Cardiovascular Diseases Prediction using Deep Neural Networks

Pruthivi Raj Behera[a] (MT20037), Shreya Goel[b] (MT20054) and Richa Dwivedi[c] (MT20104)

[a]*pruthivi20037@iiitd.ac.in*
[b]*shreya20054@iiitd.ac.in*
[c]*richa20104@iiitd.ac.in*

ARTICLE INFO

Keywords:
Cardiovascular Disease
Data Cleaning
Data Transformation
Data Analysis
Machine Learning
Neural Network
CNN
LSTM
Classification Score

ABSTRACT

Cardiovascular is a broad hypernym that describes heart infection, coronary artery disease, congenital heart defects, arrhythmia, and much more. It is estimated that around 90% of cardiovascular diseases are may be amenable to cure or preventable. Its prevention is based on 2 factors, i.e., modifiable and non-modifiable risk factors. The modifiable (poor diet, excessive alcohol, cholesterol level) can be controlled, while non-modifiable (e.x. age, genetics, etc.) are those that cannot be controlled. By considering the risk factors and severity of this disease, healthcare systems need an efficient method that can pre-impose a person regarding cardiovascular disease risk. Consequently, a person may take timely proper medication and may prevent it.

This paper focuses on Data Mining techniques for the clinical data collected to predict whether an individual has CVD. After this, various traditional classifiers such as Logistic Regression, Support Vector Machine, Decision Tree etc. and boosting algorithms such as Gradient Boost, Optimized Decision Tree, Random Forest, Catboost and neural network models are applied to build the prediction system for the risk of developing cardiovascular disease and analyzed based on diffent metrics like Accuracy, Precision, Specificity, Sensitivity, F1 Score, AUC and MCC. In this investigation of foreseeing Cardiovascular Disease, the results obtained were promising with boosting classifiers and neural network achieving the highest values for all the evaluation metrics. In this work, the proposed system is based on deep learning algorithms, i.e., Neural network, CNN, LSTM, to construct an efficient prediction system.

1. Introduction

Healthcare is emerging rapidly nowadays, and it contains a considerable amount of patient data, symptoms of the disease, treatment provided, clinical history, etc. Day-by-day advancements in technology and computer science tools can improve patient care by analyzing available data and predicting the possible disease more accurately. Disease prediction uses machine learning algorithms to save user's time and money and provide a correct diagnosis in time.

Cardiovascular diseases (CVD's) is a set of high-risk diseases, which affects the heart and blood vessels of the body. CVD is caused due to high blood pressure, high cholesterol, alcohol, smoking, etc. It can moreover affect other organs in the body such as the mind, kidneys and eyes. Sometimes when it is not cured on time, it becomes the leading cause of death. A study by Bodkhe et al. [4] shows that, in 2015, 56.4 million people, i.e., 70% of the total deaths occurred due to non-communicable diseases, in which 81% are because of cardiovascular disease, and globally, 15.5% of the people die daily. According to the world health organization [2], in India, 63% of people died due to Non-communicable diseases, of which 27% died due to cardiovascular disease. Considering the risk, an effective method must be developed to sort through heterogeneous populations and recognize individuals at risk of developing CVD.

The paper is broadly organized as follows: Section "Data Understanding" presents the preprocessing and analysis of the data, Section "Methodologies" presents the various methods used along with evaluation and results and Section "Conclusion" concludes the study and outlines the future work.

2. Problem Statement

Presently, the Medical Diagnosis System is playing an essential role in diagnosis and treatment. The paper focuses on predicting the risk of Cardiovascular disease among patients using Machine Learning and Deep Learning techniques. The information is being extracted from the clinical records of the patient collected using Data Mining. This system's automation will help the physicians for a fast and better diagnosis in the future with a reduction in costs. It also stresses Data Analysis, which plays a vital role in the medical field's progress by suggesting the proper treatment according to the patient's characteristics.

