Heart Disease Detection Using Machine Learning and Deep Learning

Mrs. B. Lalitha Rajeswari[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
M. Naga Nandini[2],**M. Venkata Gopi Jayaram**[3], **P. Lokesh**[4], **P. Divya Sri**[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
lalitharaji.raji@gmail.com[1], **moparthinandini@gmail.com**[2],
mvenkatagopi2000@gmail.com[3], **ponnekantilokesh654@gmail.com**[4],
podilidivya9559@gmail.com[5]

DOI:10.48047/IJFANS/V11/I12/216

Abstract

Heart is responsible for different functions like blood circulation and supplying oxygen. These days, heart disease has become one of the major causes of death of many people. It is caused by different reasons like having an unhealthy lifestyle and having high levels of blood pressure, cholesterol, and other conditions. When the patient is detected with the presence of heart disease, he could be monitored and treated to save their lives. An ensemble model is built to detect the presence of disease by using various machine learning and deep learning models. Initially, the unwanted features are removed from the data by using feature selection methods like correlation matrix and fisher score. The ML models are then trained with the data and are stacked with a meta model. The stacking model is ensembled with the deep learning models used. K-Fold cross validation technique is used to train the models. The built ensemble model gave higher accuracy of 87.2%.

Keywords: Deep Learning, Feature Selection, Heart disease detection, Machine Learning, Majority Voting.

1.Introduction

The heart's primary job is to circulate blood throughout the body so that tissues and organs may acquire the nutrition and oxygen they need while also getting rid of waste products. It preserves the body's health and performance. Heart failure, coronary artery disease and other illnesses that affect the vessels and the heart are all included in the category of cardiovascular disease. More than four in five CVD fatalities are due to heart strokes. Cardiovascular diseases are the most common cause of death globally, accounting for an approximate 17.9 million deathsannually [11]. Being an essential organ of the body, still the heart is prone to illness and damage. Its presence cannot be ignored due to the threat it brings to human life. Heart diseases are caused by different risk factors such as obesity, high levels of blood pressure, Cholesterol and diabetes, smoking, inactive lifestyle. There are various symptoms of heart disease such as Chest pain, difficulty in breathing and dizziness.

Many studies showed that different machine learning and deep learning models can be used to detect the existence of heart disease. In this work, machine learning models like SVM, LR, DT and Stacking of these models, deep learning models like FFNN, LSTM are used. The unwanted features are removed using different feature selection methods and these models are trained with the selected features to measure the performance.

2. Literature Review

In [1] proposed an ensemble model on a dataset consisting of 70000 records with 13 attributes. Random Forest is used for feature selection and Pearson's Coefficient to examine correlation between features. ML models such as KNN, Decision Tree and XGB are stacked. DL models such as DNN and KDNN are ensembled using Majority Voting. Finally, the proposed ensemble approach gave an accuracy of 88.7%. In [2] used SVM, Neural Network and Random Forest classifier models on a dataset with 303 records and concluded that SVM gave higher accuracy of 84.0% than other models. In [3] proposed a model to predict heart disease. They used SVM, Decision Tree, Linear Regression and K-nearest neighbour on a UCI repository dataset that has 303 records with 14 features and concluded that K-NN gave higher accuracy of 87%.

In [4] used a dataset with 70,000 records and 11 features for Heart Disease Detection. KNN, Random Forest Classifier, Decision Tree and SVM are used. Linear SVM Kernel and Gaussian SVM Kernel are used for SVM. In which Linear SVM Kernel gave 72.5% accuracy and Gaussian SVM Kernel gave 86.2% accuracy and concluded that Gaussian SVM kernel gave higher accuracy. In [5] used the deep Learning model for prediction of heart disease. Cleveland Dataset with 303 records and 14 features is used. Convolutional neural Network (CNN) is used which gave an accuracy of 75.2%. Mean Squared error is used for error calculations. In [6] proposed a model for Heart disease Prediction on UCI Dataset which has 303 records and 14 attributes. Naive Bayes, K-NN, Decision Tree and Random Forest models are used. K-NN with (k=7) gave higher accuracy of 90.789% than other models.

In [7] proposed an improved ensemble learning approach for the prediction of heart disease risk. The datasets used are Cleveland Dataset and Framingham dataset. The dataset is made to partitions by using a mean-based partitioning technique and then the Classification and Regression Tree (CART) is applied to model each partition. The Accuracy Bases average Aging Classifier Ensemble is used to compute an ensemble from various CART models. Cleveland and Framingham datasets gave accuracies of 93% and 91% respectively. In [8] proposed a Neural Network Model for Heart Disease Prediction. They used a Cleveland dataset with 303 records and 14 attributes. DNN with Talos Optimization is used. K-NN, SVM, Naïve Bayes, Logistic Regression and Random Forest models along with Talos are used and concluded that Talos gave better accuracy with 90.78% accuracy than other models.

In [9] proposed a model which is a combination of ML and DL. In this research they used three methods of evaluating. One is without any Feature Selection and Outlier Detection and second is with Feature Selection but no Outlier Detection and the third method with both. Logistic Regression, KNN, SVM, Random Forest, Decision Tree and ANN models are used. Deep Learning with ANN architecture gave an accuracy of 94.2 %. And also concluded that Machine Learning models performed better in the analysis. In [10] proposed a CNN model for prediction of heart disease on a dataset from UCI ML Repository. CNN model gave an accuracy of 94.78%. Proposed paper also compared the proposed model with LR, SVM, KNN, SVM with Linear and RBF kernel and Neural Network.

3. Problem Identification

These days heart disease has become one of the most common causes for death of people. Approximately around 17.9 million deaths are recorded annually due to heart problems. This is because of people following unhealthy diets and sedentary lifestyles. Heart is one of the important organs in the human body which does different functions like blood circulation, supplying oxygen, maintaining blood pressure and other conditions. Due to its various functionalities, the presence of heart disease cannot be ignored. Thus, it is very important to detect the presence of heart disease. When the patient is detected with heart disease, he could be monitored and treated to protect their lives. So, the main goal is to detect the presence of heart disease by using machine learning and deep learning techniques. The models will be trained after extracting the necessary features by giving the performance of the model built. [13-21]

4. Proposed Methodology

The proposed system is developed with three phases which includes Data Analysis, Feature Selection, Training the model shown in Figure 1.

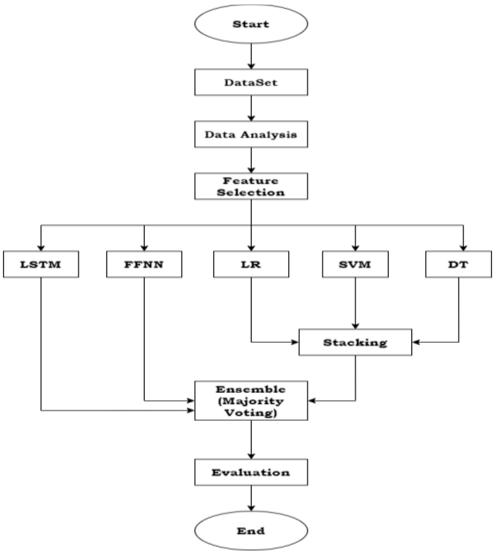

Figure 1: Working Model

4.1 Data Analysis

The data set considered has 1026 records of heart disease diagnosis. The data is collected from four locations like Cleveland, Switzerland, Long Beach, and Budapest. The dataset has 13 attributes with one binary target attribute and all the attributes are numeric values. The dataset considered has no null values. The dataset is analysed over different attributes present and the relation between the attribute with the target is visualised as graphs for better understanding.

4.2 Feature Selection

Feature selection method is used to remove the unimportant attributes from the dataset which reduces the complexity. Correlation matrix and fisher score methods are used.

Correlation Matrix:

The relationship between the attributes in the dataset is depicted by a correlation matrix, which is a matrix of correlation coefficients. The correlation coefficients range from -1 to 1. The calculated coefficients are visualised as a heat map. The attributes that are weakly correlated are not considered for further model training.

Fisher Score:

One of the most popular supervised methods for selecting features is the Fisher score. The Fisher score contrasts the between-class variance with the within-class variance of the attributes. It gives the ranks of the features and the attributes with low fisher score that is the least ranked variables are removed.

By using the both methods, the features that are eliminated are age, fbs, trestbps, chol, restecg.

4.3 Training the models

After the elimination of the features the models are trained.

K-Fold Cross Validation:

In this study, K-Fold cross-validation method is used to evaluate each model. The method first shuffles the data set randomly before splitting it into (k=5) groups, and for each unique group, that group was used as the test data set and the remaining groups were used as the training data. The model was then fit and evaluated accordingly.

The models such as Logistic regression, SVM, Decision Tree, FFNN and LSTM are trained individually. The ML models SVM, logistic regression, Decision Tree are stacked with a meta model and an ensemble model with all models is built. An ensemble of ML stacking and DL models is built by using majority voting.

5. Implementation

The machine learning models used in implementing the system are Logistic regression (LR), Support Vector Machine (SVM) and Decision Tree (DT), the deep learning models used are Feed Forward Neural Network (FFNN) and Long Short-Term Memory (LSTM).

5.1 Machine Learning Models

5.1.1 Logistic Regression

Logistic Regression is a supervising machine learning classification model used for binary classification problems, where the goal is to predict whether a certain event will occur or not. It is a type of generalised linear model that gives the result as a probability that ranges from 0 to 1 by using a logistic function. It is implemented using lbfgs solver in this project. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs) is a memory-saving optimization method.

5.1.2 SVM

A supervised machine learning classification technique called SVM can be applied to classification or regression tasks. In classification, SVM tries to separate data points into different classes using a hyperplane. It is a binary classifier by default, which can only separate data into two classes. SVM looks for the hyperplane that maximises the separation between the both classes.

5.1.3 Decision Tree

Decision Tree is a supervised machine learning algorithm used for classification tasks. It operates by repeatedly dividing the dataset into subsets according to the values of the input features, resulting in a decision-tree-like structure. Decision Tree works by splitting the data at best features until the leaf nodes have the samples from a single class. The best feature in the obtained subsets is the one which maximises the difference between the classes.

5.1.4 Stacking

Stacking is one of the ensemble learning methods that gives better performance by combining several models. The basic working of stacking is that initially the base models are trained and predictions are generated on specific input values. The base models could be any algorithm. The predictions of the base models are combined by a meta model that is trained to produce better results. Meta model used is Logistic regression.

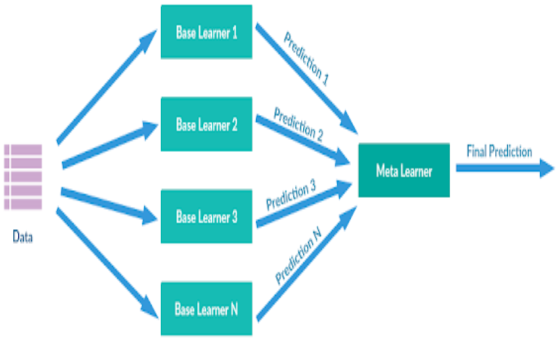

Figure 2: Stacking

5.2 Deep Learning Models

5.2.1 FFNN

A Feed Forward Neural Network (FFNN) is a simple artificial neural network. It contains an input layer, hidden layers, and an output layer. It can be assumed as a single layer perceptron. In FFNN information is processed in only one direction. Sequences of inputs entering the input layer are multiplied with weights and weighted inputs are summed to get a total. FFNN is implemented with sequence of dense layers along with ReLu and sigmoid activation functions and is compiled with adam optimizer to evaluate the performances.

5.2.2 LSTM

A sequential neural network called Long Short-Term Memory (LSTM) moves data based on timestamp from previous layers to later layers. LSTM relies on three gates namely Input gate, Forget gate, Output gate. Forget gate is useful to know previous timestamp information is taken or forgotten. Input gate is useful to add new information to neurons, Output gate provides updated information from previous timestamp to new timestamp. LSTM provides a solution to the Vanishing Gradient Problem by using timestamps.

5.3 Ensemble

Ensemble learning improves the performance by combining various models. An estimator called voting classifier is used to train the models and make predictions based on aggregating the results of each model. Majority voting is used to ensemble the models. The output is the class that has got the greatest number of votes or the class that each classifier believes has the best chance of being correctly predicted. Majority voting is usually used to improve the accuracy. Majority voting is used to ensemble FFNN, LSTM and the stacking model in this study.

6. Results

In this study, stacking of machine learning models and Ensemble of ML stacking and deep learning models are done.

Models	LR	SVM	DT
Accuracy	84.00	84.29	82.92
Precision	81.40	81.24	78.51
Recall	89.13	90.30	91.34
F1-Score	85.03	85.50	84.44
ROC-AUC	83.86	84.13	82.40

Table 1: Performance metrics of ML models (%).

Models	FFNN	LSTM
Accuracy	85.85	81.95

Precision	82.60	79.64
Recall	91.34	86.53
F1-Score	86.75	82.94
ROC-AUC	85.77	92.23

Table 2: Performance metrics of DL models (%).

Models	ML Stacking	Ensemble
Accuracy	86.14	87.21
Precision	83.40	84.31
Recall	91.04	92.20
F1-Score	87.01	88.07
ROC-AUC	86.01	86.03

Table 3: Performance metrics of ML stacking and Ensemble models (%).

The comparison metrics of the proposed approach are shown in Table 1, 2, and 3.

