Review on Cleveland Heart Disease Dataset using Machine Learning

Ruqiya[1,*], Muhammad Tayyab Yaqoob[2], Abid Muhammad Khan[2], Aftab Ahmed Shaikh[1], Noman Khan[1]

[1]Department of Computer Science, Sindh Madrasa-tul-Islam University, Karachi, Pakistan
[2]Department of Electrical Engineering, SSUET, Karachi, Pakistan
[*]Corresponding author: ruqiyaabbasi42@gmail.com

Abstract

According to the World Health Organization (WHO), heart disease has been a foremost source of death worldwide for the past 15 years. Therefore Medical diagnosis is typically performed mostly by doctors due to their training and experience. In the field of medicine, computer-aided decision support systems are enormously significant. Therefore, it is necessary to develop prediction systems that give information of different categories to readers. According to the study, hybrid intelligent algorithms increase the heart disease prediction system's accuracy. Hence, recognizing cardiovascular problems including heart attacks, coronary artery diseases, etc. by routine clinical data analysis is an important task; early identification of heart disease may save many lives. In this article, have reviewed various papers related to the Cleveland heart disease dataset that used one or more machine-learning algorithms to forecast congestive heart failure. In one of the above-mentioned papers, the result of utilizing Random Forest is almost 100%. To ensure that predictions made using machine learning algorithms produce accurate outcomes. Applying machine learning algorithms to heart disease treatment data can produce results that are just as accurate as those found in heart disease diagnosis.

Keywords—Machine learning, Cleveland heart disease dataset, Limitations

1 Introduction

The heart, which is responsible for pumping blood that is enriched with oxygen to further body organs via a network of arteries and veins, is the most significant organ in the human body. Heart disease is an illness that affects our hearts [1]. WHO reported that the foremost reason for death is congestive heart failure. Over the previous 15 years, there will be an estimated 17 billion deaths worldwide due to cardiovascular disease. [2-3] The two leading causes of death are heart disease [4] and stroke [5]. Heart disease patients have a variety of symptoms, including chest pain, dizzy sensation, and heavy sweating. The main sources of heart disease are smoking [6], hypertension [7], diabetes [8], fatness [9], and other reasons. Using invasive techniques to diagnose the disease is costly and uncomfortable. Heart disease has occurred as a severe health worry for many persons due to its high mortality rate throughout the world [10]. Identifying cardiovascular problems like heart attacks, coronary artery diseases, etc. via routine clinical data analysis is an important endeavor; early detection of heart disease may save many lives.

The death rate may be decreased with improved monitoring of heart patients. People typically consult a heart specialist quite late. So if there is a decision-support system that can forecast disease at an early phase and may help the patient to reduce the death rate also the main goal of this decision-support system is to create a system that, in the absence of a doctor, can be utilized by any literate person to make an early diagnosis of heart disease [11-12]. Even with this kind of system, the doctor may receive help in making a disease. The beauty of this approach is that it won't require a physician with expertise in heart disease because it will just be based on clinical data. According to WHO 2020, with over 17.9 million deaths per year, cardiovascular diseases (CVD) will remain to be the main source of death worldwide [13]. A key method of lowering this toll is early CVD detection. Data mining is one of the various methods for enhancing disease detection and diagnosis [14]. These methods permit the extraction of hidden information and the discovery of correlations between attributes within a dataset,

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2101.11

making them a potential approach for the classification of CVDs. One of the most important challenges facing health organizations is providing patients with high-quality clinical treatments that are inexpensive [15-16].

A decision support system to quickly and affordably identify cardiac disease using clinical data can be designed using machine learning (ML) techniques [17]. A decision-making system of this sort can help clinicians identify the disease at an initial phase. And can transform this kind of decision support system into a health chatbot system that the patient can use to detect cardiac trouble early. And an online AI-based chatbot advises the patient [18] As a result, risk management procedures must be followed to reduce safety threats. Patients who use these medical chatbot systems can monitor their health. Mobile health technologies capture patient data in real time and give better health facilities. It enhances patient monitoring without the need to visit the hospital [19]. Therefore, ML is a type of AI that uses experience to autonomously comprehend and improve itself without the aid of any other program [20] and also used to categorize hidden patterns and provide some medical expertise that would assist the specialists in arranging and providing maintenance to predict heart failure disease [21]. As a result, the patient must think about the risk issues for heart disease. Numerous lifestyle factors have a significant impact on heart disease even though it includes an inherited component. Radiation therapy for age, gender, family history, smoking, chemotherapeutic medicines, tumors, poor food, hypertension, high plasma cholesterol levels, diabetes, obesity, and pressure are some of the known risk features for heart disease. The various risk issues that patients may encounter contribute to the increase in cardiovascular disease. One of the primary challenges in the data analysis is the forecast of cardiovascular disease. Cardiovascular disease is becoming a larger issue since a few years ago. Many researchers must continue working in the same direction to identify the most significant risks of heart disease by precisely anticipating the overall risk. Heart disease is generally portrayed as a silent killer that causes the deaths of those without obvious symptoms. In high-risk patients, the initial study of cardiac disease reveals how it contributes to life-changing changes and then works to lessen the problems. This study aims to forecast the occurrence of Heart Disease by analyzing patient data that classify individuals as either having heart illness or not using ML algorithms. ML techniques might be useful in this situation. Even though heart disease can manifest itself in a variety of ways, the following list of heart disease risk factors typically has an impact on whether someone is ultimately determined to be at risk for the condition.

This paper first presents heart disease studies using ML, after which demonstrates the significance of ML in the field of medicine. This studied various papers related to the Cleveland heart disease dataset and proved which ML algorithm performs better on this dataset. This paper also builds on previous research which identified a dataset limitation and attempted to offer a solution, accordingly, in section 2 of this paper. Present methodology and highlight research from previous paper and briefly introduce the dataset and make a suggestion regarding the dataset limitation, in section 4 exhibit the answers and display data in the form of graphs. The conclusion of this paper is presented in section 5.