Various lifestyle factors are considered, such as age, blood pressure, weight, height, cholesterol, alcohol, activity, and many more, which play a significant role in the high risk of Cardiovascular Disease. Hence, a sound prediction system is required for cardiovascular disease so that people can stay more healthy and safe. The main aim is to determine the individuals at higher risk of developing Cardio Vascular disease at an earlier stage to avoid premature deaths.

3. Literature Review

Some of the earliest works in prediction of cardiovascular diseases was done by Amma [3], where the author presented a medical diagnosis system for predicting cardiovascular disease through a combination of Genetic algorithm and neural network. The Cleveland heart disease dataset was used, which included various ECG, sugar, and cholesterol attributes. The model implemented was a multilayered feed forward neural network with optimisation of the weights using the Genetic algorithm. The model then used these weights for the prediction of the severity of the cardiovascular disease. Although the author attained a respectable accuracy, the work done was on a very small dataset which could potentially show bias.

Since then a lot of research was done in this field. In one such research, Dinesh et al. [6] implemented various machine learning techniques, including support vector machine, gradient boosting, random forest, naive bayes classifier and logistic regression for prediction of cardiovascular diseases. Data was collected from Cleveland, Hungarian, Switzerland, Long Beach VA heart disease database (obtained from UCI Machine Learning Repository) consisting of various attributes ranging from basic features such as age, gender etc., to advanced features such as rest ECG, blood pressure etc. Since the dataset size was too small, the authors were able to achieve a decent score.

In the paper, Indrakumari et al. [7] used the same dataset of 303 patients with 76 features to analyze the possibility of Heart Disease. In the paper, they used various visualization tools and clustering algorithms to analyze the dataset. Then, preprocessing was done over the dataset where outliers were removed and other features were fine tuned. Later, they perform prediction on 209 patients with seven features like age, type of chest pain, resting blood pressure, fasting blood pressure etc, by using K-means clustering algorithms, where data is clustered and based on clusters, Chest pain is predicted.

In the paper, Kumar et al. [9] has used various classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN) for the classification of the patient suffering from cardiovascular disease or not. The dataset was obtained from the UCI repository, which specified attributes such as age, gender, ECG, and blood pressure. The data was integrated, transformed, reduced and cleaned using the pandas tool. In this paper, various metrics such as ROC AUC score and precision were also employed for each classifier, with the Random Forest classifier accomplishing the highest score.

Meshref [11] worked on the same Cleveland heart data set, where they applied different machine learning models, such as Naive bayes, support vector machines, and Random Forest. Multilayer perceptron (MLP) was also implemented. In the paper, several feature selection techniques were applied and by using different ML models, they achieved a decent score. For future work, Cleveland Hungarian data sets can be combined to perform the required analysis and improve the system's efficiency.

Also, Krittanawong et al. [8] has summarized all the existing papers and performed a meta-analysis of all the empirical results obtained from different experiments. Although ML algorithms showed promising results, the results obtained were far from optimal. The authors observed that SVM and boosting algorithms gave decent results. Besides that, the sample size for each type of experiments were small, and the pooled results were potentially biased. The authors also observed few other limitations in some papers, including different kind of feature selection methodologies and different evaluation metrics used. In some papers, F-score was not reported, whereas, in others, positive or negative cases and sensitivity/specificity were not reported.

We plan to encounter the above problems and limitations in this paper along with performing experiments on a much larger dataset.

4. Data Understanding

The Cardiovascular Disease dataset is obtained from Kaggle [1], in which the data was collected at the moment of medical examination. The data does not contain any directly identifiable information. The dataset consists of 70,000 records of patient data, 11 features, and one target variable, which gives the presence or absence of cardiovascular disease in the form of 1 or 0, respectively. All 12 features are broadly categorized into three categories:

- Objective: factual information;
- Examination: results of medical analysis;
- Subjective: the patient gave information.

