IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

[1]Sandeepa Oraon, [1]Suman Kumari, [2]Siba Mitra, [2]Jaya Pal, [3]Samreen Anjum

[1]MCA Student, [2]Assistant Professor, [3]MBA Student
[1, 2]CSE Department, [3]Management Department
[1]Galgotias University, Greater Noida, UP, India
[2]Birla Institute of Technology, Mesra, Ranchi, India
[3]MBA student, XISS, Ranchi, India

Abstract: Heart is a vital organ in human body that circulates blood throughout the body. It plays an important role in supplying oxygen and nutrients to the tissues and organs while removing waste products. Cardiovascular diseases, encompass a range of conditions that affect the heart and blood vessels. These diseases can disrupt with the heart's ability to function properly and may lead to serious health complications. In this study, for the purpose of predicting heart disease, we apply various kinds of machine learning algorithms. After completion of the pre-processing, several classification models were applied to the dataset. Based on Accuracy and other evaluation criteria, we can identify the most effective classifier. We found that Random Forest give the highest accuracy score.

Index Terms **– ML algorithms, Cardiac Diseases, Feature Extraction, Prediction Model, Performance Metrics**

I. INTRODUCTION

Heart disease is the leading cause of death in India and around the world, accounting for a significant proportion of mortality across all age groups. Various researches and studies have projected that the mortality rate in India due to cardiovascular diseases will take a sharp growth from 22.6 lakhs in 1990 to 47.7 lakhs in 2022 and it has actually reached there. However according to the estimates demonstrated by many researchers over the past few decades, the rate of cases of coronary heart disease in India has varied in between 1.6 percent to 7.4 percent in the rural areas and in the urban arrears it varied between 1 percent and 13.2 percent. Now in this competitive world the sharp growth rate in cardiac disease is normal because of the stress, tension, food habit and life style of the individuals. Some cardiac disease prediction models are mentioned in [22] and [23]. The muscular organ that comprises up the human heart is situated in the chest, somewhat to our left side, and is about the size of a closed fist. It is divided into four chambers: two auricles (upper chambers) and two ventricles. Age, family history, smoking, eating a poor diet, not exercising, bring obesity, having high blood pressure, high cholesterol, and diabetes are risk factors for heart disease. Lifestyle modifications, medications, and medical procedures are often used to manage and treat heart diseases. Early detection and intervention are crucial for improving outcomes and preventing complications. Advancements in computational biology and machine learning algorithms have revolutionized the early prediction of heart disease, enabling more precise and timely interventions for improved patient outcomes.

Machine learning (ML) algorithms play a pivotal role in extracting meaningful insights from complex datasets. In the research work mentioned in [14] the authors have mentioned a detailed study on deep learning attributes and various networks in computing technologies. They have also demonstrated various applications of the deep learning in the real world scenario. They can discern intricate correlations between genetic markers, lifestyle factors, and early physiological changes that may precede overt symptoms of heart disease. Early identification of high-risk individuals facilitates targeted interventions, such as lifestyle modifications, medication management, or closer monitoring, to mitigate the progression of heart disease. As mentioned in [18] the researchers have proposed a heart disease prediction model using ML approaches and have utilized the Cleveland data repository for training and testing purpose. However Random Forest gave the best result.

The research in [19] presents a vivid study on supervised ML algorithms and they have also demonstrated a comparative study of the same. Seven different types of algorithms are considered for the review process, which includes, Support Vector Machine, Decision Table, Naïve Bayes, and Decision tree, Random Forest, Neural Networks and JRip using Diabetes data set. Some of the research work mentioned in [20] and [21] also demonstrates about the life cycle of ML model and the application of ML in solving real life problems. In order to apply machine learning techniques for heart disease prediction, detection and diagnosis a good number of supervised ML algorithms available. In this research work approaches like logistic regression, decision tree, random forest Naïve Bayes, neural network and gradient boosting were employed in order to develop a model for heart disease prediction. After that a comparative study of the algorithms' performance is demonstrated so that the most effective approach can be chosen out of all. The heart disease data was acquired from the Kaggle database, and the obtained output was used to evaluate accuracy when classification and prediction processes were performed.

