Effective Prediction of Heart Disease using Machine/Deep Learning Techniques

Kanika Mittal, Akshaj Anil Patil, Dhawal Singh Pundir, and Nitindeep Singh

Department of Computer Science and Engineering
Indraprastha Institute of Information Technology
Delhi, India

{kanika18075, akshaj19111, dhawal19120, nitindeep19069}@iiitd.ac.in

***Abstract*—Heart disease is amongst the most critical conditions which have a significant impact on human life all over the world. It also accounts for being the reason for mortality and morbidity. Thus, it is necessary to diagnose heart diseases in a timely and accurate manner, which can help in the prevention of such heart attacks. But the prediction of such heart disease proposes significant challenges since a massive amount of data is produced by the healthcare industry daily. Data mining helps in transforming the extensive collection of raw healthcare data into valuable information, which can help in making versed decisions and predictions. Moreover, many factors, such as high cholesterol, diabetes, high blood pressure, etc. also contribute to the occurrence of heart diseases. These constraints have led scientists to turn to techniques like machine learning. Machine learning proves to be useful in assisting in making predictions and decisions from a vast ocean of data produced by the healthcare industry.**

In this paper, we are proposing machine and deep learning-based models for heart disease prediction, which aim to perform binary classification - people with heart disease and people without heart disease. We have used two datasets, Cleveland heart disease and Statlog heart disease datasets from the UCI repository. The structure of our methodology is as follows - preprocessing of the datasets, feature selection, cross-validation, applying machine and deep learning classifiers, evaluating classifiers' performance. We have performed classification using popular machine learning models - ANN, SVM, Random Forest, Naive Bayes, Logistic Regression, Decision Trees, and deep learning model - LSTM.

I. Introduction

Data mining has now been around for three decades. But the real importance and application of it are valued more than ever it had in the past. As we consider various definitions of data mining, we can come to a more generic definition of it i.e. the process of extracting valuable information with generic or some specific pattern from the datasets that was unknown till date and using different techniques to perform extensive analysis on the data available to get information about various aspects of the data and the best part is that the process after certain training is automated to further extract knowledge from various datasets in the specified format for the model. The data mining is used with intelligent machine learning and AI-based algorithms that make the process more efficient and fast. With the rapidly increasing number of patients suffering from heart disease, there is a need of an efficient system that can predict heart disease with high accuracy. According to WHO, 17 million people die annually because of heart attacks and strokes [1]. The health industry is facing Quality of service issues. Poor or wrong diagnostics have disastrous consequences. Diagnostics have to be accurate and efficient. The decisions are generally made on doctor's knowledge and experience rather than the abundant knowledge present in the datasets and records we have. These decisions often lead to unwanted circumstances and additional medical expenses. So to overcome all these situations and have an extra edge over this, we need a system that is efficient enough to predict the heart disease using the knowledge-rich data present.

Various systems have been developed and studies show the prediction of heart disease using data mining techniques. Some of the systems and studies are as bellow:

- An Intelligent heart disease prediction was developed by Sellappan Palaniappan et al. [2] which uses various data mining techniques such as Naïve Bayes, Neural Network and Decision Tree. Each of these methods has its strength and weaknesses for the final results. In the system, hidden patterns and the relation between them is used.
- Heon Gyu Lee et al. in [3] proposed a linear and non-linear technique for generating features of Heart Rate Variability Activity. They analyzed the recumbent positions for HRV Indices. They used SVM and Bayesian classifiers to produce state-of-the-art results.
- Classification method for extraction of multiparametric features by accessing the HRV from ECG, preprocessed data and heart disease pattern is used by Kiyong Noh et al in [4]. They used the thickness measurements of arteries as novel features alongwith the existing datasets.
- The heart disease and other predictions model with the aid of a neural network was proposed by Nitin Guru et al. [5]. The data set has each record containing 13 attributes. They used K-Means clustering algorithm to extract appropriate data from the datasets. They tried to analyze pattern among different patients to develop better classification models using Neural Network and Deep Learning.