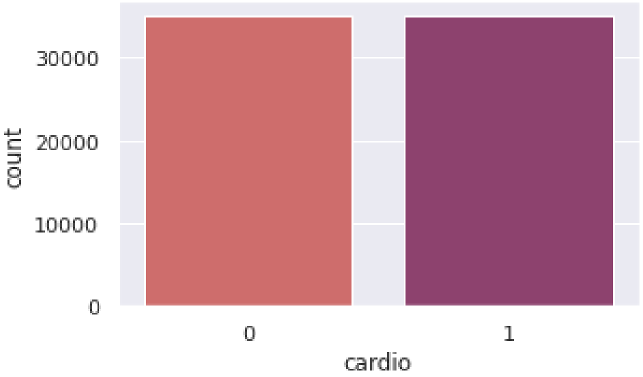

Figure 1: Class Distribution of Target Variable Cardio.

4.1. Data Preprocessing

The first step is to preprocess the data, which is performed before training and testing the model. The different preprocessing approaches have been applied based on the input data as follows:

4.1.1. Data Cleaning

The irrelevant and missing parts from the data was cleaned. Data must be cleaned by filling any missing values, identifying the noisy data or outliers 3, and resolving inconsistencies.

1. **Removal of Duplicate Data:** The duplicate rows were dropped from the data. There were 48 duplicate rows present in the data.

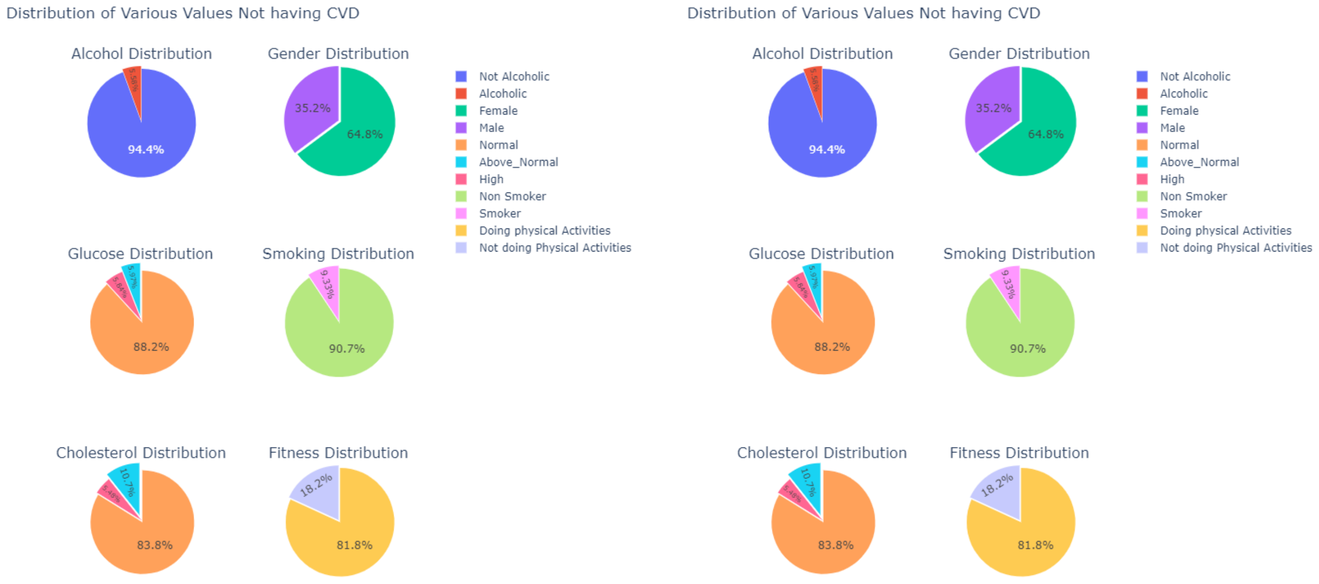

Figure 2: Distribution of Variables for Cardiovascular Disease.