II. LITERATURE SURVEY

The prediction and detection of heart disease have garnered significant research interest due to the increasing prevalence worldwide. Early prediction of heart diseases can substantially decrease the risk of mortality by enabling timely intervention and treatment. In the research article [1], in order to predict heart disease using ML algorithms, three ML modeling techniques are used to propose a cardiovascular disease prediction model- KNN, Logistic Regression, and Random Forest Classifier. This method worked by analyzing the dataset that consists of a patients' medical history such as chest pain, sugar level, blood pressure, etc. The heart disease identification system helps in identifying the risk of heart disease by analyzing on the clinical report of a patient. The average accuracy of this model is 87.5%. Among the three algorithms used, the accuracy of KNN is highest that is 88.7%. Using more data helps in developing a system that predicts disease with a higher accuracy. In [2], a dataset of heart disease patients was tested using k-modes clustering, which can increase the accuracy of classification. To obtain the result, several models are used, including XGBoost, Decision Tree classifier, Random Forest, and Multilayer Perceptron. GridSearchVC had been utilized to hyper tune the applied model's parameter to maximize the outcome. The Multilayered perceptron outperformed other models in the context of accuracy with the maximum accuracy with 87.28%. Some noted limitations were- generalizability of the model for different populations or patient groups, there was no evaluation of the model's performance on a held-out test dataset, interpretability of results and the clusters that the algorithm created weren't explained.

In the research article [3], a statistical model is introduced for heart disease prediction, which can be used to aid medical practitioners to predict disease. The proposed model is developed using the algorithm of Decision Trees and the obtained an accuracy of 97 percent. Therefore it is shown that how disease prediction, diagnosis and treatment can be done by using the ML techniques. In research work [4], where performance analysis and comparison were accomplished, several supervised ML techniques are employed to identify some classifiers with the most accuracy to predict heart diseases. These supervised algorithms were applied and compared to find out their performance and accuracy. Except Multi-Layer Perceptron (MLP) and KNN, feature importance score for each feature is estimated for all other algorithms. All the features are ranked based on the feature importance score in order to find highly predictive features. KNN, Decision Trees, Random Forest got 100% accuracy along with 100% sensitivity and specificity.

In this paper, investigation is performed how these cheap and simple algorithms might be of enough to be of clinical use and improve services. In the research work [5], have used various models of ML and deep for predicting and diagnosing diseases. The researches proposed three methods and performed some comparative analysis by applying multiple ML algorithms and deep learning algorithms. Finally difference were found in various applications. Normal dataset acquisition was done using classification directly in the first approach. But in the second method, some feature engineering and feature selection was applied on the data but as such outlier detection was not done. However in the third approach the dataset normalization was performed and along with that feature selection was done and also outliers were taken care of, and feature selection is performed. Moreover in the results third approach performed well. The comparison parameters used over here were confusion matrix, precision, specificity, sensitivity, and F1 score.

In the article [6] Kavita et al. prediction of heart related diseases were done by using ML algorithm in an automated medical diagnostic method. They used a hybrid model for the aforementioned purpose. Now a hybrid model is an approach where the probability is derived from a single ML algorithm. The proposed model is used as an input to another ML model and based on both machine learning and implementation concerns, this hybrid approach gives the researchers better and optimized outputs. The ML algorithms used in the research were namely Naïve Bayes algorithm, Random Forest and SVM, and finally the best result with an accuracy of 90.16 percent was achieved by Random Forest. A research work done in [7], presents a technique is introduced called Hybrid Random Forest with Linear Model (HRFLM). The researchers target to enhance the performance accuracy of cardiac disease prediction and diagnosis. Here various studies were conducted in order to find out the factors that put restrictions in selecting features or algorithmic use. However, the HRFLM method considers all the features without applying any restrictions of feature selection in the model. The classification is done using dataset with a radial basis function network (RBFN).

The researchers in [8] have used various ML techniques such as Neural Network, Support Vector Machines, K-Nearest Neighbors to enhance the accuracy of prediction and identify key factors contributing to heart disease. Their result indicated that the Neural Network algorithm achieved the highest accuracy of 93%. Additionally, the stability of Neural Network was tested across different dataset sizes for the best result. They also conducted a feature selection study using the correlation matrix to identify dependencies among attributes for refining the input features and enhancing model performance.

