Early Prediction of Coronary Heart Disease Using the Boruta Method

Vaibhav Satija[1(✉)], Mohaneesh Raj Pradhan[2], and Princy Randhawa[3]

[1] Department of Computer Science and Engineering, Manipal University Jaipur, Rajasthan 303007, India
vaibhav.209301028@muj.manipal.edu
[2] Department of Information Technology Engineering, Manipal University Jaipur, Rajasthan 303007, India
mohaneesh.209302365@muj.manipal.edu
[3] Department of Mechatronics Engineering, Manipal University Jaipur, Rajasthan 303007, India
princy.randhawa@jaipur.manipal.edu

Abstract. This paper discusses the application of machine learning in the healthcare sector for the prediction of heart disease. Because technology is a valuable tool in the healthcare industry, we intend to discuss the development of a machine learning model that measures many health-related characteristics in this study. The algorithm described in this research might identify whether a person is at risk of developing chronic heart disease during the following ten years. After balancing the unbalanced dataset and feature selection, the accuracy attained was 83–84% using various models such as Logistic Regression, Random Forest Classifiers, and Linear Discriminant Analysis. The paper focuses on analyzing a varied and diverse dataset whereas other papers referenced and cited have drawbacks such as the size of the dataset being too small or geographically limited. These anomalies have been kept in consideration while working on this paper.

Keywords: Healthcare · Heart Failure Prediction · Machine Learning · Algorithms · Chronic Heart Disease · Boruta · Synthetic Minority Oversampling Technique

1 Introduction

Machine learning is a branch of computer science coupled with Artificial Intelligence which employs the use of data and algorithms. This enables the software applications to predict outcomes more accurately without any explicit pre-programming. Machine Learning is now being used a lot in the Healthcare sector in numerous ways to depict how it can be used for various purposes ranging from Identification of diseases, diagnosis to drug discovery etc. Now doctors can easily monitor the health of their patients and use algorithms to derive results on the basis of their reports [1]. These not only have provided better and accurate treatment for the patients but have also made the patients more conscious of their health. Here, a machine-learning model has been designed

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 470, pp. 119–129, 2023.
https://doi.org/10.1007/978-3-031-35078-8_11

wherein various health related parameters have been measured. The model mentioned in this paper could then be used to predict if a person is prone to getting affected with a Chronic heart disease within the next 10 years.

1.1 Requirement of ML in Healthcare

Over the last few years' healthcare space is filling up quickly with progressive technological advancements due to the increasing challenges in healthcare [2]. The rising difficulties have forced the hospitals to equip themselves with modern innovations and incorporating them with machine learning and data analytics. ML has become one of the most important tools in the healthcare sector because of the factors defined below -

- Better visualization of patient's medical condition in terms of a progression or "stages" using error prone analysis.
- Medical data framework that helps hospitals and doctors in making accurate decisions.

It has also been claimed that widespread use of machine learning in health care will be one of the most essential life-saving technologies ever introduced.

1.2 Improvement Using Machine Learning

Machine Learning is defined as a part of artificial intelligence that allows a system to learn and improve automatically without being explicitly programmed. In the healthcare industry, it has been used for medical diagnosis, identification, prediction, and much more. The data collected from the datasets can be put through various machine leaning algorithms to predict future health risks an individual might face [3]. This would help people take proper preventive measures and lifestyle changes to reduce risk. Once enough datasets are run through different algorithms, and an algorithm with suitable accuracy is found, the feature of 'Future Health Risks' can be added to these smart devices.

2 Literature Review

Most of the papers that were analyzed were focused on Heart failure prediction [4] using machine learning, some of which also included prediction of the readmission of the patient and death of the patient after the treatment. Various machine learning techniques and algorithms like logistic regression neural networks support vector machines (SVM), random forest, fuzzy genetics, random forest, LASSO (Least Absolute Shrinkage and Selection Operator), cox algorithms etc. were used. Since the data was collected from various sources, data filtering, data cleaning and balancing the dataset played an instrumental role.