2. **Missing Values or Inconsistencies:** There was no missing value found in the data and any discrepancies in the values.
3. **Outliers Removal:** As the figure 3 shows, the extreme values deviating from other observations present in the data. Also, The outliers values are not possible in the real world. The Interquartile Range method was used to remove outliers from columns - ap_lo, ap_hi, weight and height. The extreme values were detected and removed based on the interquartile ranges.

Table 1
Feature Description.

Attribute	Category	Description
id	Objective	Identifier, int(unique)
age	Objective	Age , int(days)
height	Objective	Height, int(cm)
weight	Objective	Weight, int(kg)
gender	Objective	Gender, Categorical code[1]
ap_hi	Examination	Systolic Blood Pressure ,int
ap_lo	Examination	Diastolic Blood Pressure ,int
cholesterol	Examination	Cholesterol[2]
gluc	Examination	Glucose[2]
smoke	Subjective	Binary[3]
alco	Subjective	Binary[3]
active	Subjective	Binary[3]
cardio	Target	Binary[3]

[1] Value: 1-Female, 2-Male
[2] Value: 1-Normal, 2-Above Normal,3-Well Above Normal
[3] Value: 1-True, 0-False

4.1.2. Data Transformation

The data needs to be consolidated in the form which is relevant for modelling.

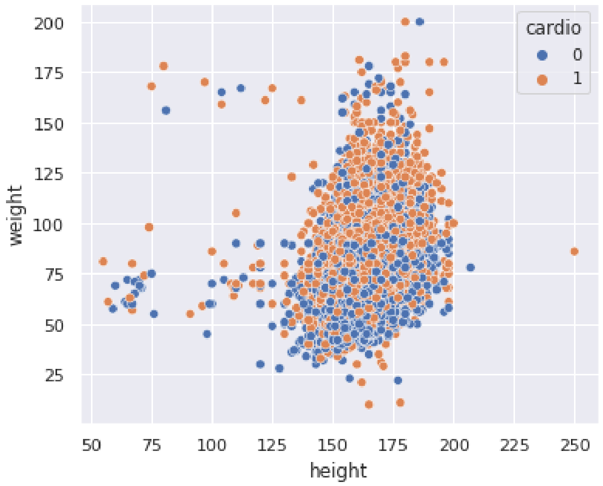

Figure 3: Distribution of weight and Height

1. The age column was converted into years by dividing the values with 365.24.
2. The weight and height columns were integrated into BMI Index as weight and height are highly correlated. A new column **bmi** was added to the data.

After preprocessing, there were 66775 patients whose records were kept for consideration.

4.2. Data Analysis

For the analysis, data visualization is done to get a better understanding of data. The class distribution of target variable cardio is visualized by a bar plot in Fig 1, where around 50% (35004) users belong to the positive class, i.e., "1." In comparison, 49.9% (34972) users belong to the negative class, i.e., "0," which implies that there is no class imbalance in cardio distribution. Age is converted into years, and cardiovascular disease risk is analyzed among various age groups, in the Figure 4. It shows that people having age more than 55 are vulnerable to cardiovascular disease.