In [15] Rehman et al. have presented a comprehensive ML framework for heart disease prediction and along with that they have also employed Synthetic Minority Oversampling Technique (SMOTE) to mitigate the class imbalance in dataset. Moreover a hybrid model is also proposed using Artificial Neural Network (ANN) and Particle Swarm Optimization for training biased data. Ratawal et al. in [16] have demonstrated a research work, where they have presented a ML model for heart disease prediction and also selecting significant features for accuracy enhancement. Finally they have achieved best results using the Random Forest technique. The researchers in [17] have proposed a novel ML model for detecting and predicting cardiac diseases. Their research have also addressed overfitting issues during the training and the testing of the data. In the research article [24] the researchers have demonstrated the use of Stratified Random Sampling approach to predict cardiac diseases in different people depending upon the gender of them. Another proposed strategy mentioned in [25] also exhibits the use of ML approach for predicting heart related illness.

III. RESEARCH WORKFLOW

Firstly, the heart disease dataset is collected from open source repository. Before using dataset for model training they must be pre-processed such as missing value should be filled, unwanted data should be removed. After cleaning, relevant data is selected using feature engineering technique. After feature selection, algorithm is chosen based on whether the dataset has both input and output label or just input labels, features are categorical or numerical, and many more. Then Performance is evaluated using classification metrics. Finally, result are generated. While training model, after balancing dataset and selecting features for training model, dataset splits into training set and testing set. For training 80% and for testing 20% of data is used. The work flow of the ML techniques is shown in the Figure 1.

