HYPER PARAMETER TUNED ENSEMBLE APPROACH FOR HEART DISEASE PREDICTION

Premisha Premananthan[1], Senthan Prasanth[2] and Kanagarathnam Mauran[1]

[1]Department of ICT, Faculty of Technological Studies, University of Vavuniya
[2]Sabaragamuwa University of Sri Lanka

Abstract

Heart disease is the one of the leading causes of death globally. Despite the fact that the causes of heart disease varied from nation to nation. However, the risk variables would be practically the same. Heart disease refers to any condition that affects the cardiovascular system. Heart disease manifests itself in a variety of ways, each of which affects the heart and blood arteries differently. Predicting the prognosis of cardiovascular diseases on early stages can assist high-risk individuals to adopt lifestyle changes and, as a result, prevent repercussions. The goal of this study is to identify the most important risk factors that influence heart disease and to detect the possibility of having heart disease in advance. The information required for this study was gathered from ongoing cardiovascular studies on the inhabitants of Framingham, Massachusetts. The prediction model development is to determine if the patient has a 10-year risk of developing coronary heart disease (CHD). The dataset has about 4,000 records with 15 parameters. Initially, the data was fed into supervised machine learning approaches like Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes (NB),Linear Discriminant Analysis (LDA),Logistic Regression (LR) and k- nearest neighbors (k-NN). In addition, bagging and boosting techniques like Random Forest (RF), CatBoost, , LightGBM, and Extreme Gradient Boosting (XGBoost) also incorporated. Furthermore, the final ensemble model was built by adapting the algorithms with good performance namely CatBoost,Random Forest, and Logistic Regression algorithms to predict the risk of heart disease. Final ensemble model resulted in an accuracy of 86.20%.

Keywords*: heart disease, machine learning, boosting algorithms, ensemble approach*

1 INTRODUCTION

Heart disease is a serious condition that has a negative impact on all societies and causes long-term misery and incapacity. The majority of the impacts of heart disease have increased over the world as a result of the Covid issue. Cardiovascular disease is an increasing socioeconomic and public health issue, with high mortality and disability rates. Annually, according to the World Health Organization, cardiac disorders affect 12 million people around the world. In the United States and other industrialized nations, cardiovascular diseases are the major cause of death, accounting for half of all fatalities. The capacity to anticipate the early identification of cardiovascular diseases can assist high-risk patients adopt lifestyle changes, allowing them to avoid serious repercussions. According to a World

Health Organization (WHO) estimate, cardiovascular diseases (CVDs) caused 17.9 million deaths in 2019, accounting for 32 percent of all global deaths [1] and an annual mortality rate of more than 17.7 million [2]. Cardiovascular disease (CVD) was the main cause of death in Australia in 2018, according to the Australian Institute of Health and Welfare (AIHW), accounting for 42 percent of all deaths [3]. Existing heart disease diagnosis methods are poor in early detection for a variety of reasons, including accuracy and computing time [4]. Researchers are working to develop an effective strategy for the timely identification of heart disorders. Here, a procedure called "Predictive Analysis" is performed, which is linked to many components of machine learning algorithms, statistical approaches and data mining techniques that use current and historical data to identify crucial facts to forecast future events. Important conclusions and excellent heart disease forecasting could be reached using predictive analysis on healthcare statistics. Machine learning approaches, both supervised and unsupervised, could be used to perform "predictive analytics." This analytics is aimed at predicting the risk of heart disease with the greatest precision feasible, improving patient health care, and enhancing resources while achieving progressive medical results [5]. Cardiac disease is a wide term that refers to a range of cardiac conditions. Coronary artery disease is the most prevalent kind and can lead to a heart attack. Heart and Stroke failure are two others.

A cardiovascular disease (CVD) risk assessment [6], also known as a heart disease risk assessment, is a sort of screening tool that determines your risk of heart disease or CVD. Heart disease is a kind of cardiovascular disease (CVD), which is a set of heart and blood vessel illnesses. A risk assessment for heart disease contains a series of questions concerning variables such as family history, age, and lifestyle choices such as food and exercise. The chance of acquiring heart disease in the future is then calculated. Even though you currently feel healthy, the evaluation can reveal if you need to take steps to avoid or lower your risk of getting heart disease in the future. Instead, the focus of this research study is on developing an effective forecasting model that uses machine learning algorithms and data mining methodologies to properly predict heart disease. Health authorities all across the world own large volume databases. These databases are likely to have structured, semi-structured, or unstructured data in them. "Big Data Analytics" refers to the process of analyzing extremely large data sets and uncovering hidden patterns and information in order to extract key details from the data. If and when the present scenario is considered.