Table 1 is the detailed review of all the papers analyzed while working on this project.In the different papers that were analysed for the proposed study, it was found that various machine learning algorithms and techniques were used with feature selection methods to achieve desired results. Feature selection methods are also essential to deal with these high dimensional tasks. Standard feature selection methods, and novel

Table 1. Overview of the reviewed sources for different algorithms

Reference	Key Findings	Algorithm	Accuracy	Drawbacks
G. Hripcsak et al. [1]	The CART model was proved to be much precise in getting desired results	• SVM • Random Forest • Fuzzy genetic • CART (Classification And Regression Trees)	• SVM-85.2% • Random Forest-85.6% • Fuzzy genetic-85.9% • CART-87.6%	Amount of sample data taken for training and testing was small
K. Dickstein et al., 2008 [2]	ML and feature selection methods improve heart rate prediction	• mRMR (Minimum Redundancy and Maximum Relevance) • Neural Networks	NA	Data available was only of patients who were 65 years or above
P Melilo et al., 2002 [3]	L2 regularized SVM using RBF kernel was proved effective in heart prediction	• SVM	• SVM without RBF-91.11% • SVM with RBF-92.22%	NA
B. H. Greenberg et al., 2019 [4]	LSTM model, with or without window sliding, provides better accuracy and performance as compared to the KNN, logistic regression, SVM and MLP models	• LSTM (- Long short-term memory) • KNN (K-Nearest Neighbour) • Logistic Regression • SVM • MLP (Multilayer perceptron)	LSTM model, with or without window sliding provides better accuracy	NA
L. Ali et al., 2009 [5]	MLP model predicted that 23.6% of HF patients were readmitted or died within 30 days	• MLP • Logistic Regression • SVM • Random Forest • Decision tree	NA	Some clinical data was not accessible due to hospital restrictions Patients under 65 years of age were excluded

(*continued*)

Table 1. (*continued*)

Reference	Key Findings	Algorithm	Accuracy	Drawbacks
R Das et al., 2009 [6]	The GLMN Machine Learning Model had the best performance in predicting hospitalizations	• LR(Logistic Regression) • GLMN (Generalized Linear Models) • SVM • Neural Network • Random Forest • CART	The GLMN model had the best accuracy	NA
G. Maragatham et al., 2019 [7]	SVM with RBF achieved the best accuracy of heart prediction	• SVM with linear • SVM with polynomial • SVM with RBF (Radial Basis Function Kernel) • SVM with sigmoid	SVM with RBF gave the best accuracy	There were less variations in the dataset so the model might not have been properly trained
E Choi et al., 2017 [8]	Various features such as ethnicity, gender and age among others were used to predict heart failure and discriminate high risk cases from low-risk cases	• A boosted decision tree algorithm	Area under the curve (AUC) of 0.88	Demographically biased as the patient chose data from the same region

(*continued*)

Table 1. (*continued*)

Reference	Key Findings	Algorithm	Accuracy	Drawbacks
K. Polat et al., 2007 [9]	Different machine learning techniques were used to classify heart failure subtypes Flexible modern tree-based algorithms had better performance than the conventional ML techniques	• Regression • Trees • Bagging • Random forest • Boosting SVM • Logistic regression	Accuracy measured by AUC was worst for regression tree and best for logistic regression classification and Random Forest has the best performance. Boosting and SVM had the poorest performance when compared to the rest of the methods	NA
S.E. Awan et al., 2019 [10]	Predicting the rate of HF readmission or death within 90 days using ML techniques 15% patients were readmitted because of acute HF and 11.5% died within 90 days after discharge	• K-Fold cross validation with SVM • RF • Gradient Boosting Machine • LASSO	The LASSO and the gradient boosting machine models performed better than other models with an AUC of 0.748 and 0.750	Single centre-based study with less data records Missing data Application of ML techniques on patients without baseline echocardiography was not testes or validated

feature selection methods were also developed for handling high dimensional imbalanced datasets. Some results that were received displayed high accuracy, however, there were some drawbacks such as the size of the dataset being too small or geographically limited. These anomalies have been kept in consideration while working on this paper.

3 Methodology

A dataset with over 4,240 records, 16 columns and 15 attributes were chosen. The methodology followed for the model development is mentioned below.

3.1 Sorting Out Data for Missing and Duplicate Entries

A very instrumental part of the data pre-processing involves the use of requisite tools to remove redundancies in the table which can hamper the overall calculation of accuracy of data for the deployed model. So, initially a check for finding duplicate and missing entries was conducted. The overall results concluded no duplication of the data, but some of the data showed missing values.