2 Literature Review

The previous research in the field of heart disease diagnostics utilizing the Cleveland dataset is described in this section. Several researchers utilized various ML algorithms on the Cleveland heart disease dataset, but they got diverse accuracy results, as discussed below.

2.1 Prediction of Heart disease

In this study, the author [22] recognized and predicted human heart disease, the author used the Cleveland heart disease dataset, which included 303 records and 6 samples with missing values. The data originally contained 76 features, however, only 13 are likely to be mentioned in all published studies, with the remaining feature describing the effect of the condition. Another generic dataset used by researchers for the prediction procedure is the Z-Alizadeh Sani dataset, which contains the data of 303 patients with 55 input components and a class label variable for each patient. To test the effectiveness of the algorithms using several measures, including classification accuracy, sensitivity, specificity, and F-measure. The author used nine machine-learning classifiers for the final dataset before and after hyperparameter tuning. To guarantee accuracy on the standard heart disease dataset, authors also do appropriate pre-processing, standardize the dataset, and optimize hyperparameters. The K-fold cross-validation technique was also built up by the author to train and validate the machine-learning algorithms. Finally, the experimental findings showed that hyperparameter adjusting increased the prediction classifiers' accuracy and that data standardization and hyperparameter tuning of the ML classifiers provided notable outcomes. The maximum level of

prediction accuracy, 90.16%, is achieved by classifiers. Multinomial Naïve Bayes (MNB) has the lowest accuracy of 59.01% and the overall worst performance. Additionally, contrast the correctness of the dataset before and after standardization. The accuracy of the ET and AB classifiers is 90.16%, demonstrating the benefit of the dataset's standardization. However, SVM has the highest accuracy (96.72%) when it comes to fine-tuning the hyper-parameters

2.2 Machine Learning-based Support System

In this study, author [23] built a machine-learning support system that can increase accuracy, the dataset used in this study was taken from the Cleveland dataset author uses the Python programming language, and also uses the libraries such as Matplotlib, Numpy, and Keras. Using four ML models, the Cleveland, Hungary, Switzerland, and Long Beach (CHSLB) and Cleveland datasets were used to predict coronary heart disease. The data was pre-processed using a variety of approved procedures and approaches to boost the detection accuracy of the applied ML models. When using the CHSLB and Cleveland datasets, the KNN model outperforms the other models with an accuracy of 100% and 97.82%, correspondingly. In the context of the CHSLB dataset, the RF, AB, and DT models' accuracy is 99.025%, 96.103%, and 100%, respectively. This kind of process intelligence method is essential in medical diagnostics. Due to the increased detection accuracy of the used ML algorithms, a computer-aided smart system was created using a readily available internet-based cloud hosting platform.

2.3 Comparison of Machine Learning Algorithms

This research compares and analyses several classifiers, pre-processing, and dimensionality reduction strategies to determine how well they predict the presence of cardiac disorders. According to the authors [24], the most important subset of features to predict the presence of heart diseases is made up of PES, EIA, CPT, MHR, THA, VCA, and OPK. Nave Bayes classifier provided the best performance prediction, and Chi-squared feature selection was the data mining method that decreased the number of features although maintaining their quality. With 165 and 138 observations, respectively, the used Cleveland dataset's 303 observations show fairly balanced positive and negative samples. Seven well-known classifiers were experimentally compared in this study using various data mining methods, such as feature extraction and feature selection. This study's proposed methodology included data collecting, data pre-processing, and balancing approaches in addition to seven basic stages (oversampling and under-sampling). Dimensionality reduction methods like PCA feature extraction and Chi-squared feature selection were also researched. The primary focus of this research is not only improving the performance of weak classifiers but also closely examining the well-known dataset and analyzing the effects of various pre-processing methods.

2.4 Classifying the Numerical and Categorical Features

The author [25] develop a model for heart disease forecast that is more precise and better. Quickly identifying new patients, speeding up diagnostics, lowering the number of heart attacks, and saving lives are the specific objectives. The Cleveland database and the heart disease database and National Cardiovascular Disease Surveillance (NCDS) System are two databases devoted to heart diseases. The four combined databases included in this study's Cleveland heart disease dataset, are Hungary, Switzerland, and VA Long Beach. There are 14 properties in the dataset, and each one has a value. A subset of this dataset contains 713 male and 312 female records, with 1025 patient records totaling a variety of ages. 75% of the training data and 25% of the test data are used in each classifier. Before and following the use of standardized datasets, classifier accuracy is also measured. The majority of the listed algorithms, including k-fold cross-validation, LR, KNN, SVM, Nu SVC, DT, RFC, AdaBoost, GBC, NB, LDA, Q DA, NN, and ensemble methods, are not neural network-based and have higher accuracy. On the consistent dataset, the Naive Bayes, as well as Support Vector Machine classifiers' accuracy, fell. On the consistent dataset, several classifiers, including RF, DT, GB, and NN, showed significant accuracy increases. The greatest prediction accuracy for the RF and DTC are 100% and 98.80%, respectively [19].

2.5 Improved Heart Disease Prediction using X2 Statistical Test

In this study, the author [26-27] improved through the application of the X2 statistical feature selection techniques. The feature selection approach was used to determine the six most important features for heart disease prediction. The 2-based SVM heart disease prediction model was created and evaluated using Python as well as the sci-kit learn module. The two acquired datasets, Cleveland and Statlog (Heart), were separated into train and test sets. The model was

developed using training data, and testing data were used to assess the model's performance. To train and test our proposed model, both datasets were divided into a train set and a test set using a split ratio of 75:25. An improved model was put into place to lessen the computational load and improve the accuracy of heart disease diagnosis and prognosis. A classification model based on the ML (SVM) algorithm was utilized to increase heart disease diagnosis. On two well-known datasets related to cardiac disease, this model was run. In the Cleveland and Statlog datasets, the results indicated improving accuracy from 84.21% to 89.47% and from 85.29% to 89.7%, respectively. Additionally, the system's feature count was lowered from 14 to 6, which translated to a 42% reduction in computational effort overall. The author hopes that this work will aid in the future creation and application of systems for the detection and prediction of cardiac disease.