The major goal of this research is to develop a model that can reliably forecast a person's risk of heart disease over the next ten years. There were studies which used the same dataset but got lesser accuracy as their final prediction accuracy. This particular study got relatively highest accuracy with basic classifier algorithms. In this study's proposed methodology only considered with Decision Tree, e Logistic regression, K-nearest neighbor, Support vector machine, and Random Forest. [3].

This research content has been modelled as follows: Section II discussed about the methodology followed to achieve the theme. Next, section IV represents the findings of the study. Finally, the paper has been ending with the conclusion and references.

2 METHODOLOGY

Various strategies were employed in this process to finalize the development of the ensemble forecasting model. Figure 1 depicts the process. Each phase of the implementation process is described in detail below.

2.1 The Dataset description

The data originates from a continuous cardiovascular research carried out by the scholars from Framingham, Massachusetts and to obtain the data the people who are resided in Framingham were considered. The data has been gathered from the Kaggle website. The dataset contains information on patients. There are about 4,000 records and 15 different characteristics in all. Every feature has the potential to be a danger factor. Demographic, behavioral, and medical difficulties are all risk factors.

- Demographic:
  - Sex: either male or female; (Nominal) age: the patient's age; (Continuous - Notion of age is continuous, even though the ages recorded have been shortened to whole numbers.)
- Behavioural
  - current Smoker: If the patient is a current smoker (Nominal)
  - cigsPerDay: If true, the average number of cigarettes smoked each day by the individual. (Can be called continuous since any number of smokes, even half a cigarette, can be consumed.)
- Medical (history):
  - BPMeds: whether the patient was taking blood pressure medication (Nominal)
  - prevalentStroke: whether the patient had already experienced a stroke (Nominal)
  - prevalentHyp: if the patient was hypertensive or not (Nominal)
  - heart disease: whether or if the patient suffered cardiovascular illness (Nominal)' 'it)
- Medical (current):
  - heartRate: heart rate (Continuous - Parameters such as heart rate are considered continuous in medical research despite the fact that they are discrete owing to the huge number of possible values.)
  - sysBP: blood pressure systolic (Continuous)
  - BMI: Body Mass Index (Continuous)
  - glucose: glucose level (Continuous)
  - diaBP: blood pressure during diastole (Continuous)
  - totChol: total cholesterol concentration (Continuous)
- Forecast variable (desired target):
  - 10 year hazard of cardiovascular disease CHD (binary: "1", means "Yes", "0" means "No")

2.2 Data Pre-processing

Data preprocessing is any type of processing done on raw data to prepare it for future processing. It was previously employed as a more advanced stage in the data mining process. For both training and inferring against machine learning and AI models, several strategies have recently evolved. These methods can be applied to a wide range of data sources, including data saved in files or databases as well as data provided by streaming data systems. This is the most critical phase in establishing a forecasting model to assess a patient's risk of cardiovascular disease. The "Python Pandas Library" was employed to implement the technique. The following tasks were done as part of the precedent procedure. Initially, a Python project was created with the PyCharm IDE. Following that, the needs machine learning and third-party libraries were loaded into the project directory. Following that, the acquired dataset was saved within the project folder. Pandas was used to import the dataset.

- Countering for the distribution of dataset along with the amount of columns and rows
- Keeping the dataset within a certain range. The ranges of features may have different values. Normalization was used to keep all of the feature values of a process inside a defined range.
  - The Min-Max Scaling normalization strategy was employed in this study.
  - Normalized Value = (Actual Value – Minimum Value) / (Maximum Value – Minimum Value)

2.3 Ensemble model implementation

Following the data preparation and interpretation phases, the dataset was separated into training and testing with a percentage of 80 and 20 correspondingly. The final ensemble model construction technique was separated into two phases, as indicated in Figure 1.

Those algorithms were used to transfer the whole dataset with all of its properties. The same dataset was put through Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes (NB),Linear Discriminant Analysis (LDA),Logistic Regression (LR) and k- nearest neighbors (k-NN), Bagging and boosting techniques like Random Forest (RF), CatBoost, , LightGBM, Extreme Gradient Boosting (XGBoost) and AdaBoost to find the best classifier based on fundamental assessment criteria.