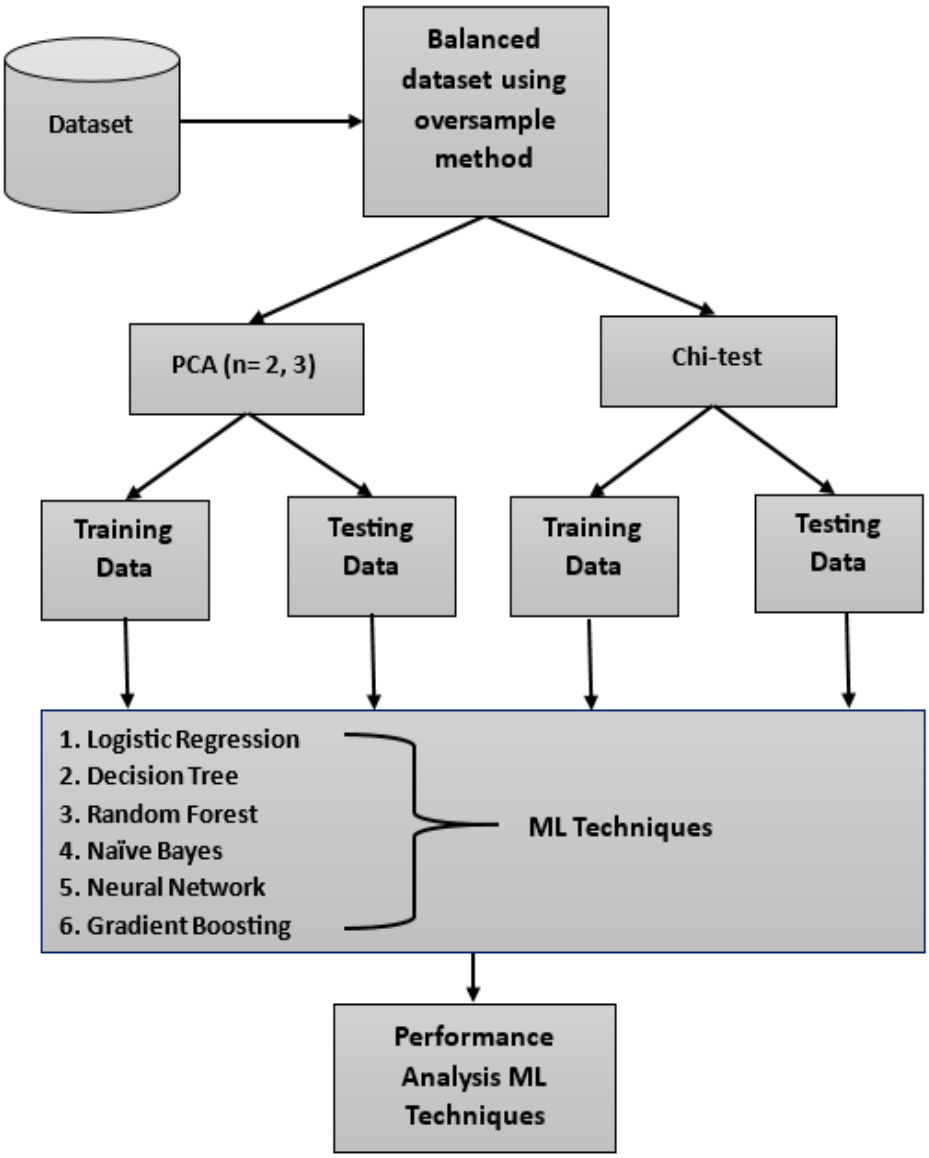

Figure 1: Workflow of ML methods

After splitting we train the model by fitting training subset of input set and target variable which contains 80% of the input set and output set respectively. The workflow of the prediction model is shown in figure 2.

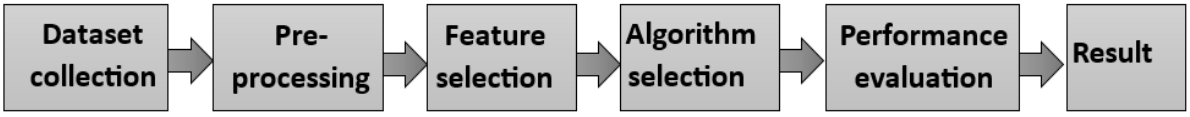

Figure 2: Workflow of model training process

IV. MODEL DEVELOPMENT METHOD

In this section the entire process of model development is mentioned. The step of model development includes data acquisition, data description, data pre-processing, feature selection and extraction, algorithm selection and finally model evaluation.

4.1 Data Collection and Data Set Description

The first step when training any ML model is data collection. We have collected data from (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset). The dataset has total of 1026 entries. There is total 14 features which is mentioned in the table 1. Amongst the used features five features hold continuous values, some five features are binary in nature with values 0 and 1 form, where 0 indicates no/false and 1 indicated yes/true, and rest of three features have range of values starting from 0 to 2 for slope, starting from 0 to 3 for CA, and starting from 1 to 3 for THAL (indicating 1 for normal, 2 for fixed, and 3 for reversible defect). There is a 'target' feature also which has 0 and 1, where 0 is false or no means that person do not have heart disease and 1 means yes/true means the person is struggling with heart disease.

Table 1: Features in the data set

Feature	Remarks
Age	Mentioned in years
Sex	Gender of patient
CP	chest pain type
TRESTBPS	resting blood pressure
CHOL	serum cholesterol
FBS	fasting blood pressure
RESTECG	resting electrocardiographic result
THALACH	maximum heart rate achieved
EXANG	Exercise induced angina
OLDPEAK	ST depression induced by exercise relative to rest
SLOPE	The slope of the peak exercise ST segment
CA	Number of major vessels coloured by fluoroscopy
THAL	Thalassemia

4.2 Data Pre-Processing

Data can be available in many formats. All the raw data that we get from real-world often contains missing data, noise data, and unreliable data which may lead to problem while analyzing the data. There might be chances that there are mistakes or some problems in the dataset [9]. We have to convert those data into usable format before using data to train our model. Data is transformed in such a way that a machine can understand the data and can analyze the features of data.

At our study, we have pre-processed the data that we have collected. There were no missing data in dataset. All the values are in numerical form. Our dataset was unbalanced. Unbalanced data can lead to several issues in training and evaluating the model. There were 526 cases for class 1 and there were only 499 cases for class 0 in target which was making dataset unbalanced. Firstly we have balanced the dataset. We have used "Random Oversampling" method to balance the dataset. The random sampling method separated the dataset into two classes based on the target variable. Then it determines the number of samples required to balance the classes by calculating the maximum of two class sizes. If a class having lesser number of samples, it oversamples the lesser class by randomly selecting samples equals the required samples until the number of samples equals the required number. By applying Random Oversampling we managed to balance the class as shown in the figure 3 below.

