International Journal of Innovative Research in Technology and Science

ISSN: 2321-1156 www.ijirts.org Volume 12 Issue 2, March 2024

Prediction of Cardiovascular Disease Using Machine Learning

Bhoopendra Singh, Subodhini Gupta

Computer Science and Engineering Department

SAM Global University, Bhopal, India

Selection and peer review of this article are under the responsibility of the scientific committee of the International Conference on Current Trends in Engineering, Science, and Management (ICCSTEM-2024) at SAM Global University, Bhopal.

Abstract - Cardiovascular disease (CVD) poses a significant threat to human health by impairing the functionality of the heart and blood vessels, often resulting in death or physical paralysis. Early and automated detection of CVD is crucial for saving lives. While numerous efforts have been made towards this goal, there remains scope for enhancing performance and reliability. This study contributes to this ongoing endeavour by employing two robust machine learning techniques, multilayer perceptron (MLP) and K-nearest neighbour (K-NN), for CVD detection, utilising publicly available data from the University of California Irvine repository. The models' performances are optimally enhanced by removing outliers and attributes with null values. Experimental results showcase a superior accuracy of 82.47% and an area-under-the-curve value of 86.41% achieved by the MLP model, outperforming the K-NN model. Consequently, the proposed MLP model is recommended for automatic CVD detection. Furthermore, the methodology presented herein holds promise for detecting other diseases, and the performance of the proposed model can be validated across additional standard datasets.

Keywords: cardiovascular disease, machine learning algorithms, K-nearest neighbour, multilayer perceptron

1. INTRODUCTION

Health is a crucial part of everyone's life. However, owing to multiple reasons like unhealthy lifestyles, work stress, psychological strain, and external factors such as pollution, hazardous work environment, and lack of proper health services, millions of people worldwide fall prey to chronic ailments like cardiovascular diseases (CVD), which affect both the heart and blood vessels, resulting in death or disability. In recent years, it was reported that the majority of human deaths were due to CVD [1, 2]. The associated conditions are hyperte9ik,m, Snsion, thromboembolism, hyperlipidaemia, and coronary heart disease, which culminate in heart failure. Hypertension is the primary cause of CVD [3]. In 2012, 7.4 million people were reported to have died from coronary heart disease, while 6.7 million people died from stroke [4]. The World Health Organization estimates that nearly 17 million people die every year from CVDs, which accounts for approximately 31% of global deaths. Early diagnosis of CVD can potentially cure patients and save innumerable lives.

Diagnosis and treatment of patients at early stages by cardiologists remain a challenge. Every traditional CVD risk-assessment model implicitly assumes each risk factor related to CVD outcome linearly. Such models tend to oversimplify

complex relationships, including several risk factors with nonlinear interactions. Multiple risk factors should be properly incorporated, and more correlated nuances between the risk factors and outcomes should be determined. No large-scale study has used routine clinical data and machine learning (ML) in prognostic CVD assessment. This study aims to determine if ML can enhance cardiovascular risk prediction accuracy in population primary care and which ML algorithm result had fairly high brevity. In recent years, multiple ML-based CVD detection models have been proposed. A review of previous studies is presented to identify each study's research problem and objective. ML helps a cardiologist to predict diseases at an early stage and treat the patient accordingly. There are many ML techniques, such as support vector machines [5], artificial neural networks, decision trees [6], and K-Nearest Neighbour (K-NN) [7], each with its strengths and weaknesses. These methods have been applied in broader areas like predicting liver [8, 9], human heart (echocardiogram signals) [10, 11], and skin diseases [12, 13, 14]. The results of each technique differ owing to several constraints. Observations from related studies reveal further scope for developing automated CVD detection using other ML models that provide improved performance. This study contains an in-depth statistical analysis of input data sets to understand the effects of data range on CVD predictions. It includes a correlation study of categorical and continuous features of patients. In addition, data visualisation and scatter plots for pairs of important features were obtained to understand the significance of the correlation between important features. These are discussed and analysed in the results section.