Using the testing dataset, the trained classifiers were evaluated, and the best classifiers were chosen based on accuracy. The same dataset were again evaluated with defined percentatge of training testing data with hyper parameter tunning. Then K fold cross validation were done for the same data set with defined percentage pf training testing data set used to get Mean Accuracy. Furthermore, all classifier models were evaluated utilizing k-fold cross-validation with 5 folds, ensuring that every observation from the original dataset has a chance of doing well in both the testing and training sets. An analysis of their performance was used to choose the best classifiers with hyperparameter tuning after hyper parameter tuning, and from K-fold cross validation.

3 Results and Discussions

This chapter begins with a discussion of the classifiers' individual performance uncovered during the studies. This section will go over the findings for the individual classifiers, as well as hyper-parameter tuning and cross-validation, in addition to the prior outcomes. Finally, the individual classifier models and the ensemble model are compared in terms of performance. The basic evaluation values were

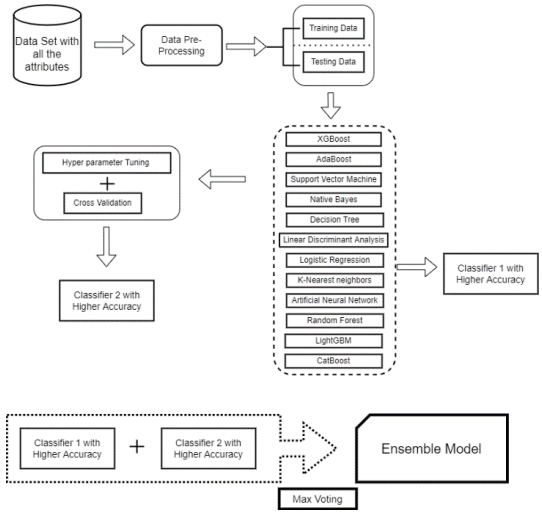

Figure 1. Major issues encountered by the tea processors

Table 1. Individual classifiers' performance before hyper parameter tuning

Classifier	Accuracy (%)	MSE (%)	Precision (%)	Recall (%)	F1- Score (%)
RF	79.22	20.77	77.61	61.17	68.42
XGBoost	76.62	23.37	71.83	60.00	65.38
LightGBM	76.62	23.37	71.23	61.17	65.82
AdaBoost	78.35	21.64	76.92	58.82	66.66
CatBoost	80.35	19.04	78.87	65.88	71.79
SVM	79.22	20.77	82.45	55.29	66.19
LDA	78.35	21.64	77.77	57.64	66.21
NB	78.35	21.64	74.64	62.35	67.94
LR	77.92	22.07	79.31	54.11	64.33
ANN	76.19	23.80	66.66	51.35	58.01
DT	70.13	29.87	61.76	49.41	54.90
k-NN	80.08	19.91	79.10	62.35	69.73

analyzed in Table I in order to identify the classifier with better performance. Based on the core assessment criteria and CatBoost was named the top classifier based on the above-mentioned data. Furthermore, the optimal values that may be supplied to the hyper-parameters of certain classifiers were identified using the GridSearchCV technique.In addition, the results of the assessments for each classifier are shown in that table. As a result of phase-1 of the approach, CatBoost was named the top classifier in terms of F1-Score value, recall, and Accuracy.

Based on the outcomes of individual classifiers (after hyper parameter tuning) depicted by the table II below, the K-fold cross-validation approach was used to select the second-best classifier with 5 folds. Finally, it was discovered and confirmed that RF outperformed all other classifiers in terms of accuracy (81.05%) and other evaluation metrics.