2.6 Review of CHD

The author [28] used CHD Database 13 HF features that are frequently used are taken into account in this study. The pre-processing of the dataset includes null value verification, loading Python libraries, as well as splitting the dataset into training and testing data. These steps were inspired by the analysis of numerous recent research papers on the prediction of heart disease using various data mining and ML techniques and algorithms. According to the author, several data mining and ML techniques are employed to forecast cardiac disease. In several trials, distinct patient datasets with heart disease are employed. The majority of tests use data from UCI's online Cleveland database. This survey taught the author how to use several machine-learning approaches to anticipate cardiac attacks. Additionally, it was discovered in several study studies that the hybridization of two or more dissimilar algorithms.

2.7 Heart Disease Prediction using Hybrid Algorithms such as MLP-PSO

The CHD dataset was used in this study to develop and compare a variety of intelligent systems based on ML algorithms for determining a person's likelihood of developing heart disease. Author [29] creates binary classification models that may be utilized as diagnostics for heart disease, this study goals to evaluate the efficacy of MLP neural network classifiers trained with PSO and several alternative supervised ML methods. A fully linked MLP network is represented by the complicated function known as MLP, which accepts numerical inputs and generates numerical outputs. It consists of three layers: the domain's raw input is taken in by the input layer, features are extracted by the hidden layer, and predictions are created by the output layer. Care must be taken when choosing the MLP's hyper-parameters, which include the number of hidden layers and neurons. All research experiments for this project were performed using Python. The author used 70% of the data for training and 30% for testing for all models. The author trained the proposed ML models utilizing the five-fold cross-validation method. Five-fold cross-validation is often used in ML to compare as well as select a model for a particular prognostic modeling assignment meanwhile it is simple to design. The experiments demonstrated that the proposed MLP-PSO outperformed all previous procedures, by an accuracy of 84.61%. The results showed that the MLP-PSO model can help medical professionals diagnose patients more correctly and suggest better therapies. Overall, the ability to detect cardiac disease using neural networks trained with PSO is promising.

2.8 Detection of Heart Disease

Coronary arteriography (CAG) is a procedure that can be used to precisely diagnose cardiovascular disease. Although this is not suited for the yearly physical, an invasive method can also be used to detect coronary heart disease. The author [30] of this paper offers an ML-based decision support system with the aim of heart disease prediction. The data collection, pre-processing, and model development phases of the proposed system are listed below. Class balancing and feature selection are carried out during the pre-processing phase. 561 cases in the sample fall under class 0, while 629 instances fall under class 1. In this dataset, there are no missing values, and all features are of the int and float data types. Then, divide the training and test datasets in a ratio of 70:30, apply all classification models to the training dataset, and determine the accuracy of each model. The ANOVA test is then applied to the dataset using the feature selection technique, and it identifies the top 7 datasets that have a strong correlation. Once you have the top 7, match the data to classification models and determine accuracy. With the use of classification algorithms employing Python ML technology, a classification model needs to be constructed using these features. Data about CHD from the UCI library was used in the study CHD. Eight of the 14 features in the sample are categorical traits, and the remaining six are numerical traits. Several ML models are needed for the methodology being deliberated. The study is done using the confusion matrix, and the best algorithm is

determined by comparing the accuracy of each one. Thus, the work's effectiveness has been established. This technique could permit the accurate and early forecast of cardiac disease. Numerous additional ML algorithms can be applied for the most accurate investigation and earlier diagnosis of cardiac illnesses in the potential future. This requires additional diagnosis.

2.9 Prediction of Diabetic Coronary Disease

In this research author [31] forecasting diabetic coronary artery disease utilizing machine-learning approaches, such as clustering, K-means, Hierarchical, Gaussian, Hidden Markov, density estimation, and credit assignment, is examined in detail in this work. Additionally utilized South African sources as well as the Cleveland HD dataset. The healthcare sector faces a big challenge in predicting diabetic coronary heart disease, and this study has looked at the available machine-learning classification procedures for doing so. The study and organization of several hundred publications on the subject of categorizing data with ML. The study quickly explains how diabetic heart disease can be predicted using supervised, unsupervised, and reinforced ML techniques. The resources and dataset used, the number of attributes and entries, and the quality of the proposed models all vary. The database will be able to make smarter decisions with more data. It is possible to increase the system's scalability and exactness by looking into several options. After this research is finished, the forecast algorithms will be trained as well as tested by data from a substantial sample of people with diabetes and coronary artery disease nationwide. Applying the prediction models to sizable datasets is necessary to evaluate their accuracy.