4.3 Feature Selection and Extraction

Many features in the dataset are unwanted which make the size of dataset bigger and reduces the learning speed. So, only those features should be selected for the training which are relevant to the topic of our model. By using the domain knowledge or getting help from expertise may also help eliminating the features that are not needed. There are also some procedures which helps in feature selection like supervised feature and unsupervised feature selection [10]. Feature selection can reduce the complexity of computation, cost incurred for storage and enhance the performance of prediction as per [11]. For feature selection method we have use Principal Component Analyses (PCA) Method and for feature extraction Chi-test (X2- test) method. PCA is a feature extraction method which is used to transform a dataset with many features into a smaller set of features. In the process of extracting features, it doesn't tell which variables are selected and only selects most relevant and important features. Chi-test (X2-test) is a feature selection method which selects the most relevant features for model training. It evaluated by calculating how much each feature contributes to the target feature by measuring the dependency between them. After balancing the dataset first, we have applied PCA method to calculate the classification metrics. In PCA method we use weight components, weights are eigenvectors of X (mean centered data). We have used the values N=2 and N=3 in PCA components and compared the values of classification metrics and computed both train data and test data. In the table 2 a comparison is demonstrated with component value N=2 and N=3, both decision tree and random forest gives the best result 100%, which means that the model correctly predicts the class labels for all instances in the dataset and the model didn't make any mistakes. 100% accuracy led to overfitting means model might be memorizing the training data rather than learning general patterns which means the model could perform poorly on unseen data.

Therefore Decision tree or random forest might not be the correct choice for model training. After decision tree and random forest gradient boosting gives the best result with accuracy 95%, precision 95%, recall 96%, and F1-score 96% when N is 2. When N is 3, accuracy is 97%, precision 97%, recall 97%, and 97%. So it is better when we choose Gradient boosting with N= 3. However in the subsequent table 2 metrics values are demonstrated after application of PCA on training data, whereas table 3 represents the same metrics' values in the testing data.

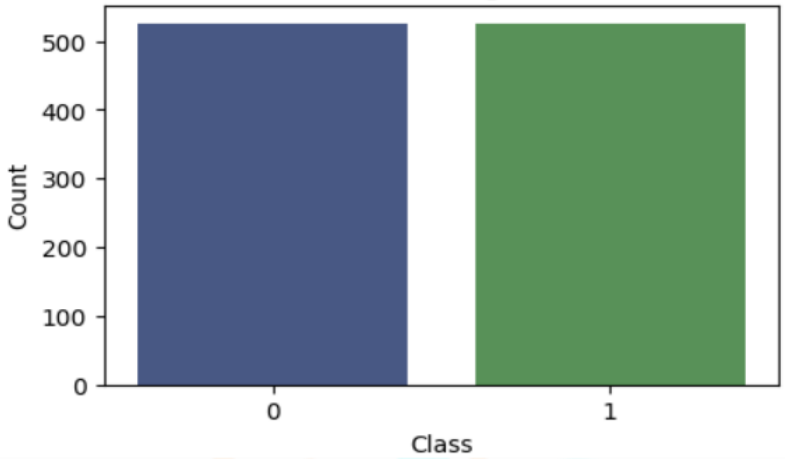

Figure 3. Distribution of Target variable

Table 2: Calculating metrics by applying PCA on Train Data

ML Techniques	N=2				N=3			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.81	0.78	0.83	0.81	0.85	0.88	0.83	0.85
Decision tree	0.72	0.74	0.72	0.73	0.79	0.80	0.79	0.79
Random Forest	0.77	0.79	0.77	0.78	0.83	0.83	0.85	0.84
Naïve Bayes	0.81	0.79	0.84	0.81	0.83	0.83	0.86	0.84
Neural network	0.88	0.85	0.91	0.88	0.88	0.90	0.85	0.87
Gradient boosting	0.74	0.76	0.74	0.75	0.80	0.80	0.83	0.81

Chi-test is feature selection method which selects the most relevant information or features from dataset. After applying chi-test, selected features are THALACH, OLDPEAK, CP, CA, EXANG, CHOL, and AGE. By comparing all the techniques in table 3. Decision tree and random forest gives the best result with accuracy of 100%, which means overfitting. After these gradient boosting gives the best result with accuracy 97%, precision 97%, recall 97%, and F1-score 97% in both train and test data. Again by comparing table 2 and 4, in table 2 when N is 3 gradient boosting gives the metrics value of 97% and in table 4 also gradient boosting gives the accuracy of 97%. But when comparing test data chi-test method gives the best result for gradient boosting in table 3.

4.4 Selection of ML Algorithm

Before selecting any algorithm, there are parameter that should be followed for determine algorithms. Different datasets come with different sizes, some of datasets have both input and output label, and some only have input label. Labels are also categorised in categorical value and numerical value. So based on these factors we have chosen our algorithm. Our dataset has both input and output label consisting total of 13 rows from which we have some selected features using Feature selection(Chi-test) and feature extraction(PCA). Output label has 0 and 1 that is yes/no, which means we have to apply classification of supervised algorithm into our dataset to train our model. Classification algorithm consists of many algorithm mentioned in [19] and [12]. This is step is to figure out which algorithm will perform well on our dataset by finding the accuracy of different algorithms.

