CARDIAC DISORDERS PREDICTION COMPARATIVE STUDY USING MACHINE LEARNING

Arpit Mohan
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
vnsmohanarpit@gmail.com

Kartik Kumar
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
kartikrajput2656@gmail.com

Abstract— **In recent times, accurate prediction of heart disease is one of the most complicated tasks in medical field. Approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. The proposed work will predict the chances of heart disease and will classify patient's risk level by implementing different machine learning techniques such as Naive Bayes, KNN, Logistic Regression and Random Forest. In addition to above mentioned algorithms we will using Convolutional Neural Network algorithm also for prediction of heart disease. Thus, the research paper will present a comparative study by analyzing the performance of different machine learning algorithms. The trial results will verify which algorithm will achieve the highest accuracy compared to other ML algorithms implemented. For performing these algorithm we will be making use of UCI heart disease repository dataset which consists of approx. 300 records based on 13 factors like age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, etc. And hence the paper will present the comparative study of all the above mentioned algorithms and find which one of them has the highest accuracy.**

Keywords— **Heart disease, medical sector, Naïve Bayes, Random Forest, KNN, Logistic Regression, CNN.**

I. Introduction .

Heart disease refers to what is really a group of conditions that affect the structure and functions of the heart and has many root causes. Heart is a muscle and its job is to pump blood around the body. Heart pumps blood through a network of arteries and veins. For most types of heart disease, maintaining a healthy lifestyle (healthy eating, physical activity, avoiding tobacco misuse) is a key part of preventing these conditions. Women's experience with heart disease is different from men's in several important ways. According to the World Health Organization more than 10 million die due to heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease.

The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences. Nevertheless, the disease, when recognized earlier, makes the treatment unconstrained for as much as identification, a challenging task depends solely on the medical community. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial. To make the task of prediction of heart disease easier the paper proposes the use of machine learning so that higher accuracy and correct results be obtained. Machine learning is a subset of artificial intelligence (AI). It is focused on teaching computers to learn from data and to improve with experience – instead of being explicitly programmed to do so. In machine learning, algorithms are trained to find patterns and correlations in large data sets and to make the best decisions and predictions based on that analysis. Machine learning applications improve with use and become more accurate the more data they have access to. Applications of machine learning are all around us –in our homes, our shopping carts, our entertainment media, and our healthcare. Machine Learning (ML) handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various Machine Learning techniques such as Naive Bayes, KNN, Logistic Regression and Random Forest for predicting heart disease at an early stage.

In addition to these four machine learning algorithms the paper will present an architecture using a convolutional

neural network for classification between healthy and non-healthy persons to overcome the limitations of classical approaches. Various clinical parameters are used for assessing the risk profile in the patients which helps in early diagnosis. Various techniques are used to avoid overfitting in the proposed network. A Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics. The input to the architecture will be the 13 features that are important in the classification of heart disease. These features are converted to a new representation called word embedding by the layer called as Embedding Layer. The input is given as in CSV file or manual entry to the system. After taking input all the algorithms apply on that input that is Naïve Bayes, Logistic Regression, KNN and Random Forest. After accessing data set the operation is performed and effective heart attack result is produced.

1.2 Formulation of Problem

Tools and Technology used

i. Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

ii. Jupyter Notebook

Jupyter Notebook is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

iii. Kaggle

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

II. LITERATURE SURVEY

Over the years, a range of works have been done related to heart disease prediction system using different machine learning algorithms by different authors. They tried to attain efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. In an efficient way different models are developed to diagnose heart disease more accurately. In this work various machine learning techniques and tools are implemented over diverse attributes to predict the presence of heart diseases. Machine Learning classification techniques for good decision making in the field of health care addressed are namely Decision trees, Naive Bayes, Logistic Regression and Random Forest algorithms. Each technique proves efficient in diagnosis of heart diseases and has its own efficiency and consistency. In this work a model is developed by comparing the efficient output features of Decision trees, Naive Bayes, Logistic Regression and Random Forest algorithms and shows promising efficiency and accuracy in the diagnosis of heart diseases.