3.2 Dataset Analysis

The analysis of the present dataset helps to determine the contribution of each of the parameters leading to the risk of the occurrence of Chronic Heart Disease in the coming decade. For this, a simple bar chart is generated with the relevant features and parameters that helps compare the number of people facing the risk of contracting a CHD and those without the risk of a CHD.

3.3 Selecting Relevant Features

Out of the 15 attributes present in the database, not all attributes change the final output on being changed. Such attributes could be removed for better training of the model. The BORUTA feature selection algorithm is applied for the selection of relevant features for the model training. BORUTA maps the level of importance of a feature in a graph and then chooses the features in which the graph is shaped like a bell curve. This portion of the graph is also called the Acceptance region and the region where the rejected features lie is known as the Rejection region. Figure 1 shows an example of BORUTA's feature mapping curve.

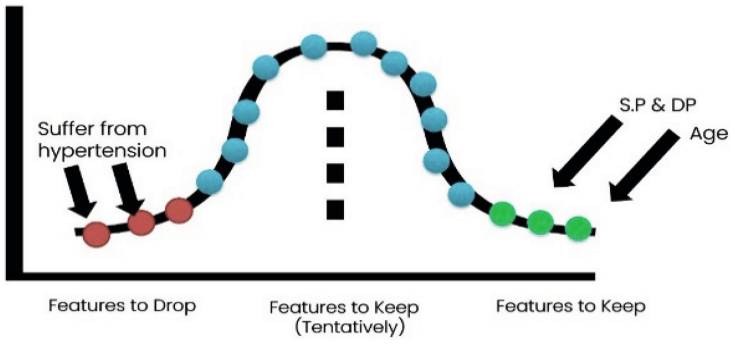

Fig. 1. An illustration depicting the feature selection curve with the acceptance and rejection region mentioning different features to be considered for prediction

After using Boruta on the dataset with 100 iterations, the z-score of each parameter was calculated and the top parameters were selected. In this case, the top parameters having the most effect on the result were 3 out of 15. These parameters were:

- Age
- SysBP
- BMI

Model deployment after this was done using only these 3 most relevant features.

4 Model Deployment

This step involves development of the model. Moving on to the training of the model by performing train test split data where around 90% of the model is allocated to train the data and the remaining 10% of the data is employed to test the model. Before using the Boruta feature classification on the modified data, LDA (Linear Discriminant Analysis),