2.10 Heart Disease prediction framework using Smote-Xgboost

The smote-xgboost method is used in this paper author [32] presents a novel framework for predicting heart disease. To identify important features from the dataset and avoid model overfitting, the author first suggests an information-based feature selection strategy. Second, the author uses the Smote-Enn algorithm to balance uneven data and create sample data with roughly equal positive and negative categories. Using sample data, compare the xgboost algorithm's propensity to forecast outcomes to those of five other conventional algorithms. The research sample for this work is the return visit data of actual patients in a hospital. Give it the Heart Disease Dataset (HDD). 37 features, comprising numeric and category features, a total of 4232 samples in the dataset. Major adverse cardiovascular and cerebrovascular events (MACCE), where 0 denotes no occurrence and 1 indicates occurrence, are the prediction target. The quality of the data will have a significant impact on how well the model can predict the future, thus preparing the data before training is essential. To deal with missing values, the author uses the following technique. For class variables, the author constructs a new class to represent null values; for numeric variables, we eliminate feature columns with missing value rates of more than 70% and classify them as invalid, after which we replace the remaining feature columns with missing values with the mean values. Also, to improve the data's relevance, normalize the data using the maximum-minimum norm technique. Exploratory data analysis findings on this dataset and exploratory data analytics were done to better understand its properties. The findings of the analysis are described in the subsection that follows. The histogram of the frequency distribution offers a quick summary of the data's dispersion and central tendency. The height of each rectangle and the rate of the amount of the values visually depict the distribution of different attributes. The degree of feature correlation also has an impact on the model's propensity to predict outcomes. The results demonstrate that, with a prediction accuracy of 93.44%, the model developed in this study achieves very well across all four assessment indicators. Also, evaluate the selected algorithm's feature relevance, which has a substantial impact on the prediction of heart illness.

2.11 Prediction using Machine Learning using Genetic Algorithm

In this study, author [33] searches were made for the most reliable heart disease prediction algorithms. To forecast patients with cardiac diseases, five classification algorithms with the use of the Cleveland and Framingham dataset, and the models' performance is evaluated. Cleveland is the best dataset, based on observation, as it uses the sklearn module of Python Jupyter Notebook and has the fewest lost values while offering all 14 qualities as predictors. This work goal is the introduction of several heart disease prediction models that mix traditional ML methods with a genetic algorithm (GA) to select the best features. The genetic algorithm used by the improved prediction model outperforms more traditional models. Another finding of the study is that the performance of the genetic algorithm-enhanced prediction models is better than that of conventional prediction models, which had a classification of 100% accuracy for the Cleveland heart disease dataset and 91.8% accuracy

for the Framingham dataset. Having features that can be used for multivariate data analysis that have both discrete and continuous values. The primary goal of the newly suggested genetic model is to improve the accuracy of the heart disease prediction model and get rid of any patient misdiagnosis. There is still room for development, though. The research will expand in the future to find and incorporate more features, and other categorization approaches, such as deep learning, will be used. Future wearable technology will be based on convolutional neural networks, and the proposed work will have access to trained and validated datasets.

2.12 Survey on Heart Disease

An author [34] survey of healthcare techniques using deep learning is presented in this work. To identify three common ailments, deep learning models were utilized in this paper's explanation of healthcare. All in all, there has been a lot of research done on deep learning models for the healthcare industry, but there are still numerous obstacles to overcome. It is a huge difficulty to design these healthcare systems approaches, but it can enhance the processing system to create gadgets that are more effective and efficient. To be employed as broad models that can handle any form of input, CNN models' structural design needs to be enhanced. Hope this survey will be a new stage in the development of inventive approaches used in healthcare and will result in more intelligent CAD systems.

2.13 Comparative Analysis using Data Mining

In this work, the author [35] uses data mining and ML approaches, the state of the art for many medical decision support systems for the prognosis of heart disease observed. Heart diseases have been predicted using a variety of classification techniques, to improve the accuracy of forecasting the early phase of the disease, more complex models are required, which combine a variety of geographically different data sources. It was discovered that the CHD dataset, which only has 303 occurrences and 14 characteristics, was utilized by the majority of investigations. As a result, developing such predictive models for people with heart disease has had only sporadic success. When it comes to accurately portraying a certain geographic area, the sample size is quite limited and constricted. A tiny number of studies that also used other data sources employed a single dataset with few heart disease features. It was unable to generalize the various classification accuracy results for heart disease prediction as a result. This is the main goal of upcoming research, which is progress-oriented. Further various heart disease datasets from geographically different sources with additional attributes should be looked into for constructing more effective ML models to obtain a more comprehensive classification and forecast accuracy. This would make it possible to classify and predict heart diseases more accurately in the early stages, which would lower the rising rates of morbidity and mortality from CVDs.

2.14 Build Decision Support System using Data Mining

Building an effective data mining-based intelligent medical decision support system is the main goal of this work. Author [36] improves the diagnosis of heart diseases through the application of data mining classification algorithms. To do this, a study of the literature on data mining studies used to diagnose cardiac disorders was done. Five classification algorithms were implemented using MATLAB. This study made use of two datasets. The first one was got from the Cleveland Clinic Foundation and has 303 records, 297 of which are complete and six of which have values that are missing. The second dataset was the Statlog dataset, which had 270 records. Both of these datasets were pre-processed to yield 14 attributes from their original 76 attributes to decrease the number of variables. With an accuracy rate of 99.0%, the decision tree performs better than other classifiers, followed by Random forest. This is true because both of them use a similar method, but the random forest can create decision tree ensembles. In our situation, the decision tree beat its ensemble variant even though ensemble learning has been shown to generate higher results.

2.15 Identifying High-risk patients suffering from Heart Disease

To compare different categorization prediction models and predict cases of heart disease, this study author [37] employed heart disease data from the UCI ML repository, the proposed model was developed to evaluate how well different feature selection strategies and machine-learning techniques predict the onset of cardiovascular disease. Feature selection algorithms like ANOVA and LASSO were used to pinpoint important properties. Following that, the effectiveness of well-known classifiers used in similar works. The training data procedure will use the cross-validation method. This study also employs various performance indicators to assess the outcomes of the statistical analysis. The suggested model is composed of four functional components: evaluation, feature selection,

ML, and data pre-processing. The most informative features were then selected using the three distinct feature selection methods (Random Forest, ANOVA, and LASSO), and the classifiers were trained using the filtered input to provide new results. Every feature was normalized as well as standardized before being applied to classifiers. Additionally, 10-fold cross-validation was used to confirm the stability and dependability of these models. The following chart compares the rankings of informative features according to various feature selection techniques. Among them, the sklearn approach for feature selection will be used to apply the ANOVA method. In this study, 10-fold cross-validation is performed, which indicates that 90% of the data will be used as a training set and 10% of the data will be used as test data in each training round. The SVM RBF performs reasonably well, with an accuracy rate of 84.5%.

