KNN grew steadily. A value of 80.74% was attained at that time for accuracy. The ANN was accurate 73.33% of the time.

Manimurugan proposed a two-stage methodology for the prediction of heart disorders using artificial intelligence and IoMT. Data was originally gathered from patients' connected medical sensors in the first step. For classifying sensor data, the HLDAMALO hybrid optimization technique was employed. The second stage involved an echocardiography. R-CNN hybrid was utilised for picture categorization. The Columbus dataset is available in the UCI machine learning repository, which has Fourteen qualities, was also used in this study. Following the deployment referring to the model, HLDA-MALO, correctly predicted both normal and abnormal sensor data with 96.85% and 98.31% accuracy, respectively. The performance criteria were used to gauge the progress of R-CNN, and it attained, respectively, 98.06%, 98.95%, 96.32%, 99.02, and 99.15% for high precision, memory, recall, specificity, and F scoring system.

ElhamNikookar et al. [4] explored employing hybrid methodologies to apply principles from data science for detecting heart disease. The author combined three algorithms to create a hybrid model. The ANN classifier was utilised initially. The SVM classifier receives the results, and then the classifier using Naive Bayes receives the results to provide the forecast outcomes. The hybrid algorithm demonstrated the great accuracy of the suggested model. 88% or so has been attained.

Poornima Singh et al. [5] suggested using a NN to approach for forecast cardiac disease. 15 different characteristics were taken into account by the algorithm for prediction. The training procedure made use of a multilayer perceptual neural network with backpropagation. The model produced 100% accuracy, and the dataset was well-organized and clean.

Gomathi et al. [6] Using naive Bayes and decision tree data mining approaches, many diseases can be predicted.. They concentrated on predicting various forms of diabetes, breast cancer, and cardiovascular disease. The outcomes were provided via the confusion metrics.

Miranda et al. [7] proposed the naive Bayes classifier technique for predicting cardiovascular illnesses. The writers identified only a few significant risk variables that can determine if someone has cardiovascular disease. In 85% of the situations, the suggested notion has proven 85% accurate, 85% sensitive, and 85% specific.

III. Methodology

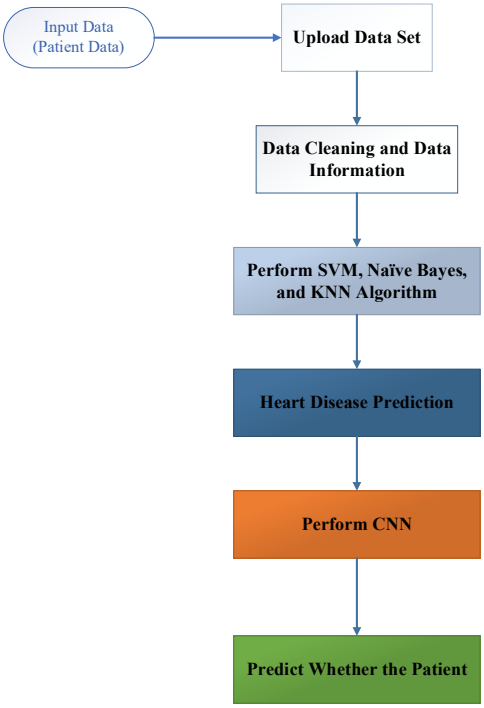

Fig. 2. Methodology for Heart-Disease Prediction

This portion of the article offers a plan for something like the quick recognition of cardiovascular illness. A variety of techniques have been used to predict heart problems. In this study, a convolutional neural network is used to forecast heart diseases (CNN). Technologies for analyzing data collected predominantly comprised of CAD procedures requires programmes that run on computers. Manually transferring

information from with a subject matter knowledgeable in computational models requires a lot of time and effort, and it mostly depending on the opinion of the expert judgement.[8] CNN, for forecasting, a computational intelligence method was employed to handle this issue successfully. The structure of the proposed technique is depicted in Fig. 2.