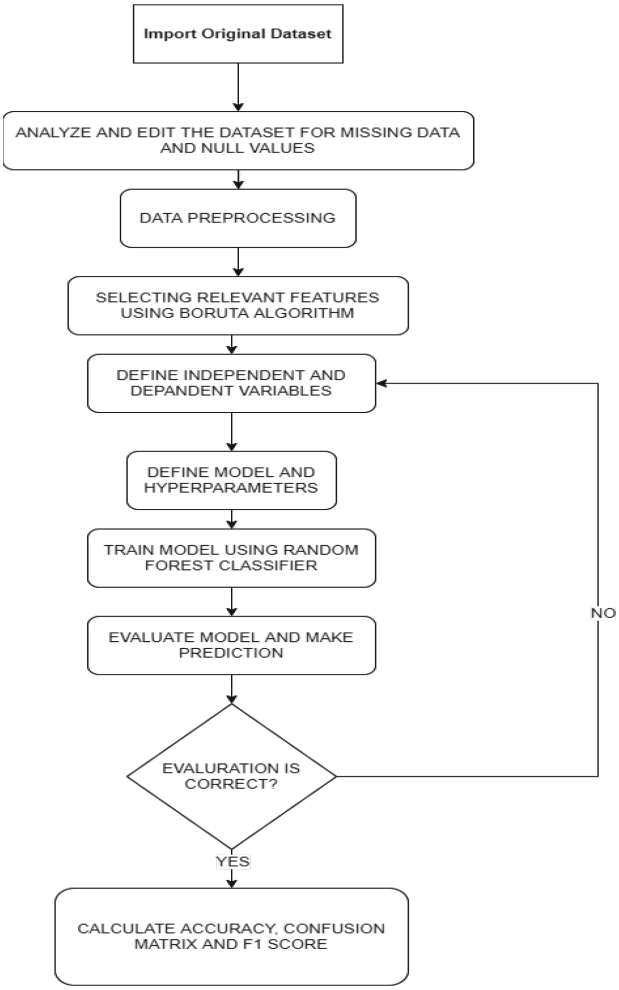

Fig. 2. A figure containing a flowchart of all the steps used in the model

Logical Regression and Random Forest Classifier techniques are used on the given data to calculate the accuracy and confusion matrix for each of the techniques After the calculation of the respective confusion matrices and accuracies for each of the techniques, the Boruta feature selection criteria is used to select features which are most useful and optimal for the process. Boruta, being a random forest base method, works best for tree models like Random Forest or XGBoost, but is also synchronous with other classification models like Logistic Regression or SVM (Support Vector Machines). The RFC (Random Forest Classifier) method is used for the classification of data. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Figure 2 below shows the flowchart of all the steps mentioned above.

After the completion of the model training, the test-train split was utilized to test the model by using the testing distribution set. The results of the test set can be summarized with the confusion matrix shown below in Fig. 3.

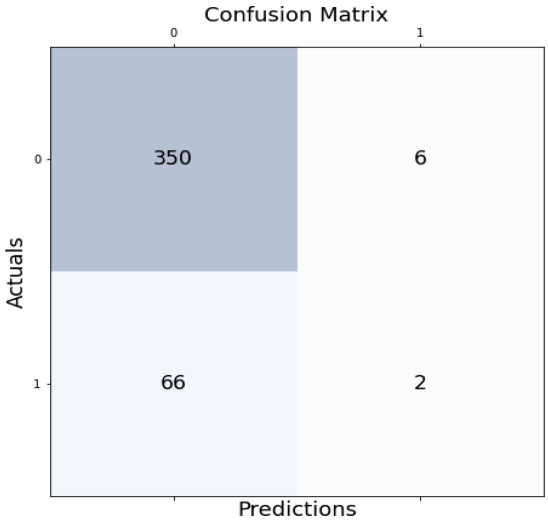

Fig. 3. A confusion matrix illustration of the test-distribution set

The following confusion matrix displays the following results

350 – True Positive (number of positive examples classified accurately)
6 – True Negative (number of negative examples classified accurately)
66 – False Positive (number of actual negative examples classified as positive)
2 – False Negative (number of actual positive examples classified as negative)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Using the above equations, it was found that the recall of the confusion matrix is 0.99 and the precision is 0.84.

5 Scope of Improvement

The scope for improvement in any model is limitless as there always exists something that can be improved within the existing model. The machine learning model created to detect the risk of Chronic heart disease can be improved by using a better dataset to train the model. The dataset should include a larger range of parameters that can influence the risk of Chronic heart disease to a person, some parameters that can be included is the level of lipoprotein, does the person exercise daily, as these parameters also majorly affect the chances of one suffering from a chronic heart disease [20]. Additionally, a dataset with a lower percentage of blank entries should be used. Using the dataset with the modifications will drastically improve the accuracy of the model. Moreover, the scope of the model should not be restricted to the prediction of Chronic Heart Disease, it can be used to predict various other medical conditions. This prediction model with more efforts and research can lead to the addition of new features in existing healthcare facilities such as prediction of an individual's future health risks.

6 Conclusion

This work presents an overview of how various Machine Learning models can be used in the healthcare sector. Here a machine learning model was created that predicts the likelihood of a person getting a Chronic heart disease, using the data collected. An accuracy of around 83–84% was achieved with the 'Framingham Heart Disease' dataset. However, in comparison to the references cited above, the dataset contained much more values and was geographically and locally unbiased. Thus, with machine learning and the different parameters, it is possible to monitor whether the individual faces any health risks such as CHD in the future. It has been observed that the use of Machine Learning has accentuated over the past few years in the healthcare sector which is expanding in this domain and has resulted in various advantages like reduction in healthcare costs, remote patient monitoring, medical data accessibility, improved treatment management, reduced errors, better patient experience etc.