2.16 Apply Neural Fuzzy Hybrid-Based System

This study author [38] proposed the neural fuzzy inference system (NFIS) as a means of describing training data made up of n-dimensional function space. The error-calculating module of the NFIS enhances learning advice once mistakes have been measured. When a membership function is first stated, its parameters are activated and learned by being used in operations. The suggested methodology, which also includes reliable and unreliable measurements, causative variables, and data matrices, has been tested. This study included more than 13000 fuzzification rules to provide the best decision-making, a normalization procedure, as well as establishing strategies to make it possible to calculate the probability of having a heart attack, and it obtained a 94 percent accuracy rate. This study can be expanded to create auto-altering and advisory systems with hardware peripheral circuit device integration. As a means of describing training data made up of n-dimensional function space in this study. The NFIS includes an error-calculating module to aid in learning instructions once errors have been measured. A membership function is initially built, and its parameters are then activated and learned as needed for an activity. The genetic algorithm and neural fuzzy inference system have been used to tackle the study job. Fuzzy sets are used to create membership functions after all the rules have been supplied into the fuzzy system. The procedure of fuzzification was applied to the input. (vi) Crisp output has been displayed as a % probability of heart attack following the training and defuzzification stages. The author discovered numerous Cleveland dataset limitations, which also revealed the study's potential future application. Here are a few of the restrictions: Out of 303 records, only 282 were deemed to be correct, and the other entries have inaccurate information. (ii) Data are not evenly distributed across age groups, with the majority of those showing cardiac disease occurring in people between the ages of 40 and 75. Male and female records are not equally represented in the data, and the amount of male records is significantly higher than that of female records. (iv) Most patient records only pertain to male patients; relatively few details concern female patients. Only 127 of the 282 valid data points were judged to be dangerous, while the remaining records related to healthy patients. There were only six reports from adult group 1 (18–40) and one from the old group2 out of 127. It is therefore obvious that the age categories are not clearly and accurately defined.

2.17 ML contribution on Coronary Heart Disease

This work aims to comprehend the detection accuracy, classification strategies, and limits of some machine-learning models (MLMs). The author [39] used different machine-learning algorithms and selected the features that would best accurately detect the disease using various feature selection strategies. 96.29% classification accuracy was attained, and to increase accuracy with the least amount of time and effort, it is necessary to construct individual MLMs for the recognition and assortment of exact features. To increase the accuracy in terms of percentages, several researchers employ hybrid systems, which layer two or more categorization techniques (based on chosen symptoms and attributes of a human being). Sometimes it isn't more time-consuming and ineffective. As a result, the author creates flexible MLMs with feature selection and reduction methods. Therefore, enhancing accuracy is the main goal of the current study. Include any potential future directions or uses for your research as well. Different ML models are used to diagnose heart disease utilizing feature selection/deduction approaches using the coronary heart disease dataset, which is taken into consideration for study reasons. According to the researcher, various strategies performed better based on all the results (with or without validation). Every approach, however, possesses the inherent ability to perform better than others depending on the circumstance. To more effectively eliminate unnecessary and redundant features, the author employed K-NN, NB, and ANN MLMs together with feature space mapping (FSM), separability split value (SSV) feature selection approaches, and Relief F and Rough Set (RFRS) method.

2.18 Prediction using ML Supervised Classifiers

To compare and analyze the accuracy of various algorithms, the author [40] of this work used the dataset from the UCI ML repository called "Cleveland Heart Disease Dataset" to deploy ML and deep learning techniques. The author intended to integrate supervised classifiers and deep learning algorithms in an optimized manner. The component of data pre-processing where the author needs to prepare or clean the data to get better results from techniques is called data pre-processing. Results of applying procedures and comparing the accuracy of several techniques to determine which one is more effective. So the author deployed techniques using supervised classifiers and deep learning techniques; deep learning techniques using Rmsprop optimizer gave the best performance, which was 94.01%. Moreover, as compared to previous work, machine-learning techniques weren't as effective. Decision trees outperformed other algorithms in ML by providing the best performance.

2.19 Monitoring Health Issues using ML and Cloud Computing

This paper author [41] introduces the Health Cloud system, which uses cloud computing as well as ML to investigate the health condition of patients. The main goal of this study is to offer the "best of both worlds" by uniting the knowledge required for the individual to comprehend the disease in proper detail with a correct prediction of heart disease whether they have or not. This was performed using the well-known heart dataset from UC Irvine called CHD and numerous machine-learning models were used and trained. The accuracy, precision, cross-validation results, sensitivity, specificity, and AUC scores were used to assess the models' performance. This study evaluates various ML algorithms to build the most precise model that satisfies Quality of Service (QoS) standards. The figure of merit of these ML models is efficiency which is compared with metrics like Accuracy, Sensitivity (Recall), Specificity, AUC scores, Execution Time, Latency, and Memory Usage. To fully validate the results, these ML algorithms underwent a 5-fold cross-validation process. Logistic regression, with an accuracy rate of 85.66%, was found to be the most appropriate model among those examined. This model's accuracy, recall, cross-validation mean, and AUC score were, in descending order, 95.83%, 76.67%, 81.68%, and 96%. On Google Cloud Firebase, the algorithm and the mobile app were evaluated using both previously observed data from the dataset and newly discovered data. Utilizing this technique can help people make self-diagnosis decisions and keep track of their health [35].