4.5 Evaluation of Model

Machine learning metrics are used to evaluate the performance of a model. The choice of metrics depends on the specific type of ML tasks (e.g., classification, and regression) and the goal of the application. Choosing the right metrics is crucial for understanding how well a model is performing and for guiding the model optimization process. In case of our dataset, since we are using classification algorithm so we have to use classification metrics such as accuracy, precision, recall, and f1-score to find how well an algorithm will work on dataset. The predictions can be quantified by using some metrics mentioned below. The formulae of these metrics are mentioned hereafter and they are dependent on parameters namely true positive predictions (TP), false positive predictions (FP), true negative predictions (TN) and false negative predictions (FN). Diego et al. in [14] have proposed a General Performance Score (GPS) a methodological technique to measure the performance of binary or multiclass prediction model. They have also demonstrated the usage of the same.

a. Accuracy- The formula for computing accuracy (given by A here) of the classification of the proposed model is given as mentioned below.

$$A = \frac{TP + TN}{TP + TN + FP + FN} (Eq. 4.1)$$

b. Precision- The precision (given by P here) metric quantifies the ratio of the TP prediction to the total number of TP and FP and the formula for computation is given by following.

$$P = \frac{TP}{TP + FP} (Eq. 4.2)$$

c. Recall – This metric recall (given by R here) is also called as sensitivity and very close to precision and it targets to compute the proportion of TP and summation of TP and FN.

$$R = \frac{TP}{TP + FN} (Eq. 4.3)$$

d. F1-score- This metric also measures the accuracy of a predictive model and it is computed as the harmonic mean of precision and recall and the formula is given as below.

$$F1 - score = \frac{2(PxR)}{(P + R)} (Eq. 4.4)$$

Table 3: Calculating metrics by applying PCA on Test Data

	N=2				N=3			
ML Techniques	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.80	0.79	0.84	0.81	0.80	0.78	0.81	0.80
Decision tree	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Naïve Bayes	0.80	0.78	0.82	0.80	0.81	0.78	0.86	0.82
Neural network	0.88	0.86	0.91	0.88	0.82	0.84	0.80	0.82
Gradient boosting	0.95	0.95	0.96	0.96	0.97	0.97	0.97	0.97

V. RESULTS AND ANALYSIS

Total of 6 different algorithm have been used in our study as mentioned in Table 2, 3 and 5. Firstly, we have imported all the libraries of algorithms and metrics from SKLEARN. We had split the dataset into two parts training and test sets. We have used 80% of data for training and 20% of data for testing. After that we trained the model using training data, made prediction on training data and test data and then metrics were calculated using training and test sets.

5.1 Descriptive Statistics of Data Set

Exploratory Data Analysis is employs use of data visualization to evaluate and investigate data sets and describe their key properties. Below is a short description of our data. For our analyses, we have filtered only those data who have 'target'= 1 and calculated their key property. In the table V below, count is the number of non-missing entries for each column. Count for each column is 1052, but since we have selected only data with target=1 so there are 526 entries in all the selected attributes and there is no missing value. Mean is the average value of entries in each column. Mean is calculated by adding all the values in a dataset and then dividing the sum to number of values. As demonstrated in Table 4, that contains the insight of 526 data with some metrics as average age value is 52.4, indicates that person having age near 52 are having higher risks of heart disease who have cholesterol above 240. STD is the standard deviation. It measures the amount of variation in a set of values. It shows the spread of the values around the mean. MIN is the minimum values in each column. 25%, 50%, and 75% gives the 25th, 50th, and 75th percentile values in each column, respectively. Max gives the maximum values in each column. Analyses of continuous value is shown in table 4 below.

5.2 Analysing the Performance of Various Algorithm

The purpose of analyzing algorithms is to see which algorithm can give best result while training the model. We applied the algorithms to our dataset to calculate metrics. Within our dataset there is 1 feature along with the output category 'target' which tells whether a person have disease or not.