Verma et al. [1] did the Prediction of heart disease using Logistic Regression Algorithm using a dataset where initially Linear Regression was used but the results obtained were not exact hence Logistic Regression was utilized for gaining the result. The machine learning model used body dimensions, physiological factors and medical problems. Then Zhang et al. [2] performed Logistic Regression Models in Predicting Heart Disease for predicting heart disease among elderly people using a dataset dependent on age, sex, cp, chol, restecg, oldpeak, slope, ca, thal. from UCI machine learning repository and achieved higher accuracy in prediction. On comparison with other machine learning algorithms like KNN and Neural Network, Logistic Regression gave more accurate prediction. Another paper with another algorithms by Singh et al. [3] Heart Disease Prediction System using Random Forest explained the prediction of heart disease using Random Forest algorithm. The non-linear behavior of Cleveland heart disease dataset was studied and obtained accurate results. Yekkala et al. [4] did prediction of heart disease using random forest and rough set based feature selection comparing Feature Selection and random forest algorithm. The dataset used is from UCI repository. Later Madhumita Pal and Smita Parija [5] performed prediction of heart disease using Random Forest on a dataset that had 303 samples based on 14 attributes on Jupyter Notebook. The work obtained the classification accuracy for prediction of heart disease and rate of diagnosis using random forest algorithm. Taqdees et al. [6] did the comparative study of heart disease prediction using Naïve bayes, KNN, Decision Tree, Random Forest on a dataset from UCI Machine Learning repository on Jupyter Notebook. The dataset was split into ratio of 75% training data and 25% testing data. On comparison of the results accuracy obtained by Naïve Bayes was higher than that of Decision Tree. Wankhede et al. [7] also did heart disease prediction using hybrid Random Forest model integrated with linear model. They used the Cleveland dataset from UCI repository. They found that Integration of linear model with Random Forest Model makes the simple estimation procedure with overall accuracy than the respective models. Accuracy obtained in Linear Model was higher than, Random Forest and accuracy in Hybrid Random Forest Model Integrated with Linear Model was more than both of them. Considering the factors of fear Karthiga et al. [8]

discussed the factors that increase the risk of Heart Disease like family history, smoking, poor diet, physical inactivity etc. and discussed classification using Decision Tree algorithm and classification using Naïve Bayes. As a result Decision Tree had more accuracy than using Naïve Bayes. Hence concluding that Decision Tree gives more accurate result. For making the problem's solution more modern Kennedy Ngure Ngare [9] built GUI based user friendly system that predicts the heart attack level using Naïve Bayes algorithm. The dataset UCI was inputted as in CSV file or manual entry to the system. The system aims at reducing treatment costs by initial diagnostics in time and can be used as training tool for medical students and serves as soft diagnostic tool for physicians and cardiologists. Polaraju et al. [10] also built a system that was implemented using C# on .NET framework, Visual Studio and winForms. Diagnostics through the system shows whether the patient has heart disease with the help of attributes of dataset like sex, chest pain type, fasting blood sugar etc. With experimental results the accuracy of Regression was better.

Dr. Poonam Ghuli [11] and students performed the comparison of Logistic Regression, Naïve Bayes, Random Forest and Decision Tree with the help of UCI HD dataset based on the factors like age, sex, cp, etc. They divided data into 80:20 training and testing data. As a result random Forest is the most efficient algorithm with higher accuracy. Ponnusamy and Thenmozhi [12] performed various decision tree algorithms such as ID3, C4.5, C5.0, J48 in classification and prediction of disease. After the study of works of different people they concluded Decision Tree classifier for disease prediction for its simplicity and accuracy. V. Sabarinathan and V. Sugumaran [13] performed the paper that discusses about heart disease prediction system using Decision Tree technique for classifying the features and for feature selection. Attributes such as age, gender, etc were used in classification. Hence accurately achieving result that predicts the cause of Heart Disease.