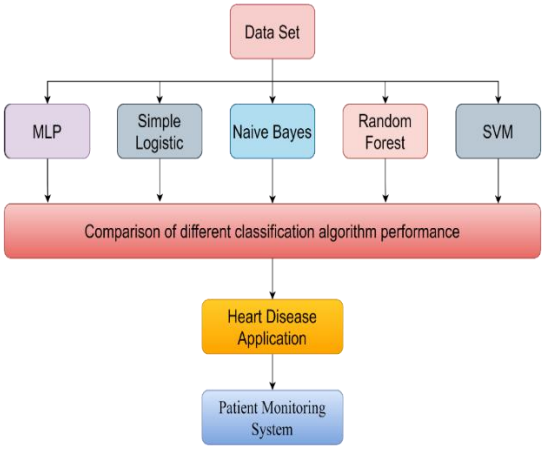

Fig. 3. The system for monitoring patients and predicting cardiac disease. [9]

The patient data collection method is the first step. After being collected, the data move onto the preparation stage. The method described in the data preprocessing section is used to complete or eliminate the missing values. Several prepossessing approaches were used to enhance the dataset. [9]The cleaned dataset was created in order to create the training and testing datasets. The model was trained using the training dataset and training data from the CNN algorithm. The effectiveness of the suggested model was assessed using the testing data. After testing and model training, the provided system provides the desired result. The suggested method places the patient in either the healthy or unwell category.

IV. System Architecture

The system architectural overview and every element, approach, and tool utilised to create the system are shown in this section. To create a method for predicting cardiac disease that is both smart and user-friendly, it is necessary to compare several machine learning algorithms and train enormous datasets with a strong software tool.

It will be decided to choose the strong algorithm with the best accuracy and performance metrics, and a programme will be developed for determining and predicting heart disease risk factors. To establish a system for ongoing patient monitoring, this is required. The block diagram for the entire system workflow is shown in Fig. 3.

1. Naive Bayes: This method for predictive modelling is remarkably effective. It is a statistical classifier that attempts to maximise the posterior probability while choosing the class while assuming no connection between attributes. Despite having the lowest error rate in theory, this classifier may not always be accurate. Assumptions result in errors due to the absence of statistical data and belongs to a particular class independence.

2. Artificial Neural Networks: Biologically inspired multilayer perceptron artificial neural networks the capacity for simulating extraordinarily nonlinear functions that are complicated. ANNs are a key tool within learning algorithms. As "neural" implies, they are systems that are focused on the brain created to replicate how people learn. The first three layers of neural networks are the hidden layers, input layer, and output layer. Components that transform the sources of information into a pattern that the output layer can manage are frequently found in a hidden layer. The use of ANNs is very effective for spotting patterns that a human programmer would find challenging to extract, let alone teach the computer to recognize.[10]

3. A technique is called a SVM for ramifying information that is both linear and non - linear. An unidirectional mapping approach is utilised to a higher dimension by transforming the data for training. A line dividing the input variable space is known as a hyperplane in SVM. The points bearing their class, either 0 or 1, can be divided by the hyperplane in the input variable space.

4. Among the most widely used well-known and effective machine learning techniques. One type of artificial intelligence algorithm is the Algorithms for Classifier and Clustering Algorithm. Bootstrapping is a powerful statistical technique for estimating a metric derived from a subset of population, the median, for instance. There are several samples of the data collected, the mean is calculated, and all of the mean values are then averaged in order to better predict the real mean value. The same technique is applied during bagging. However, decision trees are frequently employed as opposed to determining the mean of each data sample. For each training data sample taken into consideration in this case, models are constructed. Every model generates a forecast, which is required for every set of data, and these forecasts are then averaged to improve estimates in terms the measured output quantity.[11]

5. Simple Logistic Regression: Using machine learning, the science of statistics has developed the technique of LR. Using a binary system, this method can be used to divide values into two categories. The goal of LR, like that of an example of linear regression ascertain the coefficient values for each input variable. A non-linear equation called a simple regression analysis is used to create the output prediction, as opposed to linear regression. The logistic function transforms any number between 0 and 1.

V. Results

A. Data Preprocessing

Thirteen attributes are common to the dataset of all records. A few missing values were discovered during collection because clinicians manually recorded the data. Replace the missing values with the relevant data and the Replaced Missing Values filter option in the Weka 3.6.6 tool by using the mean and mode methods. To determine

classification accuracy, One gets a confusion matrix. The confusion matrix in this piece displays the number of instances assigned to each class. Dataset Attributes and Characterization is shown in table 1.