We have calculated the classification metrics such as accuracy, precision, recall, and F1-score of various different algorithms. Accuracy measures how well a model correctly predicts outcome. However, accuracy can be misleading in certain situation like having imbalanced dataset. So for these certain situations we also calculate the values for precision, recall, and F1-score. The metrics scores of algorithms are given in table 5. By comparing all the algorithms from table 5, Decision Tree and Random Forest gives highest accuracy (1.0) which is 100%. A 100% accuracy in a model means that the model correctly predicts the class labels for all instances in the dataset and the model didn't make any mistakes.

Table 4: Description of continuous data

	AGE	SEX	CP	TREST BPS	CHOL	FBS	REST ECG	THAL ACH	EXANG	OLD PEAK	SLOPE	CA	THAL
MEAN	52.41	0.57	1.38	129.25	240.98	0.13	0.6	158.59	0.13	0.57	1.59	0.37	2.12
STD	9.63	0.5	0.95	16.11	53.01	0.34	0.5	19.10	0.34	0.77	0.59	0.87	0.47
MIN	29	0	0	94	126	0	0	96	0	0	0	0	0
25%	44	0	1	120	208	0	0	149	0	0	1	0	2
50%	52	1	2	130	234	0	1	161.5	0	0.2	2	0	2
75%	59	1	2	140	265.75	0	1	172	0	1	2	0	2
MAX	76	1	3.0	180	564	1	2	202	1	4.2	2	4	3

Table 5: Evaluation of algorithms with all features

ML Techniques	Train Data				Test data			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.81	0.78	0.83	0.81	0.80	0.79	0.84	0.81
Decision tree	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0	0.80	1.0	1.0	1.0
Naïve Bayes	0.81	0.79	0.84	0.81	0.80	0.78	0.82	0.80
Neural network	0.88	0.85	0.91	0.88	0.88	0.86	0.91	0.88
Gradient boosting	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97

But 100% accuracy means the task might be very simple, or the dataset might be perfectly structured which makes it easier for the model to achieve perfect accuracy. A 100% accuracy might also lead to overfitting means model might be memorizing the training data rather than learning general patterns which means the model could perform poorly on unseen data. After Decision Tree and Random Forest, Gradient boosting gives the highest accuracy. So Gradient Boosting is the best choice based the metrics having good accuracy results. High precision value indicates gradient boosting is effective at minimizing FP cases. Recall with the value 97% meaning it correctly identifies actual positive cases which is essential for detecting as many true cases of heart disease. Finally, the F1-score 97% that balances precision and recall, meaning the model is effectively and correctly diagnosing heart disease while minimizing both FP and FN cases.

VI. CONCLUSION AND FUTURE WORK

In this section we conclude this article and we state that in this research work by using ML algorithms, we could propose predictive model that focus to identify cardiac disorders or what we know more commonly as heart diseases. Some of the well-known algorithms were employed to train our dataset and test its performance. However, after executing the training and testing process it was found that highest performance was demonstrated by Gradient Boosting Classifier Algorithm. Based on the user's input, the computer is taught to recognize whether the person is normal or has heart disease. Our goal is to provide community with an efficient and exact method to machine learning approach that might employed towards illness detection applications. Moreover the future scope of this work is to compare the current model with the other existing models and further propose a more generic novel ML model for prediction of many other common diseases by using imaging based medical data sets.

REFERENCES

[1] H. Jindal "Heart disease prediction using machine learning algorithms" IOP Conf. Series: Materials Science and Engineering, ICCRDA 2020, pages 1-10, doi:10.1088/1757-899X/1022/1/012072

[2] C. M. Bhatt, P. Patel, T. Ghetia, P. L. Mazzeo "Effective Heart Disease Prediction Using Machine Learning Techniques" Algorithms, MDPI, 2023, 16, 88, pages 1-15, doi:10.3390/a16020088

[3] A. Mall, A. Singh, A. Bansal, H. K. Gupta, "Supervised Machine Learning in Cardiology: A Predictive Model for Heart Disease", Proceedings of 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp.1-7, 2024 DOI: 10.1109/ICRITO61523.2024.10522122

[4] M. M. Ali, "Heart Disease Prediction using Supervised Machine Learning Algorithms: performance analysis and comparison", Comput. Biol. Med.(2021) DOI: 10.1016/j.compbiomed.2021.104672

[5] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Computational Intelligence and Neuroscience, vol. 2021, Article ID 8387680, 11 pages, 2021. DOI: 10.1155/2021/8387680