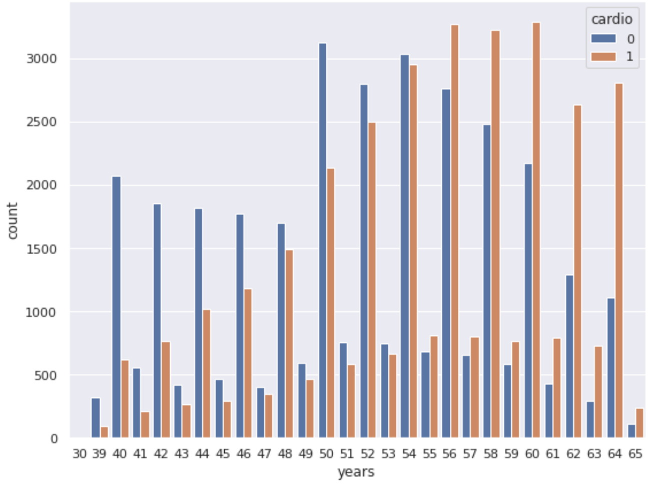

Figure 4: Risk of Cardiovascular Disease Based on Age

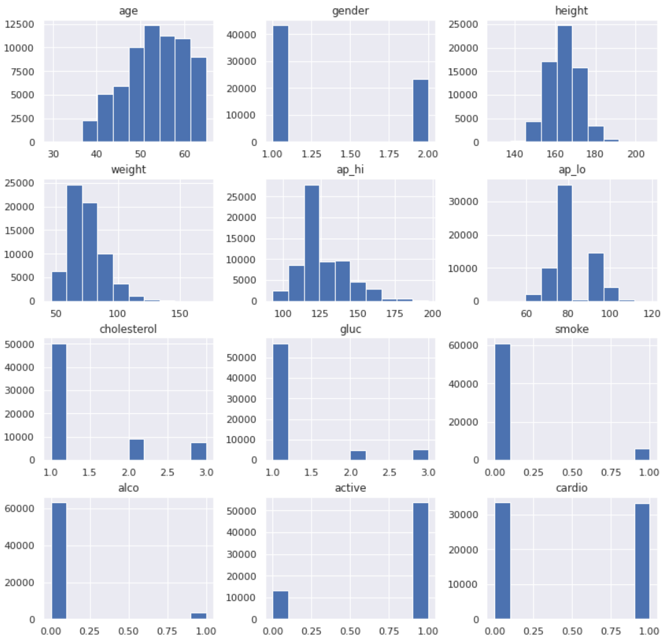

Figure 5: Histogram plot of all the Variables.

The histograms analyze Data distribution for all available attributes in Figure 5 where the bar represents the range of count of patients for every feature obtained after preprocessing.

The Box plot with the features gender, alcohol, and BMI is shown in the figure 6 in which it is observed that Alcoholic women have a higher Risk of CVD than Alcoholic men. The rate of Having Cardiovascular disease based on Alcohol, Gender, Glucose, Cholesterol Level, Smoking, and Physical Activity, are shown by a pie chart in the figure 2. The rate of not having a cardiovascular disease based on variables mentioned above is also shown in figure 2. On Comparing the plots, It is obtained that Cardiovascular diseases are not much affected by gender, while it majorly depends on body Glucose level and Cholesterol level.