3 Materials Methodology

This paper's primary goal is to compare and contrast various papers that have been written about the CHD dataset. We read 30 journals, including papers from Springer, Inderscience, Elsevier, and IEEE. In all of these papers, the authors used ML algorithms, the studies' use of this dataset had limitations, and therefore we highlighted those limitations and offered solutions. To present the graphs, we utilized Jupiter notebook as a tool and Python language.

3.1 Dataset Description

The dataset used in this paper is a CHD dataset, which is available in Kaggle [44], and the University of California, Irvine, and UCI the most popular dataset used by experts, in this dataset the total number of records is 303, and 14 attributes where 13 are independent variable, and 1 dependent variable (target or output variable). The output variable comprises the outcome of the invasive coronary angiography, which specifies whether the patient has coronary artery disease or not. Labels 0 mean the absent of CHD, and 1-4 denote the presence of CHD, respectively. The majority of studies using this dataset have focused on merely attempting to differentiate between presence (values 1, 2, 3, 4) and absence (value 0).

3.2 Limitations and Suggestion

Here are a few of the limitations of the Cleveland dataset there are only 303 records in all. Only 282 out of 303 records were verified to be accurate, and some of the others included inaccurate information. Data are not evenly distributed in different age groups, with the majority of those showing cardiac disease occurring in people between the ages of 40 and 75. Male and female records are not equally represented in the data, and the number of male records is significantly higher than that of female records. Most patient records only relate to male patients; relatively few details concern female patients. Only 127 out of 282 valid data points were judged to be dangerous, while the remaining records related to non-disease patients. There were only six records from adult group 1 (18–40) and one from the old group 2 out of 127. It is therefore noticeable that the age categories are not clearly and accurately defined. The data originally had 76 features; however, all published work is likely to relate to only 13 of these, while the remaining feature describes the impact

TABLE 1: **Cleveland heart disease dataset**

Features	Description
Age	Years of age
Sex	Sex (male = 1; female = 0)
Cp	Type of chest pain Standard angina is value 1, Atypical angina is value 2, Non-anginal pain, value 3, Value 4 indicates asymptomatic
Trestbps	Blood pressure at rest at the time of hospital admission (in mm Hg).
Chol	Per deciliter, the amount of serum cholesterol.
Fbs	Above 120 mg/dl of fasting blood sugar is present (1 is true; 0 is false).
Restecg	Findings of resting electrocardiography Value 0: regular, Value 1 indicates that the ST-T wave is aberrant (T wave inversions, ST elevation, or ST depression of greater than 0.05 mV), Value 2: demonstrating potential or actual left ventricular hypertrophy according to Estes' criteria.
Thalach	Maximal heart rate reached
Exang	Exercise-induced angina (1 = yes; 0 = no)
Oldpeak	Compared to rest, exercise promotes ST depression.
Slope	The incline of the ST segment of the summit exercise Value 1 slopes upward, Value 2 is level, and Value 3 is sliding downward.
Ca	Main vessels colored by fluoroscopy in number (0–3) (for calcification of vessels)
Thal	Results of the nuclear stress test (normal: 3, fixed: 6, and reversible: 7)
Num	Target variable (angiographic disease status) showing the presence of heart disease in any major vessel Value 0: A diameter reduction of 50%, Value 1: More than 50% diameter narrowing

of the condition. Additionally, the output attribute displays the results in the ranges of 0, 1, 2, 3, and 4, where 0 already signifies the absence of disease while ranges 1 to 4 indicate the presence of disease but are not explicitly detailed from where they are related. Therefore, in our opinion, if this dataset were to be collected again, the number of records would increase, and since no one had previously used deep learning on this dataset due to the small number of records if it were to be collected once more, deep learning can be used to improve performance.

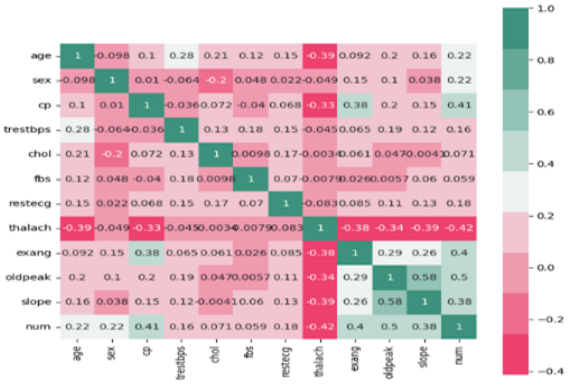

Fig. 1: The Correlation between columns

4 Results and Discussion

In this paper we use the CHD dataset and observed 303 heart disease patients' 13 attributes and 1 target attribute whereas the dataset includes 91 female patients and 212 male patients and the age range from 29 to 77 years with the average being 54 years old. 138 people in the obtained dataset are heart disease-free, while 165 people are. This dataset has no missing data. In this study, the graph was created using the Python programming language, and the libraries were using the Matplotlib and Pandas libraries. We send the data frame and this frame plot correlation based on input and output, as shown in Figure 1's correlation graph. The co-relation is virtualized in this graph so we can see how one column is related to other columns. Most strongly connected is the pink color column. Figure 2 describes the number of patients who are afflicted with the disease. The values range from 0 to 4, with 0 signifying absence and 1, 2, 3, and 4 signifying disease presence, respectively. Figure 3 pie chart displays the same output variable result in percentage form; here, values of 0 denote 54.1% of disease-free patients, 1 signify 18.2%, 2 signify 11.9%, 3 signify 11.6%, and 4 denote 4.29%. Figure 4 displays the percentage of male and female records, which is 68% male and 32% female. Figure 5 displays the age of the patient, and it is obvious that there are more records for patients between the ages of 44 and 58.

4.1 Exploratory Analysis

This heat map presents the correlation between all the variables, the green color signifies a greater correlation.