[6] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1329-1333, DOI: 10.1109/ICICT50816.2021.9358597

[7] Mohan Senthilkumar et al. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." IEEE Access 7 (2019): 81542-81554. DOI: 10.1109/ACCESS.2019.2923707

[8] D. E. Sahil, A. Tari, T. Kechadi, "Using Machine Learning for Disease Prediction" Feb 2021, DOI: 10.1007/978-3-030-69418-0_7

[9] K. Maharana, S. Mondal, B. Nemade, "A review: Data pre-processing and data augmentation technique", KEAI preprints, DOI: 10.1016/j.gltp.2022.04.020, p.n(92)

[10] M. Buyukkececi, M. C. Okur, "A Comprehensive Review of Feature selection Stability in Machine Learning", Journal Science GU Vol. 36 Issue 4 (2023) pages 1506-1520, DOI:10.35378/gujs.993763

[11] B. Venkatesh, and J. Anuradha "A Review of Feature Selection and Its Methods", DOI: 10.2478/cait-2019-0001, pages 4-13

[12] V. Sheth, U, Tripathi, A. Sharma "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose" DOI: 10.1016/j.procs.2022.12.044, p (424-426)

[13] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, J. M. Moguerza1 "General Performance Score for classification problems" Applied Intelligence, Springer (2021) Vol. 52 pages 12049–12063, https://doi.org/10.1007/s10489-021-03041-7

[14] M. M. Taye "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions" Computers, MDPI 12, 91, 2023, pages 1-26, https://doi.org/10.3390/computers12050091

[15] M. U. Rehman, S. Naseem, A. U. R. Butt, T. Mahmood, A. R. Khan, I. Khan, J. Khan & Y. Jung "Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment" Scientific Reports (2025) 15:13361, Nature Portfolio, https://doi.org/10.1038/s41598-025-96437-1 3

[16] Y. Ratawal, Sabya, Saurav, Swati "Heart disease prediction using Hybrid machine learning Model" International Journal of Research in Engineering and Science (IJRES) Vol. 10 Issue 1 2022, ISSN (Online): 2320-9364, ISSN (Print): 2320-9356, Pages 30-36

[17] A. Singh, H. Mahapatra, A. K. Biswal, M. Mahapatra, D. Singh and M. Samantaray "Heart Disease Detection Using Machine Learning Models" published in the Proceedings of International Conference on Machine Learning and Data Engineering (ICMLDE 2023), Procedia Computer Science 235 (2024), pages 937–947

[18] B Padmaja , C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, E. K. R. Patro "Early and Accurate Prediction of Heart Disease Using Machine Learning Model" Turkish Journal of Computer and Mathematics Education Vol.12 No.6 (2021), pages 4516-4528

[19] F.Y. Osisanwo, J.E.T Akinsola, O. Awodele, J. O.Hinmikaiye, O. Olakanmi, J. Akinjobi "Supervised Machine Learning Algorithms: Classification and Comparison" International Journal of Computer Trends and Technology (IJCTT) Vol. 48 No. 3 June 2017, ISSN: 2231-2803, pages 128-138

[20] S. Anjum, S. Oraon, Saloni, S. Kumari, S. Mitra "The Life Cycle of a Machine Learning Model" Journal of Emerging Technologies and Innovative Research (JETIR), Vol. 11 Issue 7, 2024, pages 232-238

[21] S. Oraon , S. Anjum, S. Kumari, S. Mitra "Machine Learning Algorithms for solving Generic Real Life Problems - A Comparative Study" International Journal of Engineering and Techniques - Volume 10 Issue 4, 2024, pages 138-148, ISSN: 2395-1303

[22] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure" Nature Reviews Cardiology, Vol. 8, no. 1, 2011, pages 30–41.

[23] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data" Neuroimage, Vol. 28, no. 4, 2005, pages 980–995.

[24] V. L. Sarraju, J. Pal, S. Kamilya "SRS: Gender-Based Heart Disease Prediction Using Stratified Random Sampling Approach", Proceedings of International Conference on Machine Intelligence with applications (ICMIA 2023), Ranchi, India,07-08 December 2023

[25] J. Pal, A. Prakash, S. Kumar "A Predictive Paradigm: Proposed Ensemble Frame Work for Cardiovascular Disease" International Journal Of Engineering Research & Technology (IJERT) 2025 Vol. 14, Issue 1