Regarding this study, it can be depicted that there are a plenty of healthcare advancement techniques that can be used by patients suffering from minor illness to major illness which make them readily available for different categories of people like doctors, patients, patient's families, insurance companies and hospital management for their respective usage. Due to the resurgences of COVID-19 it is sure that there will be a greater impact of technological advancements in healthcare for years to come. Additionally, the data stored and collected from various analysis reports can be further analysed using various machine learning techniques and this could help the heath-care professionals detect various diseases at an early stage, which could prove to be a great boon in the treatment of the patient. In conclusion, Machine learning algorithms can lead to the creation of superior healthcare devices which can be of immense benefit to the healthcare industry.

References

1. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. **20**(1), 117–121 (2013). https://doi.org/10.1136/amiajnl-2012-001145
2. Dickstein, K., et al.: ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008. Eur. Heart J. **29**(19), 2388–2442 (2008). https://doi.org/10.1093/eurheartj/ehn309
3. Melillo, P., De Luca, N., Bracale, M., Pecchia, L.: Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE J. Biomed. Heal. Inform. **17**(3), 727–733 (2013). https://doi.org/10.1109/JBHI.2013.2244902
4. Greenberg, B.H.: Heart failure epidemic. Curr. Cardiol. Rep. **4**(3), 185 (2002). https://doi.org/10.1007/s11886-002-0048-y
5. Ali, L., et al.: An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access **7**, 54007–54014 (2019). https://doi.org/10.1109/ACCESS.2019.2909969
6. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Appl. **36**(4), 7675–7680 (2009). https://doi.org/10.1016/j.eswa.2008.09.013
7. Maragatham, G., Devi, S.: LSTM model for prediction of heart failure in big data. J. Med. Syst. **43**(5), 1–13 (2019). https://doi.org/10.1007/s10916-019-1243-3
8. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. **24**(2), 361–370 (2017). https://doi.org/10.1093/jamia/ocw112
9. Polat, K., Şahan, S., Güneş, S.: Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and K-NN (nearest neighbour) based weighting preprocessing. Expert Syst. Appl. **32**(2), 625–631 (2007). https://doi.org/10.1016/j.eswa.2006.01.027
10. Awan, S.E., Bennamoun, M., Sohel, F., Sanfilippo, F.M., Dwivedi, G.: Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Failure **6**(2), 428–435 (2019). https://doi.org/10.1002/ehf2.12419
11. Lorenzoni, G., et al.: Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. J. Clin. Med. **8**(9), 1298 (2019). https://doi.org/10.3390/jcm8091298
12. Bianchi, F.M., De Santis, E., Rizzi, A., Sadeghian, A.: Short-term electric load forecasting using echo state networks and PCA decomposition. IEEE Access **3**, 1931–1943 (2015). https://doi.org/10.1109/ACCESS.2015.2485943
13. Son, Y.J., Kim, H.G., Kim, E.H., Choi, S., Lee, S.K.: Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc. Inform. Res. **16**(4), 253–259 (2010). https://doi.org/10.4258/hir.2010.16.4.253
14. Adler, E.D., et al.: Improving risk prediction in heart failure using machine learning. Eur. J. Heart Failure **22**(1), 139–147 (2020). https://doi.org/10.1002/ejhf.1628
15. Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., Lee, D.S.: 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. **66**(4), 398–407 (2013). https://doi.org/10.1016/j.jclinepi.2012.11.008
16. Sarijaloo, F.B., Park, J., Zhong, X., Wokhlu, A.: Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. Clin. Cardiol. **44**(2), 230–237 (2021). https://doi.org/10.1002/clc.23532

17. Gupta, A., Fonarow, G.C.: The Hospital Readmissions Reduction Program – learning from failure of a healthcare policy. Eur. J. Heart Fail. **20**(8), 1169–1174 (2018). https://doi.org/10.1002/ejhf.1212
18. ESC Heart Failure – 2021 – K nig – Machine learning algorithms for claims data-based prediction of in-hospital mortality.pdf.
19. Segar, M.W., et al.: Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure: a multicohort analysis. Circulation **143**(24), 2370–2383 (2021). https://doi.org/10.1161/CIRCULATIONAHA.120.053134
20. Guo, A., Pasque, M., Loh, F., Mann, D.L., Payne, P.R.O.: Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models. Curr. Epidemiol. Rep. **7**(4), 212–219 (2020). https://doi.org/10.1007/s40471-020-00259-w