4.2 Outcome results in count plot method

This graph shows the total number of outcomes in numeric form this graph shows how many patients

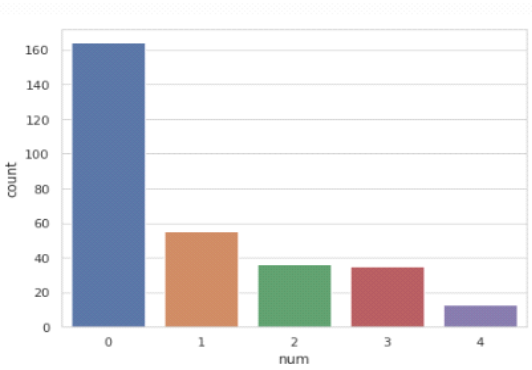

Fig. 2: The output attributes results

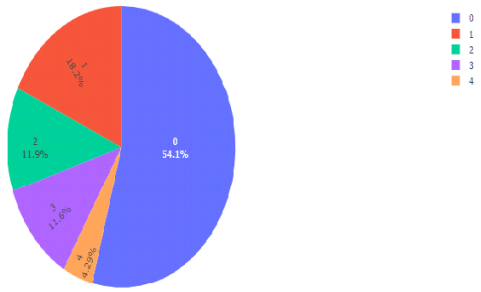

Fig. 3: The output attributes in percentage form

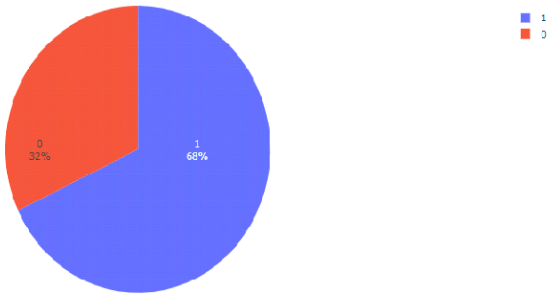

Fig. 4: The total number of male and female records

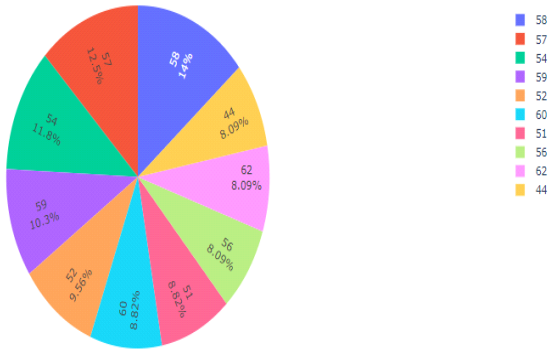

Fig. 5: The age of the patient in percentage form

affected form Heart disease.

4.3 Outcome results in chart method

This graph shows the total number of outcomes in percentage form this graph belongs to graph 4.2 but are same but this present in percentile form.

4.4 Gender results in chart method

This graph show the total number of Male and Female present in the dataset.

4.5 Different age of patient results in chart method

This graph shows the total number of patients in the dataset most of the patients is belong to 44 to 58 age.

5 Conclusion

The main cause of death worldwide is heart disease, Machine learning can be used to predict heart disease early and have a substantial influence on people who are at high risk for getting it. This helps to reduce potential problems in the field of medicine and helps to prevent the majority of people from developing heart disease each year. It is crucial to identify cardiovascular problems, such as heart attacks and coronary artery diseases, through routine clinical data analysis because doing so may help save many lives. With better cardiac patient monitoring, the death rate might drop. People regularly wait a long time to seek out a cardiac specialist. This study reveals that the Cleveland dataset has limitations and is suitable for small-scale research where Machine learning algorithms are applied. However, for large-scale investigation recurrent neural network (RNN) and Artificial neural network (ANN) modeling can be used. Moreover, a comprehensive investigation can be possible by increasing the number of records.

References

[1] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” J. Reliab. Intell. Environ., vol. 7, no. 3, pp. 263–275, 2021.

[2] A. Y. Firus Khan et al., "The Malaysian HEalth and WellBeing AssessmenT (MyHEBAT) study protocol: An initiation of a national registry for extended cardiovascular risk evaluation in the community," Int. J. Environ. Res. Public Health, vol. 19, no. 18, p. 11789, 2022.

[3] M. De Hert, J. Detraux, and D. Vancampfort, "The intriguing relationship between coronary heart disease and mental disorders," Dialogues Clin. Neurosci., vol. 20, no. 1, pp. 31–40, 2018.

[4] P. Khairy et al., "Sudden cardiac death in congenital heart disease," Eur. Heart J., vol. 43, no. 22, pp. 2103–2115, 2022.

[5] V. L. Feigin et al., "World stroke organization (WSO): Global Stroke Fact Sheet 2022," Int. J. Stroke, vol. 17, no. 1, pp. 18–29, 2022.

[6] C. W. Tsao et al., "Heart disease and stroke statistics—2022 update: A report from the American Heart Association," Circulation, vol. 145, no. 8, 2022.

[7] E. Herrett et al., "The importance of blood pressure thresholds versus predicted cardiovascular risk on subsequent rates of cardiovascular disease: a cohort study in English primary care," Lancet Healthy Longev., vol. 3, no. 1, pp. e22–e30, 2022.

[8] V. Miller, R. Micha, E. Choi, D. Karageorgou, P. Webb, and D. Mozaffarian, "Evaluation of the quality of evidence of the association of foods and nutrients with cardiovascular disease and diabetes: A systematic review," JAMA Netw. Open, vol. 5, no. 2, p. e2146705, 2022.

[9] [40] J. N. Garcia, C. N. Wanjalla, M. Mashayekhi, and A. H. Hasty, "Immune cell activation in obesity and cardiovascular disease," Curr. Hypertens. Rep., vol. 24, no. 12, pp. 627–637, 2022.