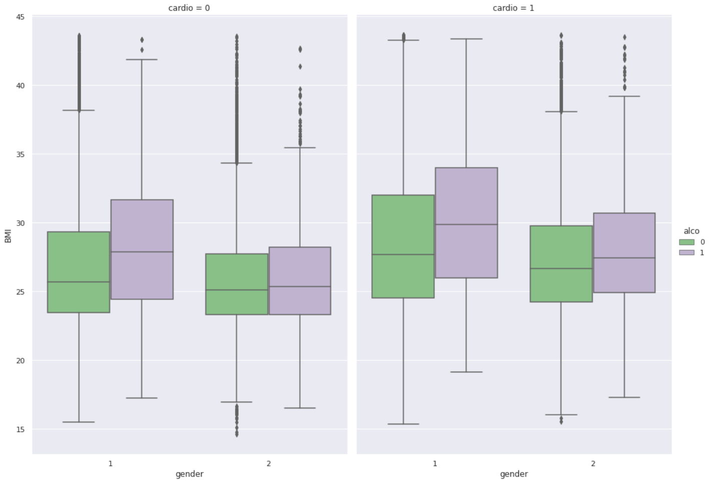

Figure 6: Distribution of Gender, Alcohol and BMI for cardiovascular disease

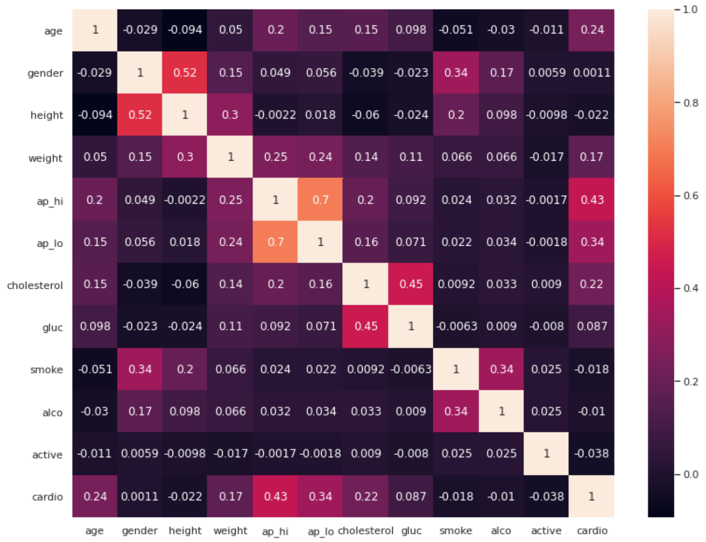

Figure 7: Pearson Correlation of the Features

A Heatmap in figure 7 displays the correlation between features and also with target variable. The feature with higher correlation with target variable are given more importance than features correlated with independent variables. The figure 7 also shows that ap_hi and ap_lo are highly correlated with the target variable, and gender is the least correlated variable with target. Also, height, glucose, smoke have less amount of a correlation with the target variable cardio.

Table 2
Experiment 1: Best results obtained from different Machine Learning and Deep Learning Models using split validation.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC	MCC
Logistic Regression (LR)	72.799	69.846	66.847	78.917	71.355	79.382	46.062
Support Vector Machine (SVM)	72.220	68.444	63.646	81.032	69.901	79.415	45.316
k-Nearest Neighbors (KNN)	72.854	70.213	67.763	78.087	71.674	79.217	46.063
Naive Bayes (NB)	71.207	67.121	61.114	81.578	68.269	78.187	43.559
Decision Tree (DT)	72.924	69.762	66.423	79.605	71.319	79.593	46.391
Decision Tree - Optimised (DT-O)	72.779	69.027	64.513	81.275	70.608	79.383	46.392
Random Forest (RF)	73.354	70.686	68.304	78.542	72.209	80.256	47.060
Gradient Boosting (GB)	73.014	70.973	69.506	76.619	72.305	79.768	46.219
Light Gradient Boosting (LGBM)	73.458	70.979	68.915	78.127	72.466	80.307	47.212
XGBoosting (XGB)	73.543	70.867	68.511	78.714	72.413	80.360	47.439
CatBoost (CB)	**73.558**	**71.192**	**69.319**	**77.915**	**72.657**	**80.393**	**47.380**
Neural Network (NN)	**72.784**	**70.315**	**68.137**	**77.560**	**71.733**	**72.849**	**45.872**

5. Methodologies

Inspired from Martins et al. [10] various models have been implemented including Logistic Regression (LR), Support Vector Machine (SVM), k Nearest Neighbour (k-NN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Gradient Boosted Tree (GBT), Light Gradient Boosted Tree (LGBM), XGBoosting (XGB) and lastly CatBoost (CB) for the baseline implementation. All experiments were performed after a 70:30 training and testing split as it yielded a decent amount of testing data to capture most of the variance in the data. For the experiment 1, all attributes were considered. For the experiment 2, attributes with less feature importance which includes gender, gluc, alco, smoke and active were removed as shown in Figure 8.