In addition, to develop the final ensemble model K-fold cross validation was performed with

Table 2. Individual classifiers' performance after hyper parameter tuning

Classifier	Accuracy (%)	MSE (%)	Precision (%)	Recall (%)	F1- Score (%)
RF	80.95	19.04	78.87	65.88	71.79
XGBoost	78.35	21.64	76.92	58.82	66.66
LightGBM	80.51	19.48	80.30	62.35	70.19
AdaBoost	79.22	20.77	77.61	61.17	68.42
CatBoost	79.65	20.34	78.78	61.17	68.87
SVM	78.35	21.64	76.92	58.82	66.66
LDA	77.92	22.07	71.25	67.05	69.09
NB	78.79	21.21	75.71	62.35	68.38
LR	78.35	21.64	80.70	54.11	64.78
ANN	77.05	22.94	66.66	56.75	61.31
DT	77.06	22.94	74.24	57.64	64.90
k-NN	76.19	23.80	73.43	55.29	63.08

Table 3. Performance of individual classifiers after hyper parameter tuning and cross-validation with 5 folds

	Accuracy attained using K-fold cross-validation with 5-fold					
Classifier	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Mean Accuracy
RF	74.67	68.18	77.27	84.31	75.81	77.73
XGBoost	75.32	68.83	77.27	81.04	77.12	75.39
LightGBM	77.27	70.77	77.27	79.73	76.47	77.08
AdaBoost	74.02	70.12	80.51	80.39	75.71	76.69
CatBoost	75.97	72.07	79.87	81.69	75.81	77.47
SVMLR	76.62	73.37	77.92	83.66	74.50	77.21
LRSVM	75.97	68.18	76.62	82.35	77.12	76.17
NB	75.32	72.72	75.32	80.39	74.50	75.39
ANN	68.25	63.17	60.47	54.56	58.83	61.05
DT	71.42	64.28	70.12	78.43	74.50	72.15
k-NN	73.37	66.88	74.67	79.73	75.81	74.22
LDA	75.32	63.63	73.37	76.47	75.16	72.79

afore mentioned classifiers after the hyper-parameter tuning. The final ensemble model was built after the result obtained from the individual classifiers for the cross validation. Based on the mean accuracy obtained for the five folds for the considered classifiers CatBoost, RF and LR produced the prominent results. The max vote technique was utilized to combine the aforementioned classifiers. As a consequence, the ensemble model built with the Max-voting approach attained an accuracy of 86.20 percent, which is much better than the performance of individual classifiers.

4 CONCLUSION

An ensemble model was implemeted in this work to predict risk of heart diseases in advance. The procedure of creating the final ensemble model was done in two steps. Several supervised machine learning techniques were employed to analyze the dataset during the early part of the process, including Support Vector Machine, Naive Base, Decision Tree, Logistic Regression Linear Discriminant Analysis, k-NN, Random Forest, ANN, XGBoost, AdaBoost, LightGBM, and CatBoost. Following that, during development part 2, RF achieved the best classification accuracy of 80.95 percent. Hyper parameter tuning was performed to get the optimum value for the accuracy to the considered classifiers. Moreover K-Fold cross validation with 5 folds also being performed with the dataset to build up the final ensemble model. From all the results discovered CatBoost, RF and LR were identified as the algorithms to produce the precise results. Finally to develop-p the ensemble model max voting technique was accommodated and thus model produced the classification results with an accuracy of 86.20%. This research would be a promising start for the future researchers to give more insights on focusing on the health sector by adapting the techniques from machine learning.

REFERENCES

[1] Sunitha Guruprasad, Valesh Levin Mathias, and Winslet Dcunha. "Heart Disease Prediction Using Machine Learning Techniques". In: *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*. IEEE. 2021, pp. 762–766.

[2] Amit Krishna Dwivedi, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. "Algorithms for automatic analysis and classification of heart sounds–a systematic review". In: *IEEE Access* 7 (2018), pp. 8316–8345.

[3] Emrana Kabir Hashi and Md Shahid Uz Zaman. "Developing a hyperparameter tuning based machine learning approach of heart disease prediction". In: *Journal of Applied Science & Process Engineering* 7.2 (2020), pp. 631–647.

[4] Geoffrey Rose. "Familial patterns in ischaemic heart disease". In: *British journal of preventive & social medicine* 18.2 (1964), p. 75.

[5] Syed Immamul Ansarullah, Syed Mohsin Saif, Syed Abdul Basit Andrabi, et al. "An Intelligent and Reliable Hyperparameter Optimization Machine Learning Model for Early Heart Disease Assessment Using Imperative Risk Attributes". In: *Journal of healthcare engineering* 2022 (2022).

[6] Jhabindra Khanal, Dae Young Lim, Hilal Tayara, et al. "i6ma-stack: a stacking ensemble-based computational prediction of dna n6-methyladenine (6ma) sites in the rosaceae genome". In: *Genomics* 113.1 (2021), pp. 582–592.