Viney and Bindu [14] used Naïve Bayes and Laplace Smoothing for classification. They also compared the accuracy of prediction when the number of medical attributes used for prediction is decreased. Data used was from Cleveland Heart Disease Database. Accuracy of higher values was gained hence Laplace Smoothing technique makes more accurate results than Naïve Bayes alone to predict patients with heart disease. Singh et al. [15] for modern solution developed a web application that enables users to share their heart connected problems. It then specifies the details related to it. Result is given based on the prediction whether the risk of heart disease is low, average or high. Marathe et al. [16] developed a system evaluating the available data (based on sugar level, age, blood pressure etc) collection using Naïve Bayes Algorithm. Apart from the Rstudio's R shiny addon for web UI design, R for coding and the dataset was taken from the University of California at Irvine's Repository. Golande et. al. [17] studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning. Fahd Saleh Alotaibi [18] has designed a ML model comparing five different algorithms. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random Forest and Naive Bayes classification algorithms were compared. Decision tree algorithm had the highest accuracy. Jabbar and Shirina [19] claimed that hidden Naïve Bayes algorithm can be used to predict heart disease and it achieved higher accuracy and dominated naïve Bayes.

In Lutimath et al. [20] two more additional attributes like obesity and smoking are used other than frequently used 13 attributes such as sex, blood pressure, cholesterol and so on for the prediction of coronary diseases. This work is simulated on WEKA 3.6.6 tool. Decision trees, Naïve Bayes are analyzed for heart disease prediction. On the basis of their accuracies, performance of these techniques is compared. This work shows accuracy of Decision Trees to be higher than Naive Bayes is. By analysis of this research work out of these classification models Decision tree algorithm outperformed in heart disease prediction accuracy.

Sharanyaa et al. [21] diagnosed the heart diseases by applying Machine Learning (ML) algorithms such as K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and finally a hybrid of all afore mentioned techniques on a public dataset repository, UCI. The investigation showed hybrid of ML techniques exhibited high accuracy in heart disease prediction. Alotaibi [22] implemented a machine learning model to predict heart failure disease accurately. The investigation was conducted on Cleveland dataset with 14 attributes. The ML algorithms considered for the study are NB, DT, SVM, Logistic Regression, and RF. The class precision and class recall for the ML algorithms as confusion matrix form were tabulated and compared. Comparative results depicted DT outperformed other four models. Harkulkar et al. [23] proposed which detects heart disease based in CNN using Cleveland dataset with 303 instances and achieved a highly accurate result. In a similar work Hussain et al. [24] proposed a novel deep learning architecture using 1D Convolutional Neural Network for classification between healthy and non-healthy persons to overcome the limitations of classical approaches, the system contained an embedding layer which converts the feature vector into the embedding which helps in classification. Later Barhoom et al. [25] performed heart disease prediction on a dataset which consisted of 319795 different patients with different age groups with medical features such as BMI, alcohol, smoking, strokes, physical health, mental health, etc. Their study helped in concluding that the accuracy obtained by deep learning techniques is highest in comparison to other models. Reddy and Sharma [26] performed a work which includes machine learning based classifier to asses classification accuracy and deep learning techniques such as Convolutional Neural Network – Unidirectional Risk Prediction(CNN-UDRP) to enhance the accuracy of heart disease prediction. Durgesh [27] presented Bi-direction long short term memory with Random Forest method in the thesis to enhance the precision of heart disease prediction using Cleveland dataset. Varkala [28]

performed a work with an overall objective to predict the heart disease patients with more accuracy using two datasets Cleveland with 303 records and Statlog with 270 records and concluded that the accuracy can be enhanced by increasing the number of records with attributes to provide better accuracy.

III. WORKING

The suggested study examines the four classification algorithms stated above and performs performance analysis to predict heart disease. This study's goal is to accurately determine whether a patient has heart disease.

The healthcare provider inputs the data from the patient's health report. The information is incorporated into a model that foretells the likelihood of developing heart disease. Figure 1 depicts the full procedure.