B. Dataset Attributes

TABLE 1. DATASET ATTRIBUTES AND CHARACTERIZATION

Name of Feature	Characterization
Id	A distinctive identification number
Gender	Men or Women
Age	An individual's age
heart_disease	Present: nil absentee:1
Stroke	0 or 1

C. Performance Metrics

Some performance measures could be used to evaluate how well the suggested model is performing. There are many criteria for measuring a system's performance in deep learning. The following sections address certain performance indicators, such as accuracy, sensitivity, specificity, and F-Measure.

D. XGboost

F measurement, level accuracy, accuracy, and search are all formula components, rule development, and performance level. Class A means you have cardiomyopathy; Class B means you don't ia shown in table 2.

To analyse these measurements, real negative values (TN), fake positive values (FP), false negative values (FN), and actual transaction prices (TP) are employed (FN).

$$\text{Accuracy: } p = \frac{tp}{tp+fp} \quad (1)$$

$$\text{F-Measure: } f = \frac{2pr}{p+r} \quad (2)$$

$$\text{Sensitivity: } se = \frac{tp}{tp+fn} \quad (3)$$

$$\text{Specificity: } sp = \frac{tn}{tn+fp} \quad (4)$$

$$\text{The Total Accuracy : } f = \frac{tp+tn}{tp+tn+fp+fn} \quad (5)$$

TABLE 2. CLASS A MEANS YOU HAVE CARDIOMYOPATHY; CLASS B MEANS YOU DON'T

	A	B
A	+ve	-ve
B	+ve	- ve

The 573 records in the data set have been separated into 25 classes, each comprising 23 records. To gauge the system's precision, training and test sets are divided equally. To categorise the system using the interval method and symbolic learning, i.e., (, +), 13 attributes are selected to measure the degree of certainty.

The clinical data set was used to train the classifiers to categorise it as having or not having heart illness There are two groups of five classifiers: the general and particular confusion matrices (i.e., normal and abnormal). Access matrix for confusion Google Classifier is shown in table 3.

TABLE 3. ACCESS MATRIX FOR CONFUSION GOOGLE CLASSIFIER

	A	B
A	257	05
B	08	271

The Naive Bayes, SVM, K-NN, ANN, and CNN classifiers were compared. Table 4 below shows the results of the mistake classifiers with high accuracy and low mean value. Similar to this, Table 5 provides access to the data set's overall accuracy for the unreliable K values, and Fig. 3 and 4 show their comparison graphs, respectively. Sensitivity, Specificity, and F-Measure Comparison Chart is shown in Fig 5.

TABLE 4. COMPARISON OF THE 5 CLASSIFICATIONS' ACCURACY AND MEAN VALUE DEVIATION

Classifiers	Accuracy	Mean
K-NN	56.42	0.042
Naïve Bayes	76.47	0.002
SVM	78.46	0.006
ANN	86.54	0.004
XGBoost	88.25	0.003
CNN	96.23	0.002

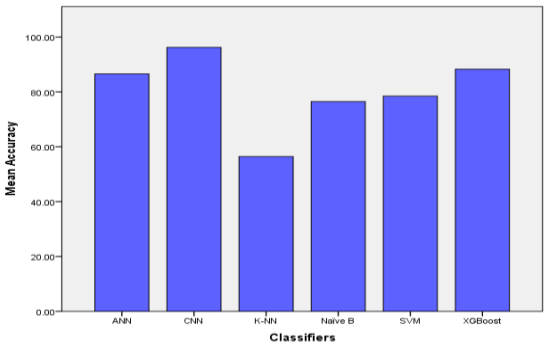

Fig 4: Comparison of Accuracy

TABLE 5. TABLE OF SENSITIVITY, SPECIFICITY, AND F-MEASURE COMPARISON

Classifiers	Sensitivity	Specificity	F-Measure
K-NN	0.9327	0.9023	0.8402
Naïve Bayes	0.9406	0.9112	0.8920
SVM	0.9521	0.9345	0.9322