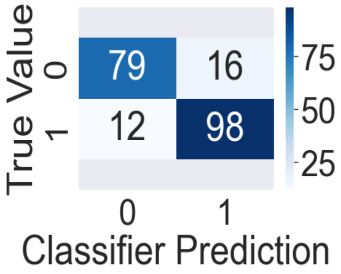

Figure 3: Confusion Matrix of Stacking classifier.

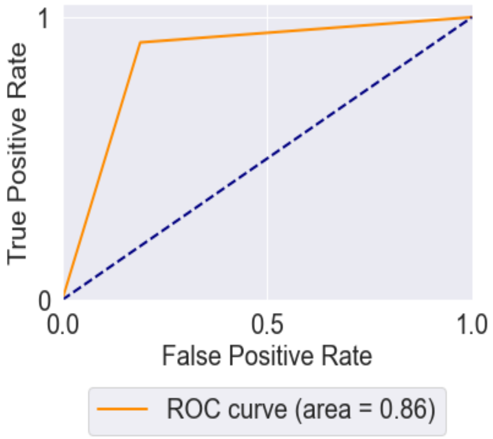

Figure 4: ROC-AUC curve of Stacking Classifier.

Figure 3 and 4 show the Confusion matrix and ROC-AUC curve of stacking classifier.

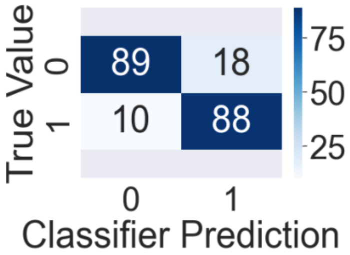

Figure 5: Confusion Matrix of Ensemble classifier.

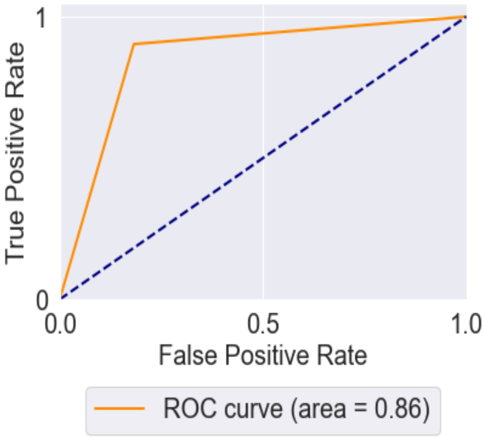

Figure 6: ROC-AUC curve of Ensemble classifier.

Figure 5 and 6 show the Average Confusion matrix and ROC-AUC curve of Ensemble classifier. In this study, an Ensemble model of ML stacking and DL models gave higher accuracy of 87.2%.

7. Conclusion

The Machine Learning and Deep Learning models in the study are designed to detect the presence of heart disease. The Ensemble model built gave an accuracy of 87.21%. The scope of this study can be extended by using new datasets that may detect specific heart disease like Myocardial infarction etc. And may use alternate models to detect the heart disease by various pre-processing techniques, resampling procedures and also by applying various ensemble techniques.

References

[1] Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular Disease Detection using Ensemble Learning. Computational Intelligence and Neuroscience, 2022.

[2] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart Disease Prediction Using Machine Learning. Heart Disease, 5(1).

[3] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

[4] Chithambaram, T., & Gowsalya, M. (2020). Heart disease detection using machine learning.

[5] Poonam Vengurlekar1 Swati Nadkarni, Dr. Bhavesh Patel "Heart Disease Prediction using CNN, Deep Learning Model", IJSRD, October 2020, ISSN (online): 2321-0613..

[6] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1, 1-6.

[7] Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20, 100402.

[8] Sharma, S., & Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(3), 2244-2248.

[9] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021.

[10] Sajja, T. K., & Kalluri, H. K. (2020). A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. Rev. d'Intelligence Artif., 34(5), 601-606.

[11] W. H. Organization, Cardiovascular Diseases. 2020.

[12] Alalawi, H. H., & Manal, S. A. (2021). Detection of cardiovascular disease using machine learning classification models. International Journal of Engineering Research & Technology (IJERT) ISSN, 2278-0181.

[13] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[14] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[15] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[16] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[17] Sri Hari Nallamala, et.al, “Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment”, International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[18] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., “Accurate Hand Gesture Recognition using CNN and RNN Approaches” International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[19] Sri Hari Nallamala, et al., “A Review on ‘Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management’”, Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[20] Nallamala, S.H., et al., “A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems”, IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[21] Nallamala, S.H., Mishra, P., Koneru, S.V., “Breast cancer detection using machine learning approaches”, International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.