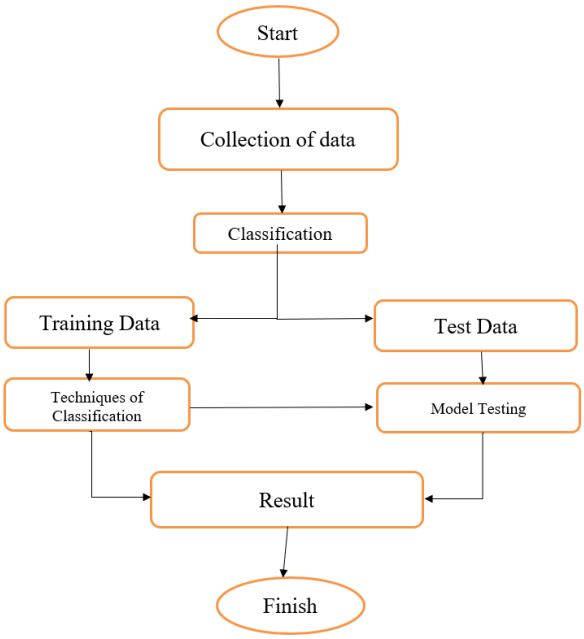

Figure 1

A. Data Collection and Preprocessing

The dataset is from UCI Heart Disease Repository. The dataset contains 302 records based on 13 factors. The data dictionary is below-

1. Age
2. Sex
3. Chest pain Type (4 values)
4. Blood pressure
5. Cholesterol
6. FBS
7. maximum heart rate achieved
8. exercise induced angina
9. oldpeak=ST depression induced by exercise relative to rest
10. the slope of the peak exercise ST segment
11. number of major vessels(0-3) colored by fluoroscopy
12. thal: 3=normal; 6=fixed defect; 7=reversible defect
13. target.

B. Classification

The many ML algorithms, including Random Forest, Decision Tree, Logistic Regression, and Naive Bayes classification methods, receive the attributes as input. Eighty percent of the input dataset is used as training data, and the remaining twenty percent is used as test data.

The dataset used to train a model is referred to as the training dataset.

The performance of the trained model is evaluated using the testing dataset. Accuracy, precision, recall, and F-measure scores are just a few of the various metrics that are used to calculate and analyze the performance of each algorithm.

KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. A supervised machine learning algorithm (as opposed to an unsupervised machine learning algorithm) relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Random Forest

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

$$Gini\ Index = 1 - \sum_{i=1}^{n}(P_i)^2$$
$$= 1 - [(P_+)^2 + (P_-)^2]$$

Logistic Regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. It's an extension of the linear regression model for classification problems. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

$$p = \frac{e^{\beta_0+\beta_1 X_a}}{1+e^{\beta_0+\beta_1 X_a}}$$
$$logit(p) = \log\left(\frac{p}{(1-p)}\right)$$

Naïve Bayes

Naïve Bayes algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B) as shown:

$$P\left(\frac{A}{B}\right) = \frac{\left(P\left(\frac{B}{A}\right) * P(A)\right)}{P(B)}$$

Convolutional Neural Network(CNN)

A Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics.

The architecture of a CNN is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area.

$$x = \frac{w + 2p - f}{s} + 1$$

Where x is the dimension of output features and w is the size of input features. f indicates the size of the filter used for convolutions. 'p' indicates padding which are values added on the boundary before applying convolution. 's' indicates stride which is the value traveled after applying convolution operation.

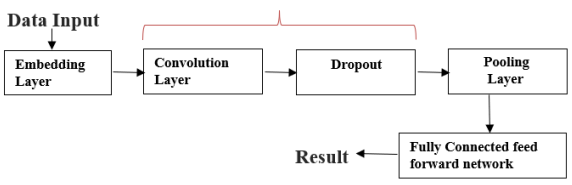

Figure 2

III. Tools and Technology used

i. Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

ii. Jupyter Notebook

Jupyter Notebook is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

iii. Kaggle

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

IV. Result and Discussion

The results will be obtained by applying Random Forest, KNN, Naive Bayes and Logistic Regression as mentioned earlier. The metrics used to carry out performance analysis of the algorithm are Accuracy Score, Precision (P), Recall (R) and F-measure. Precision metric provides the measure of positive analysis that is correct. Recall defines the measure of actual positives that are correct. F-measure tests accuracy. The confusion matrix detects the count of **TP** (True Positive), **TN** (True Negative), **FP** (False Positive), **FN** (False Negative) in the predictions of a classifier.

Accuracy: From Confusion matrix we can derive the **accuracy** which is given by **the sum of the corrected predictions** divided by **the total number of predictions:**

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

Logistic Regression:

The results and accuracy obtained from the Logistic Regression Model Algorithm when tested on the inputs has a accuracy of 77%.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn import metrics
from sklearn.metrics import roc_auc_score, confusion_matrix, accuracy_score, roc_curve

logit = LogisticRegression()
logit.fit(X_train, y_train)

predicted_logit = logit.predict(X_test)

LogisticRegressionScore = accuracy_score(predicted_logit, y_test)


#plt.figure()
#metrics.plot_roc_curve(logit, X_test, y_test)
#plt.title("Receiver Operating Characteristic (ROC)")
#plt.show()

print("Logistic Regression score: ", LogisticRegressionScore)
```

```
Logistic Regression score:  0.7708333333333334
```

KNN:

The results and accuracy obtained from the KNN Model Algorithm when tested on the inputs has a accuracy of 72.9%.

KNN

```
[26] from sklearn.neighbors import KNeighborsClassifier


KNC = KNeighborsClassifier(n_neighbors=2)
KNC.fit(X_train, y_train)

KNC_pred = KNC.predict(X_test)

KNC_accuracy = metrics.accuracy_score(y_test, KNC_pred)

print("KNeighbourClassifier score: ", KNC_accuracy)
```

```
KNeighbourClassifier score:  0.7291666666666666
```

Naïve Bayes:

The results and accuracy obtained from the Naïve Bayes Model Algorithm when tested on the inputs has a accuracy of 62.5%.

Naive Bayes

```
[35] from sklearn.naive_bayes import GaussianNB

gauss = GaussianNB()
gauss.fit(X_train, y_train)

gauss_pred = gauss.predict(X_test)

gauss_score = accuracy_score(gauss_pred, y_test)


#plt.figure()
#metrics.plot_roc_curve(gauss, X_test, y_test)
#plt.title("Receiver Operating Characteristic (ROC)")
#plt.show()


print("Gaussian Naive Bayes score: ", gauss_score)
```

```
Gaussian Naive Bayes score:  0.625
```

Random Forest:

The results and accuracy obtained from the Random Forest Model Algorithm when tested on the inputs has a accuracy of 68.75%.

Random Forest

```
[27] from sklearn.ensemble import RandomForestClassifier

rnd_clf = RandomForestClassifier(n_estimators=500, max_leaf_nodes=16, n_jobs=-1)
rnd_clf.fit(X_train, y_train)

rnd_clf_pred = rnd_clf.predict(X_test)

rnd_clf_accuracy = metrics.accuracy_score(y_test, rnd_clf_pred)
print("RandomForest score: ", rnd_clf_accuracy)
```

```
RandomForest score:  0.6875
```

CNN:

Along with the above four algorithms, the result obtained from the Convolutional Neural Network(CNN) model, the accuracy obtained is 80.33%. The accuracy obtained in case of CNN is highest in comparison to any other machine learning algorithm.

```
score_nn = round(accuracy_score(Y_pred_nn,Y_test)*100,2)

print("The accuracy score achieved using Neural Network is: "+str(score_nn)+" %")

#Note: Accuracy of 85% can be achieved on the test set, by setting epochs=2000, and number of noc
```

```
The accuracy score achieved using Neural Network is: 80.33 %
```

The focus of our study was on using machine learning techniques in healthcare for heart disease. We performed some experiments on our data set of heart disease by applying four machine learning algorithms. Through implementation of different classification algorithms we try to find out that which algorithm is best in predicting heart disease. And which one gives the best accuracy. There are four experiments we performed and these experiments are designed for the same purpose, the purpose is to compare the results of Logistic Regression, KNN, Naive Bayes and Random Forest. In addition to these four models the experiment is performed using CNN model also to obtain higher accuracy.

Through implementation we can know which classification algorithm is best for predicting heart disease. After the implementation of different algorithms the second step is the comparison between different machine learning algorithms used in these experiments and choose the best one which gives most accuracy. In order to do comparison of these experiments different performance measures are used for example, Accuracy True Positive, False Positive, False Negative, True Negative is used. Summary of classification algorithms is shown in the following table.

Table 1

Algorithms	Accuracy	TN	TP	FN	FP
Logistic Regression	77	20	26	5	9
Naïve Bayes	62.5	17	23	10	14
Random Forest	68.75	18	26	8	12
KNN	72.9	20	27	8	10

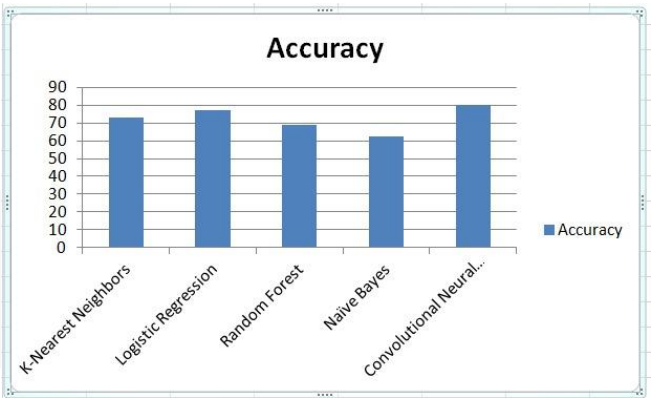

Figure 3

Conclusion:

To summarize, the project presented a heart disease prediction approach using Logistic Regression, Random Forest, Naïve Bayes, KNN. In addition to these models Convolutional Neural Network is also used for prediction of heart disease and hence for comparative study which algorithm is better for prediction. The present work is evaluated on Heart Disease Prediction from Cleveland UCI repository dataset collected from Kaggle. When compared to the performance of the conventional machine learning approaches regarding heart disease prediction, Logistic Regression approach resulted in improved and reliable heart disease prediction as the accuracies obtained by Logistic Regression is 77%, by Random Forest is 68.75%, by KNN is 72.9%, by Naïve Bayes is 62.5%. But when Convolutional Neural Network(CNN) algorithm was used the accuracy obtained was highest amongst all the other four algorithms with the accuracy of 80.33%. Henceforth, the proposed approaches have probable of diagnosing the heart disease to aid cardiologists but the CNN technique is best amongst all machine learning techniques for heat disease prediction.

REFERENCES

[1] K.V. Siva Prasad Reddy, G. Yamini, T. Nandini, V. Thanuja, International Journal for Research in Applied Science and Technology (IJRASET) ISSN:2321-9653; IC Value:45.98; SJ Impact Factor: 7.538 Volume 10.

[2] Yingjie Zhang et al 2021 J. Phys. :Conf. Ser 1769012024 'Logistic Regression Models in Predicting Heart Disease'

[3]Yeshvendra Singh, Nikhil Sinha, Sanjay Kumar Singh, July 2017. Communications in Computer and Information Science DOI:10.1007/978-981-10-5427-3_63 International Conference on Advances in Computing and Data Sciences

[4]Indu Yekkala, Sunanda Dixit, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection", International Journal of Big data and Analytics in Healthcare, 3(1):1-12, DOI: 10.40181IJBDAH.2018010101

[5]Madhumita Pal, Smita Parija, "Prediction of Heart Disease using Random Forest", ICCIEA 2020, J.Phys.Conf.Ser.1817012009

[6]Sibgha Taqdees, Nayab Akhtar, Kanwal Dawood, "Heart Disease Prediction", https://www.researchgate.next.publication/349140147