ANN	0.9572	0.9416	0.9381
XGBoost	0.9602	0.9496	0.9426
CNN	0.9623	0.9535	0.9512

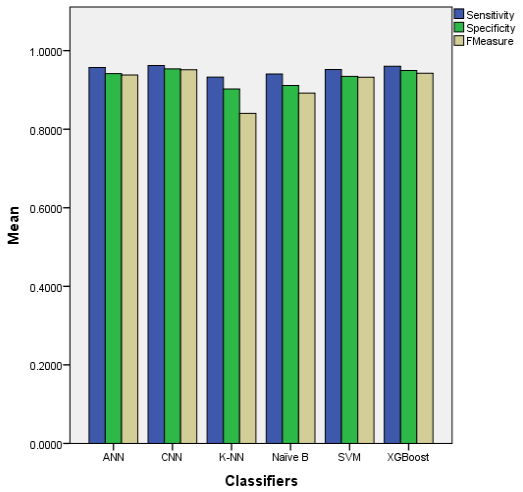

Fig 5: Sensitivity, Specificity, and F-Measure Comparison Chart

VI. CONCLUSION

The neural network classification methodology is effectively used in our proposed system. Many accidents can be avoided by early heart disease detection. The convolution technique is used to train the neural network for the classification problem of medical data. The experimentation with single and multilayer neural networks is demonstrated using the dataset for heart illness. This study's main goal is to forecast people who may develop heart disease better correctly. The dataset includes data depending the patient's hypertension, age, and gender, kind of employment, blood body mass index, blood sugar levels, etc., to forecast the likelihood that they would pass away from heart failure. The algorithms' performances have been compared using Sensitivity, Specificity, F-Measure, and Accuracy. CNN is used to get an accuracy of 96.23%. The cloud-based cardiac disease diagnosis application and database for this kind of patient monitoring application will be the main topics of future work.

REFERENCE

[1] VARKALA, KRISHNAIAH. "Heart Disease Prediction System Using Convolutional Neural Networks." (2022).

[2] Hussain, Shadab, Santosh Kumar Nanda, Susmith Barigidad, Shadab Akhtar, Md Suaib, and Niranjan K. Ray. "Novel Deep Learning Architecture for Predicting Heart Disease using CNN." In 2021 19th OITS International Conference on Information Technology (OCIT), pp. 353-357. IEEE, 2021.

[3] Mehmood, Awais, Munwar Iqbal, Zahid Mehmood, Aun Irtaza, Marriam Nawaz, Tahira Nazir, and Momina Masood. "Prediction of heart disease using deep convolutional neural networks." Arabian Journal for Science and Engineering 46, no. 4 (2021): 3409-3422.

[4] Nikookar, Elham, and Ebrahim Naderi. "Hybrid ensemble framework for heart disease detection and prediction." International Journal of Advanced Computer Science and Applications 9, no. 5 (2018).

[5] Singh, Poornima, Sanjay Singh, and Gayatri S. Pandi-Jain. "Effective heart disease prediction system using data mining techniques." International journal of nano medicine 13, no. T-NANO 2014 Abstracts (2018): 121.

[6] Gomathi, K., and Dr D. Shanmuga Priyaa. "Multi disease prediction using data mining techniques." International journal of system and software engineering 4, no. 2 (2016): 12-14.

[7] Miranda, Eka, Edy Irwansyah, Alowisius Y. Amelga, Marco M. Maribondang, and Mulyadi Salim. "Detection of cardiovascular disease risk's level for adults using naive Bayes classifier." Healthcare informatics research 22, no. 3 (2016): 196-205.

[8] Reddy, V. Archana, and K. Venkatesh Sharma. "Heart Disease Classification And Risk Prediction By Using Convolutional Neural Network." Int. J. of Aquatic Science 12, no. 2 (2021): 1973-1986.

[9] Nashif, Shadman, Md Rakib Raihan, Md Rasedul Islam, and Mohammad Hasan Imam. "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system." World Journal of Engineering and Technology 6, no. 4 (2018): 854-873.

[10] Du, N.; Cao, Q.; Yu, L.; Liu, N.; Zhong, E.; Liu, Z.; Shen, Y.; Chen, K. FM-ECG: A fifine-grained multi-label framework for ECG image classifification. Inf. Sci. 2021, 549, 164–177.

[11] Yin, X.; Goudriaan, J.; Lantinga, E.A.; Vos, J.; Spiertz, H.J. A flflexible sigmoid function of determinate growth. Ann. Bot. 2003, 91, 361–371.