The limitations of these existing systems vary from one

aspect to the other, but the major drawbacks in the systems were, one being the systems are developed for a certain number of attributes and are not capable of taking any number of attributes and utilizing the best attributes for prediction. The second limitation is that most systems use the categorical data, but for some conditions and diagnostics, continuous data may be necessary. The last being that the systems are majorly using only data mining techniques while as in today's era, the data mining techniques can be incorporated with machine learning and AI techniques to fetch better and accurate results.

We have tried to overcome the limitations by projecting the data from high-dimensional subspace to lower dimensions. It was achieved by projecting the data into the direction of maximum variance of Eigen vectors. It also helps to reduce the noisy dimensions and hence improves the classification performance on the various metrics which was discussed in [6]. We have also converted the categorical values to continuous values using different methods like Standard Scaling and Feature hashing. Standard Scaling of the feature values is done so that none of the feature value dominates among other specially in Neural network and deep learning models. In the proposed prediction model, in addition to data mining techniques, we have incorporated state-of-the-art deep learning models like LSTM as also proposed in [7] to improve the classification performance of the system.

II. Materials and Methods

The proposed architecture is explained visually in Fig. 1. The pipeline consists of various stages like data collection, feature extraction, cross validation and model selection. We have used various evaluation metrics to rank the models. The details about each section is explained below.

A. Dataset Used

Heart disease prediction is performed over two different datasets. The datasets are obtained from the UCI Machine learning repository [8]. These datasets are merged into a final dataset. The first data set is taken from the Cleveland Heart Disease database [9] which consists of 303 records and the second dataset is taken from Statlog Heart Disease database [10]. The second dataset contains 270 records. The datasets obtained from both the sources have the same 13 features as described in Table I. The datasets were merged and shuffled for all the experiments.

B. Feature Generation Techniques

We explored various feature generation and selection techniques for getting state-of-the-art classification performance. We analyzed various techniques like Principal Component Analysis (PCA) for transforming the features to separate lower-dimensional subspace. We also implemented SelectKBest feature techniques along with conversion of categorical to continuous values. We experimented and selected the techniques that performed the best on evaluation. For categorical to continuous values conversion, we employed various techniques like scaling the values using Z-score, cosine-transformation, One-hot encoding. In PCA, we experimentally found that using about 77% of Eigen energy produces the best results.

TABLE I
Dataset attributes description

	Name	Description	Type	Range/Domain
1	age	Age of patient	Continous	28 <age<78
2	sex	Genger of patient	Categorical	Female = 0 Male = 1
3	cp	Type of chest pain	Categorical	Typical angina = 1 Atypical angina = 2 Non-anginal pain = 3 Asymptomatic = 4
4	trestbps	Resting blood pressure (in mm/Hg)	Continous	93 <trestbps <201
5	chol	Serum cholestoral (in mg/dl)	Continous	125 <trestbps <565
6	fbs	Fasting blood sugar >120 (in mg/dl)	Categorical	False = 0 True = 1
7	restecg	Resting electrocardiographic results	Categorical	Normal = 0 ST-T wave = 1 Hypertrophy = 2
8	thalach	Maximum heart rate achieved (in mg/dl)	Continous	70 <thalach <203
9	exang	exercise induced angina	Categorical	No = 0 Yes = 1
10	oldpeak	ST depression induced by exercise relative to rest	Continous	0.0 <oldpeak <6.3
11	slope	Slope of the peak exercise ST segment	Categorical	Upsloping = 1 Flat = 2 Downsloping = 2
12	ca	Number of major vessels colored by flouroscopy	Categorical	[0,1,2,3]
13	thal	Type of defect	Categorical	Normal = 3 Fixed = 6 Reversable = 7

C. Cross-validation technique

K Fold cross-validation is a method that divides the data in the test and validation set. Suppose the data set is divided into k subsets. Now, the validation method iterates for k times and in each iteration it will consider one of data fold as the testing data set and the remaining k-1 data folds as the training data set. As data set attributes have a large variation in their values or in classification there may be more negative values than positive. To handle such a situation a slight change in K Fold cross-validation is introduced. Due to this change, each fold from k folds contains an approximately equal percentage of data points of each class. The mean response value is also approx. equal for all the k folds while predicting the result. This change introduced in K Fold is known as Stratified K Fold cross-validation.

We used 5-fold cross validation technique. For every iteration, 80% data is available for training and remaining 20% is used for validation. [11] used similar technique but with increased number of folds (k=10), however, that results in very less data available for validation giving a false impression of better classification by the models.

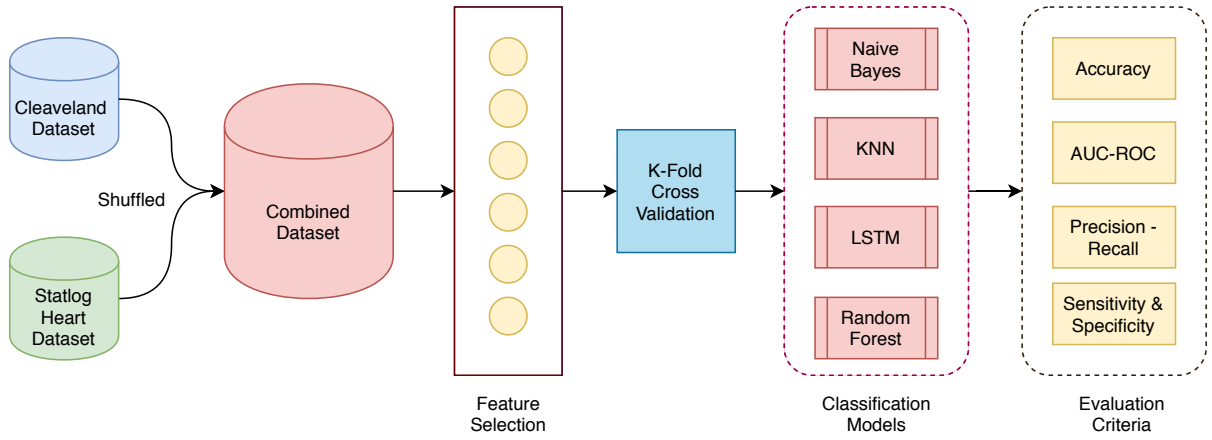

Fig. 1. Model Architecture

D. Classification Techniques

The above dataset analysed using various data mining classification techniques such as Artificial Neural Networks, Decision Trees, Naïve Bayes, Random Forest, SVM, Logistic Regression, etc.

*1) **Artificial Neural Networks**:* A machine learning technique that is used to predict the output through computational methods is artificial neural networks as proposed in [12]. It is based on the biological neural network, which consists layer of interconnected neurons. Similarly, the artificial neural network also consists of layers i.e. input layer, an output layer, and if required some hidden layers. Artificial neural networks required appropriate weight to classify the training data in the correct way. ANN processes the training data to map it on the appropriate result. A multilayer perceptron neural network is used to predict the output.

In this model we have used one input layer, 3 hidden layers and one output layer to predict a patient is having heart disease or not. Nodes present in one layer are interconnected with the nodes present in the next layer i.e. nodes of the input layer are interconnected with the hidden layer nodes and hidden layer nodes are interconnected with the output layer nodes. This connection exists between the nodes of different layers hold some weights. To produce an output this model will use simple threshold functions.

*2) **Naïve Bayes**:* It is the most famous supervised classifier which is based on the Naïve Bayes theorem. This classifier is based on the probabilistic prediction which uses conditional independence. The values of attributes on a class are independent of the values of attributes of other classes. This model is efficiently working for data which is having unrelated features and high noises. So, the features assumed to be independent and equally important to produce the outcome. Bayes theorem finds the probability of an occurring event given the probability of an event that has already occurred. Bayes theorem is expressed as –

$$Pr(\frac{H}{X}) = \frac{Pr(\frac{X}{H}) * Pr(H)}{Pr(X)}$$

Where H and X are events and Pr(X) not equal to 0. This network can also predict the result for the incomplete data as it is a directed acyclic graph of random variables. Performance of Naïve Bayes classifier for heart datasets is well experimented in [13] and hence, we used it for our experiments as well.

*3) **Decision Trees**:* The decision tree is an efficient classification and prediction tool as proposed in [14], [15]. In this model first, a decision tree is created then that tree is applied on the dataset. The decision tree contains some rules which finally predict the result. The nodes of the decision tree contain some set of questions related to the features of the dataset and these questions help in the classification of the data. Classification of the dataset is basically performed by finding a path from the root to the leaf node in the decision tree. A number of algorithms are present that follows the logic of the decision tree. During the classification of data, there are chances of over-fitting of data that will be reduced with the help of a technique called pruning. The algorithm for decision tree recursively classifies data until the appropriate prediction result occurred.

*4) **Recurrent Neural Network**:* This neural network is also a feed-forward neural network and it contains an internal memory. A recurrent neural network as the name suggests is recurrent in nature and uses the previous output as input to compute results. This network considered both the previous output and the current output to predict the result of a particular iteration until it gets the most appropriate result. This model uses its internal memory to process the input sequences. This network has many types. Here, we have used Long Short-Term Memory networks. LSTMs are designed in such a way that these avoid long-term dependency problem. This model remembers the information for a long period of time. Modules are repeated in the recurrent neural networks.

LSTMs have been widely used in heart disease prediction as they tend to find out more crucial insights and patterns in the datasets which are useful to find out early symptoms of heart disease as explained in [16], [17].

*5) **Random Forest**:* Random forest [18] contains one or more regression trees of unpruned classification which is made

by a random selection of training data samples. By induced process, random features are selected. Aggregation is used to predict the result. For growing the tree N number of random samples used is taken from the original data. Total m number of variables are selected from the M number of input variables randomly such that m«M. This value of m will remain constant during the growth of the tree. Without using pruning trees grows as large as possible. Random forest performs better than a single tree classifier.

6) ***Logistic Regression****:* This is an important classification algorithm in context to heart disease prediction [19]. Different algorithms for supervised classification problems are used to find out the decision boundaries for the different classes. These decision boundaries divide data of one class from others. In logistic regression the shape of these decision boundaries is linear. The feature vector for a training sample has a dimension equal to the number of elements in the training sample. S-shaped logistic function is used to assign values 0 or 1 to each feature of the vector. The probability of data belongs to a particular class is the value. To correctly classify the data learning algorithm will assign some weights to each data. After assigning weights, the logistic function is applied to the remaining dataset to predict the class of that data.

7) ***Support Vector Machine****:* This is a type of classifier based on the hyperplane and has been used widely in heart disease prediction like [20], [21]. A support vector machine is a supervised learning method that results in an optimal hyperplane. Here, a hyperplane is used to categorize the new data. A hyperplane is defined as a line that divides data of two different classes in a 2D plane. To find hyperplane in an N-dimensional plane that divides the data points of different classes is the objective of the support vector machine. It is possible that between the data points of two different class more than one hyperplane can be possible. But to find a hyperplane that has the maximum distance between the points of different classes is the main objective of SVM so that it becomes easy to find hyperplane for new data points.

E. Evaluation Parameters

Various Evaluation metrics are used in the paper. We have used confusion matrix so that we can predict every observation in testing set. We get 2x2 matrix which contains TP(True Positive), FP(False Positive), FN(False Negative) and TN(True Negative).

1) ***Accuracy****:* Accuracy shows how many data points are predicted correctly. It represents the overall performance of classification system.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

2) ***Specificity****:* Specificity is also call as TNR (True Negative Rate). It is calculated as number of True negative prediction divided by total negative prediction.

$$Specificity = \frac{TN}{TN + FP}. \tag{2}$$

1-specificity is used to plot ROC curve.

3) ***Sensitivity****:* Sensitivity is also called as RECALL. Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category. It is the TPR(True Positive Rate).

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

4) ***Precision****:* Precision represents the percentage of the results which are relevant. Precision is calculated as :

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

5) ***AUC-ROC****:* AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) is performance measurement of classification problem at various threshold. ROC is probability curve and AUC represents Degree of measure of separability.

$$TPR = sensitivity \tag{5}$$

$$FPR = 1 - specificity \tag{6}$$

ROC curve is plotted with TPR against FPR, TPR on y-axis and FPR on x-axis.

III. Results

TABLE II
PERFORMANCE OF CLASSIFIERS ON DIFFERENT EVALUATION METRICS

Models	Accuracy	Specificity	Sensitivity	AUC	Precision
Base (Naive Bayes)	94.44	93.54	95.65	-	91.66
Base (Decision Tree)	96.66	96.50	96.85	-	96.09
Base (ANN)	99.25	98.69	100	-	98.31
ANN	89.56	95.77	79.54	91.35	92.10
Decision Tree	**99.91**	**99.41**	**100**	99.34	**99.94**
Naive Bayes	90.43	96.82	75	92.49	95.12
Logistic Regression	91.30	95.31	96.27	92.55	93.61
SVM	72.17	85.50	52.17	77.85	70.58
Random Forest	99.84	99.25	99.87	**99.47**	**99.94**
KNN	86.08	86.79	85.48	91.84	88.33
LSTM	91.42	92.43	84.36	92.72	83.24

This section explains and analyses the various evaluation metrics on the various classification models we used for the experimentation. We have performed various experiments and tabulated the results in Table II. It clearly shows that our proposed model perform better than the baseline model [22] all the evaluation metrics. The Table II shows a clear view of our proposed models outperforming the baselines by a large margin on specificity and precision and substantial gain on Accuracy.

Fig.2 shows the ROC curves for binary classfication of all the classifiers performing good in our experiments. The ROC Curve for Decision Tree and Random Forest are the best as they produce near perfect results as shown in Table II. The ROC curve for SVM is very close to random guess and hence it signifies poor performance of SVM on the dataset.

The Precision-Recall curve (Fig. 3) is very important objective function for medical science related experiments. It is because the model must always try to reduce the false negatives even at the cost of increase in false positives. We can see that

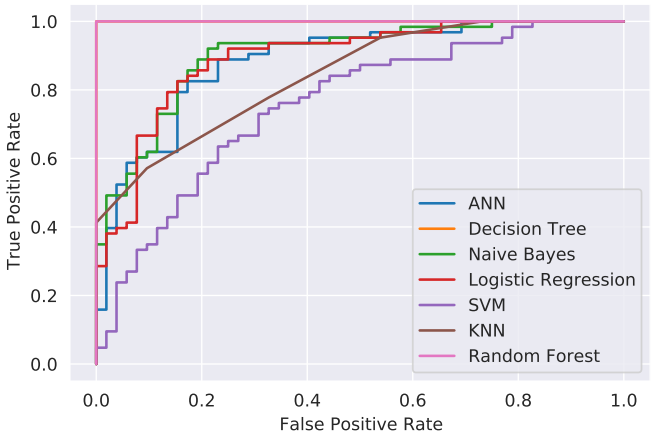

Fig. 2. ROC Curve for different classifiers

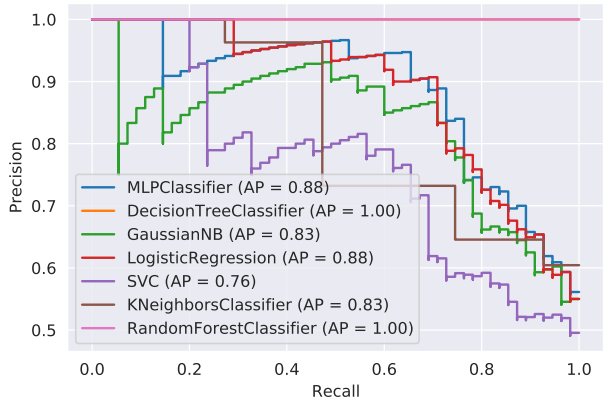

Fig. 3. Comparison of Precision Recall Curve for different classifiers

SVM again produces unsatisfactory results. Gaussian Naive Bayes although has more than 90% classification accuracy but has a recall of 75% which is not acceptable for a model predicting heart disease.

Classification Accuracy is considered one of the important characteristics for evaluating a model's performance. Fig. 4 shows a comparison of accuracy for all the models.

Matthews correlation coefficient is an effective way to measure the quality of a classifier in case of binary classification. Fig. 5 shows that the Decision Tree and Random Forest perform best among all the classifiers.

Fig. 6 shows the confusion matrix for all the classifiers. We can observe that we get identical no false negatives for both Decision Tree and Random Forest and that is one of the most sought after criteria for any medical science related machine learning models. We get slightly better results with Naive Bayes but it is very far from being used as a real-time model.

IV. Discussion

In this paper, we have proposed machine and deep learning based system for heart disease prediction.The proposed system

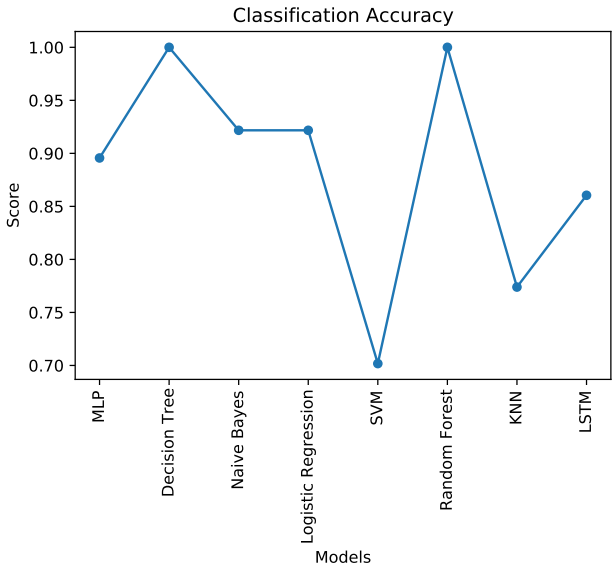

Fig. 4. Comparison of Classification Accuracy for different classifiers

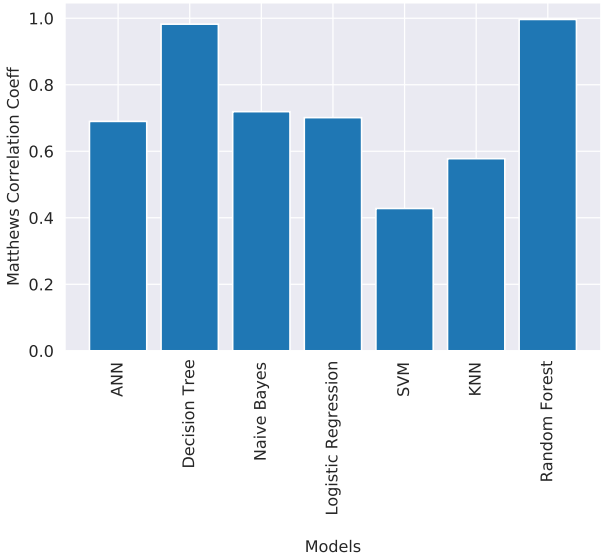

Fig. 5. Comparison of Matthews Correlation Coefficient for different classifiers

is tested on the Cleveland and Statlog datasets taken from the UCI repository. Various evaluation measures are taken into account for the performance of the system. The models used in the system are - ANN, Decision Tree, Naive Bayes, Logistic Regression, SVM, Random Forest, KNN, LSTM.

Decision Tree outperforms all other classifiers in accuracy, Specificity and Sensitivity while Random Forest performs the best in AUC. Decision Tree and Random Forest have performed equally better with very negligible difference in all the evaluation metrics. Decision Tree being reasonably fast to train performs very well with prediction systems while as the naive bayes works well in domains like Computer Vision

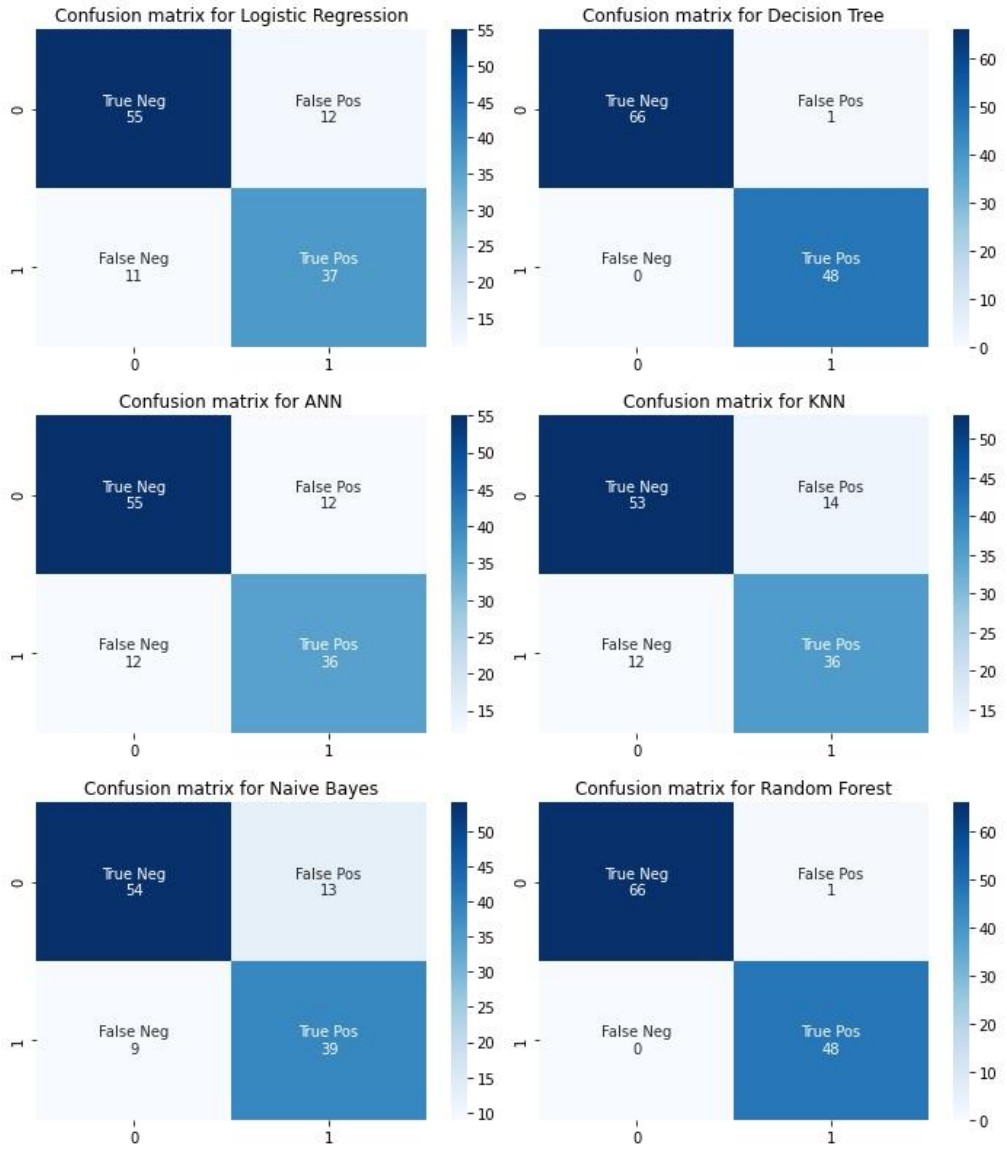

Fig. 6. Confusion Matrix for different classifiers

and Robotics. Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. Two models have worked identically in terms of Precision measure i.e Decision tree and Naive Bayes. Both of the models are well versed working with large data sets and prediction systems, the decision tree being more fast in training time than random forest model.

Advanced feature selection techniques like Recursive Feature Selection, Minimal-Redundancy-Maximal-Relevance Feature Selection, LASSO can be used in the system to improve the accuracy of various parameters. In the future, we can perform more experiments with the system incorporating new and efficient feature selection and deep learning models for achieving highly accurate and efficient heart disease prediction system.

V. Contribution of each author

Section Wise Contribution of each author:

- **Dataset Collection and Preprocessing:** Akshaj Patil, Kanika Mittal
- **Feature extraction and Feature Selection:** Dhawal Singh Pundir, Nitindeep Singh
- **Experiments related to cross validation:** Nitindeep Singh
- **Implementation of Models for classification:** Kanika Mittal
- **Analyzing Models for classification:** Akshaj Patil

- **Generation of Graphs, Tables:** Dhawal Singh Pundir, Kanika Mittal
- **Presentation:** Dhawal Singh Pundir, Nitindeep Singh, Kanika Mittal
- **Research Paper Writing:** Akshaj Patil, Dhawal Singh Pundir, Kanika Mittal, Nitindeep Singh

REFERENCES

[1] "Cardiovascular diseases (CVDs) WHO update." [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, no. April, pp. 108–115, 2008.

[3] H. Lee, K. Noh, H. Park, and K. Ryu, "Predicting coronary artery disease from heart rate variability using classification and statistical analysis," 11 2007, pp. 59–64.

[4] H. G. Lee, K. Noh, and K. H. Ryu, "A data mining approach for coronary heart disease prediction using hrv features and carotid arterial wall thickness," *2008 International Conference on BioMedical Engineering and Informatics*, vol. 1, pp. 200–206, 2008.

[5] B. Shantakumar and Y. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of Scientific Research*, vol. 31, pp. 642–656, 01 2009.

[6] N. Ziasabounchi and I. Askerzade, "A comparative study of heart disease prediction based on principal component analysis and clustering methods," *Turkish Journal of Mathematics and Computer Science*, vol. Article ID 20140043, 11 pages, 09 2014.

[7] I. Javid, A. K. Z. Alsaedi, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110369

[8] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[9] "Heart disease data set: Cleveland." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[10] "Statlog (heart) data set." [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/

[11] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018.

[12] J. P. Bigus, "Data Mining with Neural Networks," *New York*, vol. 5, no. 1, pp. 1–154, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=231007

[13] J. Bohacik and M. Zabovsky, "Naive bayes for statlog heart database with consideration of data specifics," in *2017 IEEE 14th International Scientific Conference on Informatics*, 2017, pp. 35–39.

[14] A. Karthiga, S. Mary, and 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, 04 2017.

[15] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," in *Information and Communication Technology for Competitive Strategies*, S. Fong, S. Akashe, and P. N. Mahalle, Eds. Singapore: Springer Singapore, 2019, pp. 447–454.

[16] E. Zriqat, A. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," 04 2017.

[17] M. Khan, "An iot framework for heart disease prediction based on mdcnn classifier," *IEEE Access*, vol. PP, pp. 1–1, 02 2020.

[18] I. Yekkala and S. Dixit, "Prediction of heart disease using random forest and rough set based feature selection," *International Journal of Big Data and Analytics in Healthcare*, vol. 3, pp. 1–12, 01 2018.

[19] M. Thirugnanam, "A heart disease prediction model using svm-decision trees-logistic regression (sdl)," *International Journal of Computer Applications in Technology*, vol. 68, pp. 11–15, 04 2013.

[20] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 3107–3111.

[21] J. Vankara and G. L. Devi, "PAELC: Predictive analysis by ensemble learning and classification heart disease detection using beat sound," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 31–43, Jan. 2020. [Online]. Available: https://doi.org/10.1007/s10772-020-09670-6

[22] C. Dangare and S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47, pp. 44–48, 06 2012.