[7]Jaishree Pandhari Wankhede, Palaniappan S, Magesh Kumar S, "Heart disease Prediction using Hybrid Random forest Model Integrated with Linear Model", Advances in Parallel Computing Technologies and Applications, D. J. Hemanth et al. (Eds.)

[8]A. Sankari Karthiga, Safish Mary, M. Yogasini, "Early Prediction of Heart Disease Using Decision Tree Algorithm", International Journal of Advanced Research in Basic Engineering Sciences and Technology(IJARBEST), Vol.3, Issue. 3, March 2017, ISSN: 2395-695X

[9] Kennedy Ngure Ngare, " Heart Disease Prediction System", https://www.researchgate.net/publication/331589020

[10]K. Polaraju, D.Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research, ISSN: 2321-9939, 2017 IJEDR| Volume 5, Issue 4

[12]Deepika Ponnusamy, K. Thenmozhi, "Heart Disease Prediction Using Classification with different Decision Tree Techniques", International Journal of Engineering Research and General Science, Volume 2, Issue 6, October-November, 2014, ISSN: 2091-2730

[13]V. Sabarinathan, V. Sugumaran, "Diagnosis of Heart Disease Using Decision Tree", International Journal of Research and Computer applications and Information Technology, Volume 2, Issue 6, November-December, pp.74-79, ISSN: 2347-5099, DOA: 27122014

[14]Viney Churian, Bindu M.S., "Heart Disease Prediction using Naïve Bayes and Laplace Smoothing Technique", International Journal of Computer Science Trends and Technology(IJCST)- Volume 5, Issue 2, Mar-Apr2017

[15] Garima Singh, Kiran Bagwe, Shivani Shanbhag, Shraddha Singh, Sulochana Devi, "Heart Disease Prediction Using Naïve Bayes", International Journal of Engineering and Technology(IRJET), Volume:04, Issue:03, Mar-2017, ISSN: 2395-0056

[16]Ninad Marathe, Sushopati Gawade, Adarsh Kanekar, "Prediction of Heart Disease and Diabetes using Naïve Bayes Algorithm", International Journal of Scientific Research in computer Science, Engineering and Information Technology, ISSN: 2456-3307

[17] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[18] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[19]Jabbar Akhil, Shirina Samreen, "Heart disease prediction system based on hidden naïve Bayes classifier", https://www.researchgate.net/publication /309735105_Heart_disease_prediction_system_based_on_hidden_naive_Bayes_classifier, October 2016.

[20] Kim, H. Y. (2014). Analysis of variance (ANOVA) comparing means of more than two groups. Restorative dentistry & endodontics, 39(1), 74-77

[21]Sharanyaa S, Lavanya S, Chandhini MR, Bharathi R, Madhulekha K. Hybrid Machine Learning Techniques for Heart Disease Prediction. International Journal of Advanced Engineering Research and Science. 2020;7(3):44-8.

[22]Alotaibi FS. Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications. 2019;10(6):261-8.

[23]Nilam Harkulkar, Swati Nadkarni, Dr. Bhavesh Patel, "Heart Disease Prediction using CNN, Deep Learning Model", "International Journal for Research in Applied Science and Engineerig Technology" (IJRASET), ISSN:2321-9653; Vol 8 Issue XII Dec 2020

[24]Shadab Hussain, Susmith Barigdad, Shadab Akhtar, Md Snaib, "Novel Deep Learning Architecture for Heart Disease Prediction using CNN"

[25]Ali M.A. Barhoom, Abdulbaset Almasri, Bassum S. Abu Nasser, Samy S. Abu Nasser, "Prediction of Heart

Disease using a collection of machine learning and deep learning algorithms".
[26]V Archana Reddy, K Venkatesh Sharma, "Heart Disease Classification and Risk Prediction By Using Convolutional Neural Network", International Journal of Aquatic Science, ISSN:2008-8019, Vol 12, Issue 02, 2021
[27]Durgesh Kumari, "Study of Heart Disease Prediction using CNN algorithm", ISSN:2349-5162, Volume 8, Issue 7, JETIR July 2021
[28]Dr. Krishaiah Varkala, "Heart Disease Prediction System Using Convolutional Neural Networks", DOI: https://doi.org/10.21203/rs.3.rs-2009078/v2