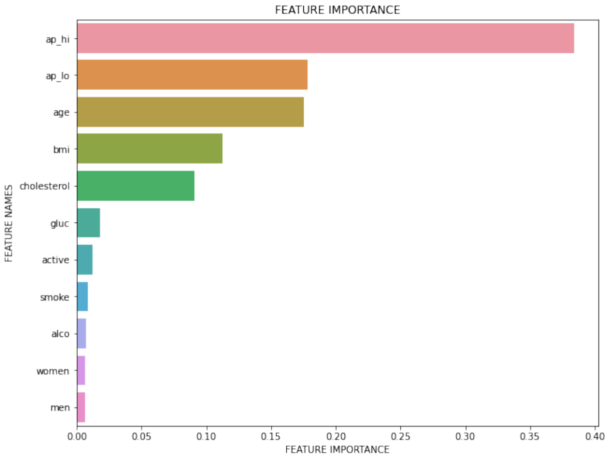

Figure 8: Distribution of Gender, Alcohol and BMI for cardiovascular disease

6. Evaluations

To compute all metrics, True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) [12] were calculated as they will be used for creation of confusion matrix M as shown below:

$$M = \begin{bmatrix} TP & FP \\ TN & FN \end{bmatrix}$$

Then for the evaluation purposes, various metrics were calculated including accuracy, precision, sensitivity, specificity, AUC metric along with F-Score. Metrics were computed using the below formulas:

$$\text{Accuracy} = \frac{(TN + TP)}{(TN + TP + FN + FP)}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)}$$

$$\text{F1 Score} = \frac{TP}{\left(TP + \frac{1}{2}(FP + FN)\right)}$$

Matthews Correlation Coefficient (MCC) was also used for evaluation as it is regarded as more reliable statistical measure [5] for binary classification.

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

For evaluation, receiver operating characteristic curve (ROC curve) was also made as shown in the figure 9.

7. Present Work Status

For present work, we are currently working on a neural network implementation. Using only 2 fully connected layers with RMSProp Optimiser and early stopping, we were able to achieve a accuracy of around 72-73% comparable to baseline implementations. Hence, we were planning to implement our proposed model. i.e. CNN and LSTM based

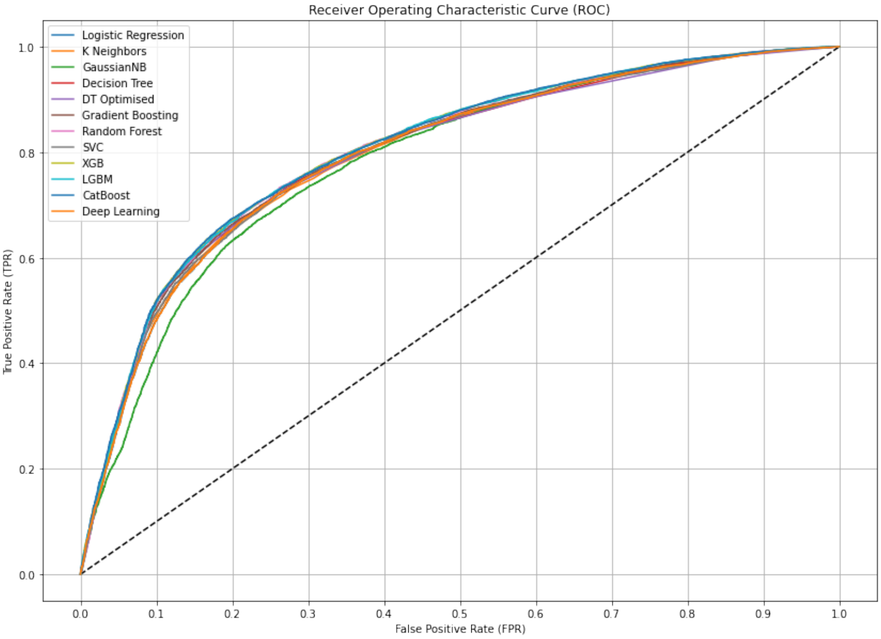

Figure 9: Experiment 1: ROC Curve

Table 3
Experiment 2: Best results obtained from different Machine Learning and Deep Learning Models using split validation.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC	MCC
Naive Bayes (NB)	71.542	67.419	61.508	81.852	68.660	78.630	44.228
Decision Tree - Optimised (DT-O)	72.695	69.247	65.300	80.293	70.795	79.255	46.068
Random Forest (RF)	73.194	70.529	68.117	78.410	72.034	79.905	46.743
XGBoosting (XGB)	73.269	70.641	68.304	78.370	72.145	80.126	46.879
CatBoost (CB)	**73.274**	**70.884**	**68.925**	**77.742**	**72.330**	**80.073**	**46.821**
Neural Network (NN)	**72.755**	**69.926**	**67.132**	**78.532**	**71.409**	**72.832**	**45.928**