[10] B. Şahin and G. İlgün, "Risk factors of deaths related to cardiovascular diseases in World Health Organization (WHO) member countries," Health Soc. Care Community, vol. 30, no. 1, pp. 73–80, 2022.

[11] A. Gupta, R. Kumar, H. Singh Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," IEEE Access, vol. 8, pp. 14659–14674, 2020.

[12] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," NPJ Digit. Med., vol. 3, p. 17, 2020.

[13] M. Saw, T. Saxena, S. Kaithwas, R. Yadav, and N. Lal, "Estimation of prediction for getting heart disease using logistic regression model of machine learning," in 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020.

[14] S. Ahuja and N. Masih, "Application of data mining techniques for early detection of heart diseases using Framingham heart study dataset," Int. J. Biomed. Eng. Technol., vol. 38, no. 4, p. 334, 2022.

[15] N. Townsend et al., "Epidemiology of cardiovascular disease in Europe," Nat. Rev. Cardiol., vol. 19, no. 2, pp. 133–143, 2022.

[16] M. Amini, F. Zayeri, and M. Salehi, "Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017," BMC Public Health, vol. 21, no. 1, 2021.

[17] G. L. Simegn, W. B. Gebeyehu, and M. Z. Degu, "Computer-aided decision support system for diagnosis of heart diseases," Res. Rep. Clin. Cardiol., vol. 13, pp. 39–54, 2022.

[18] J. N. Jothi, S.Poongodi, V.Chinnammal, L.Kannagi, M.Panneerselvam, and R. T. Prabu, "AI based humanoid chatbot for medical application," in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), 2022.

[19] N. H. Harlapur and V. Handur, "Text-based prediction of heart disease doctor chatbot using machine learning," in ICT with Intelligent Applications, Singapore: Springer Nature Singapore, 2023, pp. 231–245.

[20] A. Iftikhar et al., "Risk classification in global software development using a machine learning approach: A result comparison of Support Vector Machine and K-nearest neighbor algorithms," J. Inf. Technol. Res., vol. 15, no. 1, pp. 1–21, 2022.

[21] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," Artif. Intell. Med., vol. 128, no. 102289, p. 102289, 2022.

[22] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," Mob. Inf. Syst., vol. 2022, pp. 1–9, 2022.

[23] N. Absar et al., "The efficacy of machine-learning-supported smart system for heart disease prediction," Healthcare (Basel), vol. 10, no. 6, p. 1137, 2022.

[24] N. Alotaibi and M. Alzahrani, "Comparative analysis of machine learning algorithms and data mining techniques for predicting the existence of heart disease," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 7, 2022.

[25] G. N. Ahmad et al., "Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features," Appl. Sci. (Basel), vol. 12, no. 15, p. 7449, 2022.

[26] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, "Enhanced heart disease prediction based on machine learning and 2 statistical optimal feature selection model," Designs, vol. 6, no. 5, p. 87, 2022.

[27] K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a heart disease risk prediction model using machine learning," Comput. Math. Methods Med., vol. 2022, p. 6517716, 2022.

[28] M. Thorat and S. Munot, "Review on heart disease prediction using machine learning algorithms," Ijrpr.com. [Online]. Available: https://ijrpr.com/uploads/V3ISSUE11/IJRPR8196.pdf.

[29] A. Al Bataineh and S. Manacek, "MLP-PSO hybrid algorithm for heart disease prediction," J. Pers. Med., vol. 12, no. 8, p. 1208, 2022.

[30] D. S. Nerkar, "Non-invasive detection of coronary heart disease using machine learning," Ijrpr.com. [Online]. Available: https://ijrpr.com/uploads/V3ISSUE9/IJRPR6947.pdf.

[31] S. Madhumalar and S. Sivakumar, "A study on prediction of diabetic coronary heart disease using machine learning algorithms," Journal of IoT in Social, Mobile, Analytics, and Cloud, vol. 4, no. 2, pp. 119–132, 2022.

[32] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-xgboost algorithm," Information (Basel), vol. 13, no. 10, p. 475, 2022.

[33] B. A. M. Metwally, N. E. Mekky, and I. M. Elhenawy, "Heart disease prediction using genetic algorithm with machine learning classifiers," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 5, p. 1887, 2022.

[34] D.E. M. Nisar, R. Amin, N.-U.-H. Shah, M. A. A. Ghamdi, S. H. Almotiri, and M. Alruily, "Healthcare techniques through deep learning: Issues, challenges and opportunities," IEEE Access, vol. 9, pp. 98523–98541, 2021.

[35] L. Yahaya, N. David Oye, and E. Joshua Garba, "A comprehensive review on heart disease prediction using

data mining and machine learning techniques," American Journal of Artificial Intelligence, vol. 4, no. 1, p. 20, 2020.
[36] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," arXiv [cs.CY], 2017.
[37] W. Sun"Using machine learning approach to identify and analyze high risks patients with heart disease," in 2022 International Conference on Biotechnology, Life Science and Medical Engineering (BLSME 2022), 2022.
[38] R. K. Jha et al., "Neural fuzzy hybrid rule-based inference system with test cases for prediction of heart attack probability," Math. Probl. Eng., vol. 2022, pp. 1–18, 2022.
[39] V. Lamba, T. Agarwal, A. Gupta, and J. S. Chitkara, "A review on role of machine learning models on coronary heart disease detection accuracy," Ijcrt.org. [Online]. Available: https://ijcrt.org/papers/IJCRT2202193.pdf. [Accessed: 27-Aug-2023].
[40] F. Fatima, A. Jaiswal, and N. Sachdeva, "Heart disease prediction using supervised classifiers," SSRN Electron. J., 2022.
[41] Forum Desai et al., "HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing," Internet of Things, vol. 17, no. 100485, p. 100485, 2022.
[42] A. Avigan, "Cleveland clinic heart disease dataset."https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset.