Fuzzy Systems and Soft Computing
ISSN : 1819-4362

COMPARATIVE ANALYSIS FOR HEART DISEASE DATASET USING SUPERVISED MACHINE LEARNING ALGORITHMS

Dr. K. Radhakrishnan Associate Professor, PG & Research Department of Computer Science Dr. Ambedkar Government Arts College, Vyasarpadi, Chennai – 600039

Abstract:
The development of machine learning and data mining-based methods for the prediction and diagnosis of cardiac disease presents a significant clinical challenge. A large percentage of cases are misdiagnosed in most nations because to a lack of cardiovascular competence, which might be addressed by creating accurateand effective early-stage cardiac illness prediction through clinical decision-making with digitized medical data supported by analytics. Finding the machine learning classifiers with the best accuracy for these kinds of diagnostic applications was the goal of this work. The effectiveness and accuracy of many supervised machine-learning algorithms were used to predict cardiac disease and then compared. With the exception of MLP and KNN, all applied methods evaluated the feature significance scores for each feature.To identify the factors that offered the highest likelihood of heart disease, all of the variables were graded according to significance. This study discovered that the RF approach achieved 100% accuracy coupled with 100% sensitivity and specificity utilizing a heart disease dataset obtained from Kaggle three-classification based on k-nearest neighbor (KNN), decision tree (DT), and random forests (RF) algorithms. Thus, we discovered that heart disease predictions may be made with extremely high accuracy and good potential value using a relatively basic supervised machine learning technique.

Keywords*:*
Machine learning; Heart disease; Random forest; k-Nearest Neighbour; Decision tree

I. INTRODUCTON

Heart disease continues to be a major global source of morbidity and death, presenting significant challenges to international health systems. The ability to detect cardiac disease early and accurately is crucial for prompt therapies that can save lives and lower medical expenses. Developments in supervised machine learning algorithms provide a viable way forward for this important role in healthcare. These algorithms are able to accurately forecast the possibility of heart disease in new patients by analyzing vast datasets to find patterns and risk factors connected with the illness. The use of supervised machine learning algorithms for cardiac disease prediction is examined in this introduction, with a focus on how these approaches have the potential to revolutionize preventative healthcare and enhance patient outcomes.

Heart disease refers to a wide range of heart-related disorders, such as congenital heart abnormalities, arrhythmias, and coronary artery disease. High cholesterol, smoking, diabetes, obesity, and hypertension are important risk factors. Because lifestyle modifications, medication, and other interventions can dramatically lower the risk of serious cardiovascular events, early identification is essential for effective care. The prediction of heart disease is a prime option for machine learning techniques because to the intricate and interconnected nature of these risk variables.

Training algorithms on labeled datasets—where the input characteristics and related outputs such as the presence or absence of heart disease—requires supervised machine learning. For the purpose of predicting cardiac disease, supervised learning techniques such as logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks are frequently employed. Because it can estimate the likelihood of a binary outcome, logistic regression is a useful tool for predicting the existence of cardiac disease. Decision trees give an intuitive grasp of the decision-making process by using a structure like a tree to generate judgments depending on input attributes. An ensemble technique called random forests combines several decision trees to decrease over fitting and increase prediction accuracy. SVM models provide strong performance in high-dimensional spaces by identifying the ideal hyper plane that divides various classes. Deep learning models in

particular, which use neural networks, are capable of capturing intricate correlations between variables and frequently attain high accuracy in predicting tasks.

Effectiveness of machine learning models is heavily reliant on the caliber and pertinence of the input data. Relevant aspects in the prediction of heart disease might be clinical measures (blood pressure, cholesterol levels), lifestyle variables (smoking status, physical activity), and demographic data (age, gender). Choosing, altering, and producing new features from unprocessed data is known as feature engineering, and it is essential to improving model performance. Features that are well-designed can greatly enhance the model's capacity to recognize trends and generate precise forecasts.

Supervised machine learning has the ability to forecast cardiac disease, but there are still a number of obstacles to overcome. These include incorporating predictive models into healthcare operations, protecting data privacy and security, and handling unbalanced datasets. Biased models that perform badly on minority classes might result from imbalanced datasets, where the number of heart disease patients is significantly lower than the number of non-cases. Because medical data is sensitive, it is imperative to ensure data privacy and security. When incorporating predictive models into clinical workflows, it's important to take into account how medical professionals will utilize these tools and how they might enhance current diagnostic procedures.

Cardiovascular diseases (CVD) rank as the leading cause of mortality globally, accounting for around 17.9 million deaths annually, according to estimates from the World Health Organization [1]. One essential strategy for lowering this toll is the early identification of CVD. Data mining is one of the numerous methods for enhancing illness diagnosis and detection. These related approaches are a potential approach for CVD classification because they enable the extraction of hidden information and the identification of correlations among parameters within the dataset [2-4].One of the biggest problems facing health organizations is providing patients with clinical treatments of the highest caliber at a reasonable cost. In order to provide quality care, patients must be correctly diagnosed, and an effective treatment plan must be identified, all the while avoiding incorrect diagnoses [5].

II. LITERATURE REVIEW

Researchers have employed several data mining techniques, including association rules, classification, and clustering, to construct a model aimed at predicting heart disease. Using the data mining approach, Shiva Kazempour Dehkordi1 & Hedieh Sadeghi created a prediction model that utilized the prescription [7]. They suggested the Skating algorithm to improve the system's accuracy. Skating may be compared to boosting and bagging as an ensemble strategy. Four distinct methods for classification, including DT, Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Skating in a different label, were compared. They demonstrated that staking is the most accurate classifier available. The accuracy of this categorization method was 73.17%. Comparatively speaking to other classification algorithms and approaches, this one performs quite poorly.Jan et al. (2018), for instance, used two benchmark datasets—Cleveland and Hungarian—that were gathered from a UCI repository to implement an ensemble data mining approach. The ensemble of five distinct classification algorithms—including RF, neural networks, NB, regression analysis, and support vector machines (SVM)—was used in this study [8]. In that investigation, they found that RF produced a very high accuracy of 98.136%, whereas regression approaches produced the lowest performance.

To increase classification performance, Jyoti Soni et al. used DT in conjunction with a genetic algorithm in 2011; this was contrasted with two other algorithms, including NB and classification using cluster approaches [9]. 99.2% accuracy was found for the suggested system. In 2017, Hend Mansoor et al. examined how well the LR and RF classification algorithms performed in assessing the risk exposure of patients with CVD [10]. They demonstrated that the LR Model outperformed the RF classification algorithm in terms of performance. The accuracy of the LR Model was 89%, and the RF Model was 88% accurate. The performance of regression trees and traditional classification trees was examined by Austin et al. in 2013 [6].Excellent results were obtained in assessing the possible occurrence of HD using conventional LR.

In 2018, Le et al. used three different classification techniques for the dataset that was gathered from the UCI Machine Learning Repository, which had 58 specified properties [11]. They demonstrated that, with an accuracy of 89.93, a support vector machine (SVM) using a linear kernel performed better. A hybrid technique with 12 features was presented by Tarawneh and Embarak [12],

who also compared its performance to that of KNN, J48, GA, DT, artificial neural network (ANN), SVM, and NB. 89.2% accuracy was produced by the suggested hybrid technique, the best result when compared to other applicable algorithms. A classifier known as a cascaded neural network (CNN) was suggested by Chitra and Seenivasagam in 2013 as a way to improve the accuracy of heart disease prediction [13].A CNN has a cascade design, meaning that when neurons are added to the hidden network, the network is augmented one at a time with cached neurons that remain unchanged. The suggested approach's outcome was compared to SVM, which yielded 82% accuracy and CNN 85% accuracy as well as 0.87 and 0.775 specificity, respectively. After taking these factors into account, they recommended CNN since the CNN classifier predicts heart disease more accurately than SVM and the model it uses has greater accuracy.

Using the Cleveland dataset, Latha and Jeeva developed an ensemble classification strategy and combined Majority vote with MP, RF, BN, and NB utilizing the feature selection method to increase the classifier's accuracy [14]. The six sets of attributes were used to evaluate the performance. To determine which ensemble model performed the best, they constructed many ensemble models and contrasted the results. They presented an ensemble approach to predict heart disease after discovering that the Majority vote with MP, RF, BN, and NB using attribute selection technique gave the greatest performance with 85.48% accuracy. But now, approaches that provide greater accuracy than their suggested model can be found. A model for predicting cardiac disease using hybrid machine learning approaches was presented by Mohan et al. in 2019 [15].Rattle, a Graphical User Interface tool for Data Mining with R, was used in this work to categorize HD using the dataset that was gathered from the Cleveland UCI repository. This led to a higher performance level, wherein the prediction model for HD with the hybrid RF with a linear model (HRFLM) had an accuracy of 88.7%. They demonstrated that their implemented model produced superior results than previous classification algorithms by comparing it with other proposed models and methods.

A few studies that used data mining and machine learning techniques to forecast the course of cardiac disease are included in this part. The explanation above makes it rather evident that the precision attained in individual research projects is currently insufficient. It is possible to get superior performance using certain algorithms over others. Through 10-fold cross-validation, the research study has effectively identified three algorithms that provide 100% accuracy. Therefore, the goal of the project is to identify classifiers that can accurately predict cardiac disease in a way that is relevant for clinical settings.

III. METHODOLOGY

This section discusses the problem definition for this research work. Organizing the many forms of unorganized information in a customer review is the main challenge for data mining activities. It is necessary to comprehend the patterns and important phrases in the customer review to analysis the disease. The dataset is preprocessed to remove redundant data, missing data, and unnecessary features. The cleaned heart disease reviews dataset is then used in the prediction process to determine whether the individual affected would be identified using Random Forest, Decision Tree and k-nearest neighbor. The.csv file format for the heart disease dataset can be obtained from the Kaggle repository.

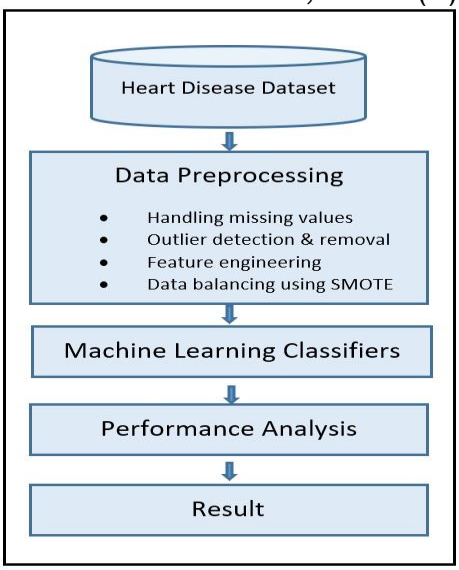

Figure 1: Research Methodology

According to the disease among the various input data forms, three distinct machine learning algorithms Random Forest, Decision Tree and KNN are used to do the research. Based on how the data points are separated from one another, each feature has been categorized and arranged [15]. A range of colors are used to display each categorized data point, and the execution time is expressed in seconds. Figure 1 shows the overall methodology of this research work.

A. Description of Dataset

To create the anticipated model for this investigation, a dataset on heart disease was analyzed. The collection of the dataset came from Kaggle [16]. This dataset has 14 characteristics. All feature information are included in figure 1.1025 patient records total from 713 males and 312 females of various ages make up the collection. Of them, 499 (48.68%) have normal hearts and 526 (51.32%) have heart disease. Of the patients suffering from heart disease, 226 (52.97%) are female and 300 (57.03%) are male.

SN	Attribute name	Description
1	age	Age in years
2	sex	Male = 1; Female = 0
3	cp	chest pain type (4 values)
4	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholesterol in mg/dl
6	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	resting electrocardiographic results
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0–3) colored by fluoroscopy
13	thal	1 = normal; 2 = fixed defect; 3 = reversible defect
14	Target (Class)	0 = no disease and 1 = disease

Figure 2: Sample of dataset

Figure 2 displays a sample of the dataset for reviews of musical instruments before preprocessing.

B. Data Preprocessing

This work uses Python version 3.8.5 in exploratory data analysis (EDA) and visualization, and Weka version 3.8.3 for data mining. Any machine learning or data mining strategy must include data preprocessing as the quality and organization of the dataset determines how well a machine learning methodology performs.After handling missing data with a ReplaceMissingValues filter, an Interquartile Range (IQR) filter was used to identify outliers and extreme values during the pre-

processing stage. A technique for calculating a dataset's variability around the median is the IQR.An outlier is a data point which falls outside of the data's predicted range and is presumed to be the result of recording mistakes or other unrelated occurrences for analytical reasons [17]. To obtain a better analytical or statistical outcome, it is crucial to exclude such outliers from data mining or machine learning (ML) techniques [18]. The data is divided into three quartiles, Q_3, Q_2, and Q_1, in order to discover outliers. The data boundaries in this case are Q_1 and Q_3. IQR was determined using the formula $IQR = Q_3 – Q_1$. Next, the following equations [19 - 22] were used to derive the upper boundary Bu and lower boundary B_l:

$$Bl = Q1 – 1.5*IQR \quad (1)$$

$$Bu = Q3 + 1.5*IQR \quad (2)$$

An outlier is defined as a result that is larger than Bu and less than Bl. To balance the unbalanced dataset, the synthetic minority oversampling method (SMOTE) was also utilized. In order to ensure that the dataset is free of outliers, some exploratory data analyses (EDA) were carried out, such as box plots. The data was also represented as an IQR and heatmap to find correlations between the features, along with a KDE plot for both diseased as well as non-diseased individuals based on age distribution [23, 24].

C. Performance Metrices

The dataset was subjected to six (06) classification algorithms in order to identify the top performing method using 10-fold cross-validation, which compares accuracy and other statistical factors. Multilayer perceptrons (MP), K-nearest neighbors (KNN), random forests (RF), decision trees (DT), logistic regression (LR), and AdaboostM1 (ABM1) were the techniques used. The algorithms were compared using measures for evaluating their performance. This part offers a succinct description of various performance assessments.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F-Measure} = \frac{2*precision*recall}{precision + recall} \quad (7)$$

The dataset's performance is shown by performance metrics. The presentation of the suggested system was assessed using the following criteria: Sensitivity, Specificity, Accuracy, F-measure, Precision, and Recall. Conventional count values, such as True Positive (Tp), True Negative (Tn), False Positive (Fp), and False Negative (Fn), are utilized here.These metrics are carefully utilized to

evaluate the algorithms' performance when compared to the analysis's evaluation of the data that was selected set.

D. Supervised Machine Learning Algorithms

In this research, many supervised machine learning techniques were used. The labeled training dataset is used mostly to practice the basic method in supervised machine learning techniques. In order to group the research dataset into comparable categories, this qualified model is next put into an unlabeled testing dataset [25]. The related part provides a brief summary of these suggested supervised machine learning formulae for illness detection.

D.1*K-nearest neighbor (KNN)*

One of the simplest and most traditional classification algorithms [26, 27] or statistical learning methods [28] is KNN. K stands for the number of nearest neighbors utilized, which may be computed by utilizing the upper limit given by the given value [24] or by manually defining it in the object constructor. As a result, related examples are classified similarly [29], and a new instance is categorized by comparing it to every existing instance [30]. The closest neighbor method looks for the k training samples that are next to an unknown sample in the pattern space when it receives an unknown sample.Two different approaches are shown to convert the distance into a weight, and predictions from several neighbors may be computed from the test instance based on their distance [28, 31]. Among the many benefits of the method are its ease of implementation and analytical tractability [28]. Because it only requires one instance, the classifier is incredibly efficient and performs well in illness prediction, particularly in HD prediction. The analysis found that the values of leaf_size 40 and n_neighbors 2 provided the greatest fit for the dataset.

D.2 *Random Forest (RF)*

Based on DT, RF is an ensemble learning technique for data classification [32]. When it is in the training stage, it generates a lot of trees as well as a forest of decision trees [33]. During the testing phase, each tree in the forest predicts the class label for each and every occurrence. Majority voting is utilized to determine the ultimate choice for each test data when each tree predicts a class label [34]. When it comes to the test data, the class label with the most votes is deemed to be the most appropriate one. This cycle is repeated for each piece of data that is collected.For this investigation, the best suited random state value was 123, which provided the best results for the used dataset.

D.3 *Decision Tree (DT)*

Among the most popular and established machine learning algorithms is DT.A decision-making logic known as a decision tree (DT) is designed to assess and correlate data item categorization findings into a tree-like structure [25]. A decentralized graph (DT) often consists of several layers of nodes, with the root or parent node at the top and other levels being child nodes. All internal nodes with at least one child node indicate the assessment of input variables or characteristics. The classification algorithms branch to the appropriate child node based on the evaluation result, and this process of branching and evaluation continues until the leaf node is reached [34]. The decision's results are denoted by the leaf or terminal nodes. DT is widely acknowledged as being simple to comprehend and acquire, and it forms the foundation of several medical diagnostic processes [35]. For the dataset used in this investigation, the classifier with a maximum depth value of 7 yielded the best results. This maximum depth value was defined for the classification procedure.

D.4 *AdaboostM1 (ABM1)*

ABM1 is a popular type of supervised machine learning classifier that is based on ensemble learning. It combines many weak classifiers into one strong classifier using an adaptive improvement technique, which improves classification results [36]. Every observation is given the same weight during the initial phase. The coefficient of the weak classifiers affects the weights of the data, and the estimation error value is used to estimate the coefficient of the applied classifiers.Accordingly, the classifier's coefficient is defined as the value of error that the classifier produces. As a result, the ABM1 algorithm may increase the weight of incorrectly categorized data while decreasing the weight of

properly recognized observations. It will give the improperly categorized observations additional weight in the next repetitions. In order to achieve accurate classification performance, all of the weak classifiers that were produced are finally merged into a stronger classifier utilizing a linear combination approach [37]. In this investigation, the classifier that performed the best was identified as having a value of 200 for n estimators.

D.5 ***Logistic regression (LR)***

LR is an extension of generic regression modeling that, when applied to a dataset, reflects the likelihood of occurrence or nonoccurrence of a certain instance [39]. It is a powerful classifier of supervised machine learning algorithms [38]. Since LR is a probability, it determines the likelihood that a new observation will fall into a particular class. The outcome can be anything from 0 to 1. As a result, in order to apply the LR as binary classification, a threshold is set that determines the division into two classes. For example, a probability value more than 0.5 is classified as "class A," whereas a value less than 0.5 is classified as "class B." The LR model may be extended as a multinomial logistic regression to provide a categorical variable with more than two values [40].This study revealed that, for the applied dataset, the best fit random state value was 1234 and the best fit maximal iteration number was 100.

D.6 ***Multilayer perceptron (MLP)***

Three or more layers make up the well-known neural network-based categorization technique known as Maximum Linguistic Product (MLP): an input layer, an output layer, & one or more hidden layers that sit between the input and output levels [41]. A large number of "neurons" interconnect each layer with the adjacent levels. Because training data may learn and generalize [2] using training data utilizing backpropagation learning methods [42], MLP is a universal multivariate non-linear mappings calculator. Enough input variables, network type specification, relevant data pre-processing as well as partitioning, network infrastructure configuration, success parameter specification, training algorithm (relation weight optimization) specification, and model evaluation are all necessary for building MLP classifiers [43].

E. Importance of Feature

In the realm of machine learning, feature importance & its visualization constitute a significant and popular analytical technique. Because feature rating and risk analysis are so straightforward and easy to understand, they are especially used in fields like biology and the social sciences. Each feature's coefficient value determines the feature's relevance and ranking. MLP and KNN do not produce any feature significance or coefficient scores, despite the fact that the majority of supervised algorithms do. In addition to these two classifiers, the appropriate sections identify and display feature significance or coefficient ratings.

IV. RESULTS AND DISCUSSION

A dataset on heart disease has been prepared for this work. Outliers have been found and eliminated, and many classification methods, such as MLP, KNN, DT, RF, LR, and ABM1, have been used. These classifications all on the dataset, algorithms using 10-fold cross-validation techniques were used. We analyzed the various cross-validation performance factors to find the optimum method for predicting the occurrence of heart disease. Figure 1 shows the entire procedure. The following outcomes are the product of the procedure.

Table 1: Classification results of different algorithms

Algorithms	Accuracy	Sensitivity	Specificity
LR	83.03	95.06	90.63
ABM1	92.05	98.90	96.03
MLP	99.42	98.60	98.67
KNN	100	100	100
DT	100	100	100
RF	100	100	100

The performance outcome metrics of the used classification algorithmsspecificity, accuracy, and sensitivityare displayed in Table 1.All of these parameters yield strong results; KNN, RF, and DT offer the highest levels of accuracy, sensitivity, and specificity. MLP performs better than LR and ABM1, in second place.

Table 2: Precision, Recall, F-measure results of algorithms

Algorithms	Precision	Recall	F-measure
LR	90.62	91.62	92.52
ABM1	96.02	97.75	98.27
MLP	99.36	98.50	97.84
KNN	100	100	100
DT	100	100	100
RF	100	100	100

LR and ABM1 perform worse than MLP when accuracy, recall, and f-measures are taken into account, as shown in Table 2. Simultaneously, 100% performance is shown by KNN, RF, and DT.

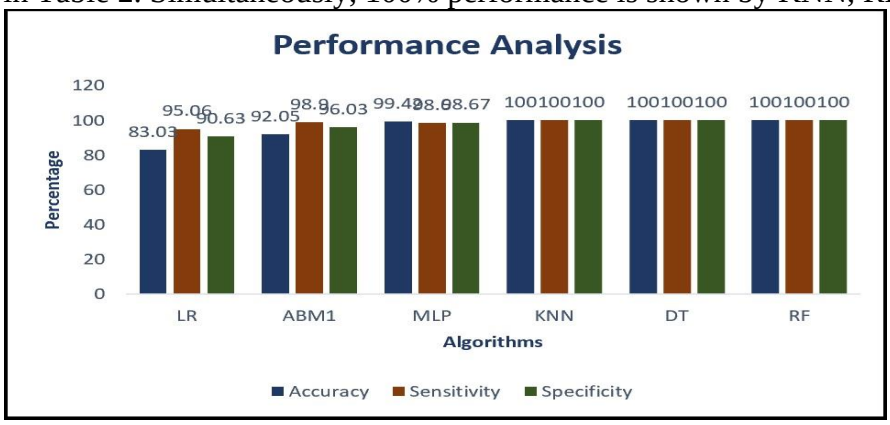

Figure 3: Classification results of all algorithms

Figure 3 shows the accuracy, sensitivity and specificityof all six methods, with the Logistic regression (LT) achieving the values 83.03%, 95.06% and 90.63% respectively.The AdaboostM1 (ABM1) obtained the values of 92.05%, 98.90% and 96.03% respectively. The Multilayer perceptron (MLP) achieves 99.42%, 98.60% and 98.67% respectively. The k-nearest neighbor (KNN), Decision Tree (DT) and Random Forest (RF) achieves the value of 100% on all metrices.

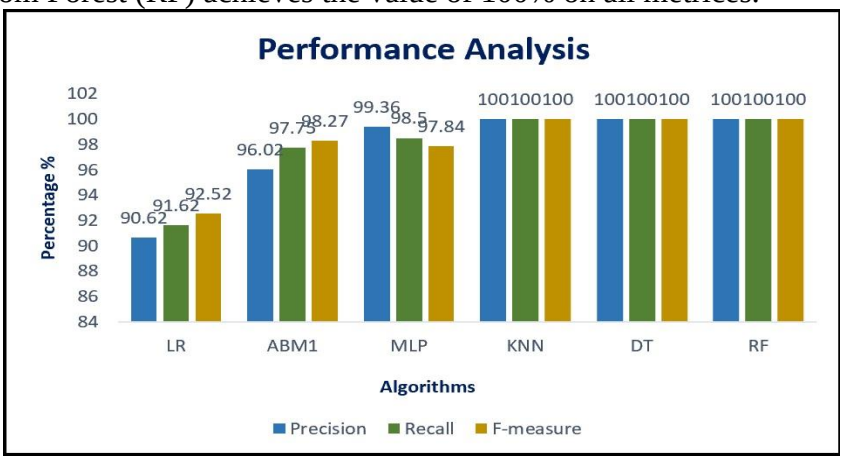

Figure 4: Precision, Recall, F-measure results of all algorithms

Figure 4 shows the precision, recall and f-measure of all six methods, with the Logistic regression (LT) achieving the values 90.62%, 91.62% and 92.52% respectively. The AdaboostM1 (ABM1) obtained the values of 96.25%, 97.75% and 98.27% respectively. The Multilayer perceptron (MLP) achieves 99.36%, 98.50% and 97.84% respectively. The k-nearest neighbor (KNN), Decision Tree (DT) and Random Forest (RF) achieve the value of 100% on all matrices.

V. CONCLUSION

Heart attacks are among the potentially lethal consequences of heart disease, making it a potentially dangerous condition. Owing to the possibility of an accurate illness prediction rate, machine

learning and data mining techniques are important because they may be used to forecast the presence of this disease. Here, we tested the effectiveness of machine learning techniques for heart disease prediction using a dataset on heart disease. We discovered that three classification algorithmsKNN, RF, and DTperformed exceptionally well with 100% accuracy. Furthermore, feature relevance ratings were computed for every feature for every deployed method, except for MLP and KNN. The feature significance score was used to rank these features.The goal of this research was to identify the most effective machine learning approaches. It found that, at least for this dataset, a variety of widely used and simple to use algorithms performed well. Although applying ML techniques is still in its infancy, there is reason to believe that it may be a very useful addition to patient care.

REFERENCE

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 02 June 2021].
[2] R.D. Canlas, "Data Mining in Healthcare: Current Applications and Issues", School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.
[3] Christoph Helma, Eva Gottmann, Stefan Kramer, "Knowledge discovery and data mining in toxicology", Stat. Methods Med. Res.,Vol. 9 (4), PP. 329–358, 2000.
[4] I.-N. Lee, S.-C. Liao, M. Embrechts, "Data mining techniques applied to medical information", Med. Inf. Internet Med,.Vol. 25 (2),PP.81–102, 2000.
[5] L. Parthiban, R. Subramanian, "Intelligent heart disease prediction system using CANFIS and genetic algorithm", Int. J. Biol., Biomed. Med. Sci.,Vol. 3 (3), 2008.
[6] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes", J. Clin. Epidemiol., Vol. 66 (4), PP. 398–407, 2013.
[7] S.K. Dehkordi, H. Sajedi, "Prediction of disease based on prescription using data mining methods", Health Technol.,Vol. 9 (1),PP. 37–44, 2018.
[8] M. Jan, A.A. Awan, M.S. Khalid, S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms", Res. Rep. Clin. Cardiol., Vol. 9,PP. 33–45, 2018.
[9] J. Soni, U. Ansari, D. Sharma, S. Soni, "Predictive data mining for medical diagnosis: an overview of heart disease prediction", Int. J. Comput. Appl., Vol. 17 (8), PP. 43–48, 2011.
[10] H.M. Islam, Y. Elgendy, R. Segal, A.A. Bavry, J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach", PP. 1–7, 2017.
[11] H.M. Le, T.D. Tran, L.A.N.G. Van Tran, "Automatic heart disease prediction using feature selection and data mining technique", J. Comput. Sci. Cybern.,Vol. 34 (1),PP. 33–48, 2018.
[12] M. Tarawneh, O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques, Acta Sci. Nutr. Health, Vol. 3 (7), PP. 147–151, 2019.
[13] R. Chitra, V. Seenivasagam, "Heartdisease prediction system using supervised learning classifier", Bonfring Int. J. Softw. Eng. Soft Comput.,Vol. 3 (1), PP. 01-07, 2013.
[14] C.B.C. Latha, S.C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Info. Med., Vol. 16, PP. 100203, 2019.
[15] S. Mohan, C. Thirumalai, G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", IEEE Access,Vol. 7, PP. 81542–81554, 2019.
[16] https://www.kaggle.com/johnsmith88/heart-disease-dataset [Accessed 02 June 2021].
[17] M.R. Rahman, T. Islam, T. Zaman, M. Shahjaman, M.R. Karim, F. Huq, J.M. Quinn, R.D. Holsinger, E. Gov, M.A. Moni, "Identification of molecular signatures and pathways to identify novel therapeutic targets in alzheimer's disease: insights from a systems biomedicine perspective", Genomics, Vol. 112 (2),PP. 1290–1299, 2019.
[18] Four Techniques for Outlier Detection, https://www.kdnuggets.com/2018/12 /four-techniques-outlier-detection.html.
[19] Md Satu, Syeda Atik, Mohammad Moni, A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus, 2019.

[20] S. Asaduzzaman, M.R. Ahmed, H. Rehana, S. Chakraborty, M.S. Islam, T. Bhuiyan, "Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective", BMC Bioinf., Vol. 22 (1), PP. 1–17, 2021.
[21] T. Akter, M.S. Satu, M.I. Khan, M.H. Ali, S. Uddin, P. Lio, J.M. Quinn, M.A. Moni, "Machine learning-based models for early stage detection of autism spectrum disorders", IEEE Access, Vol. 7, PP. 166509–166527, 2019.
[22] S.M. Vieira, U. Kaymak, J.M.C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection", International Conference on Fuzzy Systems, 2019.
[23] Z. Lei, Y. Sun, Y.A. Nanehkaran, S. Yang, M.S. Islam, H. Lei, D. Zhang, "A novel data-driven robust framework based on machine learning and knowledge graph for disease classification", Future Generat. Comput. Syst., Vol. 102, PP. 534–548, 2020.
[24] X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu, Z. Huo, M. Yu, J. Peng, "Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features", Sci. Rep. Vol. 9 (1),PP. 1–13, 2019.
[25] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Med. Inf. Decis. Making, Vol. 19 (1), PP. 1–16, 2019.
[26] T. Cover, P. Hart, "Nearest neighbor pattern classification", IEEE Trans. Inf. Theor., Vol. 13 (1),PP. 21–27, 1967.
[27] B.V. Dasarathy, "Nearest neighbor (NN) norms: NN pattern classification techniques", IEEE Comput. Soc. Tutorial, 1991.
[28] K.H. Raviya, B. Gajjar, "Performance Evaluation of different data mining classification algorithm using WEKA", Indian J. Research, Vol. 2 (1),PP. 19–21, 2013.
[29] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, "Supervised machine learning: a review of classification techniques", Emerg. Artif. Intel. Appl. Comput. Eng., Vol. 160, PP. 3–24, 2007.
[30] R.L. De Mantaras, E. Armengol, "Machine learning from examples: inductive and Lazy methods", Data Knowl. Eng., Vol. 25 (1–2), PP. 99–123, 1998.
[31] S. Vijayarani, S. Sudha, "Comparative analysis of classification function techniques for heart disease prediction", Int. J. Innov. Resear. Compute. Commun. Eng., Vol. 1 (3),PP. 735–741, 2013.
[32] L. Breiman, "Random forests", Mach. Learn., Vol. 45 (1), PP. 5–32, 2001.
[33] S.M.M. Hasan, M.A. Mamun, M.P. Uddin, M.A. Hossain, February.,"Comparative analysis of classification approaches for heart disease prediction", International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, PP. 1–4, 2018.
[34] J.R. Quinlan, "Induction of decision trees", Mach. Learn., PP. 81–106, 1986.
[35] J.A. Cruz, D.S. Wishart, "Applications of machine learning in cancer prediction and prognosis", Canc. Inf., Vol. 2, 2006.
[36] K. Li, G. Zhou, J. Zhai, F. Li, M. Shao, "Improved PSO_AdaBoost ensemble algorithm for imbalanced data", Sensors, Vol. 19 (6), PP. 1476, 2019.
[37] C. Zhang, Y. Chen, "Improved piecewise nonlinear combinatorial adaboost algorithm based on noise self-detection", Comput. Eng., Vol. 43, PP. 163–168, 2017.
[38] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, "Applied Logistic Regression",John Wiley & Sons,Vol. 398, 2013.
[39] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Med. Inf. Decis. Making, Vol. 19 (1), PP. 1–16, 2019.
[40] S. Dreiseitl, L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", J. Biomed. Inf., Vol. 35, PP. 352–359, 2002.
[41] K. Kwon, D. Kim, H. Park, "A parallel MR imaging method using multilayer perceptron", Med. Phys., Vol. 44 (12),PP. 6209–6224, 2017.
[42] S. Tajmiri, E. Azimi, M.R. Hosseini, Y. Azimi, "Evolving multilayer perceptron, and factorial design for modelling and optimization of dye decomposition by bio-synthetized nano CdS-diatomite composite", Environ. Res., Vol. 182, PP. 108997, 2020.
[43] Y. Azimi, "Prediction of seismic wave intensity generated by bench blasting using intelligence committee machines", Int. J. Eng., Vol. 32 (4), PP. 617–627, 2019.