model for improved results. Also, we were planning on trying with different activation functions, hyper-parameter tuning, experimenting with batch normalization layers and different regularization and Dropout techniques.

8. Results and Observations

We can observe that the in traditional ML techniques, we got accuracy of around 72% whereas in boosting algorithms and neural network, we got 73% accuracy (a minor improvement). However, accuracy is not a reliable measure in healthcare domain due to a large difference in true negatives and false positives. The number of false positives must be minimized as much as possible as the impact due to false positives is much more than impact due to true negatives. Hence, sensitivity and precision are much better measure than accuracy and specificity. Boosting algorithms tend to work better than traditional machine learning algorithms as noted in Table 2.

In experiment 2, as observed in Table 3 some algorithms like Naive Bayes and Neural Network performed better contrary to boosting algorithms where results were dropped across some metrics. A partial reason that could be attributed to the fact that neural network worked better is due to less features hence less complex network. Also, we note that the MCC score, along with the F1 score, gives way better representation of which classifier works better matching with the study done by Chicco and Jurman [5].

9. Conclusion

In conclusion, it was observed that machine learning models were able to achieve around 70 percent score in balanced dataset. Overall, it can be noted that boosting algorithms and neural network will perform better. Some achieved decent score also hence, viable to use in prediction of CVD's.

Compliance with Ethical Standards

Conflict of interests Pruthivi Raj Behera, Shreya Goel, Richa Dwivedi declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] , a. Cardiovascular disease dataset. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

[2] , b. Cardiovascular diseases in india. https://www.who.int/india/health-topics/cardiovascular-diseases.

[3] Amma, N.G.B., 2012. Cardiovascular disease prediction system using genetic algorithm and neural network, in: 2012 International Conference on Computing, Communication and Applications, pp. 1–5. doi:10.1109/ICCCA.2012.6179185.

[4] Bodkhe, S., Jajoo, S.U., Jajoo, U.N., Ingle, S., Gupta, S.S., Taksande, B.A., 2019. Epidemiology of confirmed coronary heart disease among population older than 60 years in rural central india—a community-based cross-sectional study. Indian heart journal 71, 39–44.

[5] Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics 21, 1–13.

[6] Dinesh, K.G., Arumugaraj, K., Santhosh, K.D., Mareeswari, V., 2018. Prediction of cardiovascular disease using machine learning algorithms, in: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1–7. doi:10.1109/ICCTCT.2018.8550857.

[7] Indrakumari, R., Poongodi, T., Jena, S.R., 2020. Heart disease prediction using exploratory data analysis. Procedia Computer Science 173, 130–139.

[8] Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., et al., 2020. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific reports 10, 1–11.

[9] Kumar, N.K., Sindhu, G.S., Prashanthi, D.K., Sulthana, A.S., 2020. Analysis and prediction of cardio vascular disease using machine learning classifiers, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 15–21. doi:10.1109/ICACCS48705.2020.9074183.

[10] Martins, B., Ferreira, D., Neto, C., Abelha, A., Machado, J., 2021. Data mining for cardiovascular disease prediction. Journal of Medical Systems 45, 1–8.

[11] Meshref, H., 2019. Cardiovascular disease diagnosis: A machine learning interpretation approach. International Journal of Advanced Computer Science and Applications 10. URL: http://dx.doi.org/10.14569/IJACSA.2019.0101236, doi:10.14569/IJACSA.2019.0101236.

[12] Zhu, W., Zeng, N., Wang, N., et al., 2010. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland 19, 67.