2 METHODOLOGY

This study aimed at the confusion matrix of each technique, and out of 303 occurrences in the dataset, 243 (80%) were used to train the two models. In order to test the trained models, 60 instances are fed to know the class. This study intends to predict the likelihood of developing CVD via a computerised prediction route that can be useful to health professionals. The materials required for CVD detection are the test data of patients from publicly available standard CVD data from the UCI repository [15]. The classification algorithms used are MLP and K-NN. Generally, the method comprises training the proposed model via respective learning algorithms using relevant input test data of patients and then validating these models based on test data of patients. Finally, performance measurements are evaluated and compared. The following steps are carried out to predict CVD:

Step 1. The relevant CVD data set is first collected from the UCI repository.

Step 2. Data samples are preprocessed by eliminating null values, filtering for demonising, and removing outliers present in samples.

Step 3. Attributes more useful in CVD forecasting are selected, and strongly correlated features are dropped.

Step 4. Two ML algorithms that are simple but effective are chosen to classify the selected features.

Step 5. Various performance measures are evaluated to compare and find the better method.

This study aims to predict the probability of heart disease through computerised heart disease prediction, which can benefit medical

professionals and patients. We employed various machine learning algorithms on a dataset to achieve this objective and present the results in this study report. To enhance the methodology, we plan to clean the data, eliminate irrelevant information, and incorporate additional features such as MAP and BMI. Next, we will separate the gender-based dataset and implement K-mode clustering. Finally, we will train the model with the processed data. The improved methodology will produce more accurate results and superior model performance, as demonstrated in Data Source. Clustering is a machine learning technique where a group of instances is grouped based on similarity measures. One common algorithm used for clustering is the k-means algorithm, but it is ineffective when working with categorical data. In order to overcome this limitation, the k-modes algorithm was developed. The k-modes algorithm, introduced by Huang [29] in 1997, is similar to the k-means algorithm but utilises dissimilarity measures for categorical data and replaces the means of the clusters with modes. This allows the algorithm to work effectively with categorical data. Since our data have been converted to categorical data, we will use k-mode analysis. We will first use the elbow curve with Huang initialisation to find the optimal number of clusters. An elbow curve creates a k-mode model with that number of clusters, fits the model to the data, and then calculates the cost (distance between the attribute modes of each cluster and the data points assigned to the cluster). The costs are then plotted on a graph using the "elbow method" to determine the optimal number of clusters. The elbow method looks for a "knee" or inflexion point in the plot of costs, which is often interpreted as the point where adding more clusters does not significantly improve the model's fit. Splitting the dataset based on gender can be advantageous for prediction due to the existence of significant biological disparities between men and women that can impact the manifestation and progression of diseases. For instance, men tend to develop heart disease at an earlier age than women, and their symptoms and risk factors may differ. Studies have shown that men have a higher risk of coronary artery disease (CAD) compared with women and that the CAD risk factors and presentations may differ between the sexes [30]. By analysing the data separately for men and women, it is possible to identify unique risk factors and patterns of disease progression that may not be discernible when the data are consolidated.

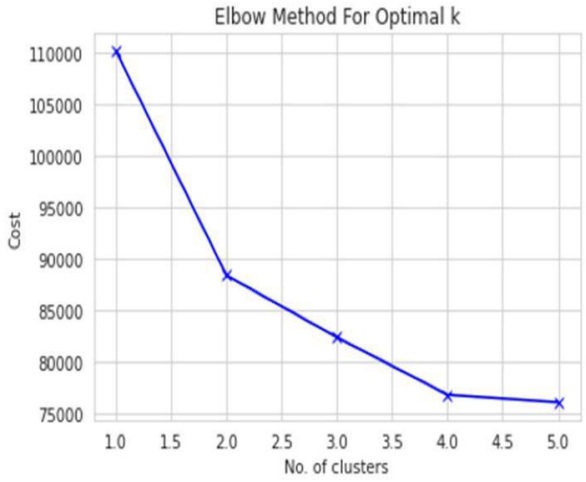

Figure 1. Described male dataset with elbow method

Additionally, heart disease has a varying prevalence rate among men and women, indicating a need for gender-specific analysis. Subsequently, we utilised the elbow curve method to determine the optimal number of clusters for both the male and female datasets. The knee joint, as depicted in Figure 1 and Figure 2, was located at 2.0 in both cases, indicating that 2

was the optimal number of clusters for both the male and female datasets. Further, a correlation table was prepared to determine the correlation between different categories. From the mean arterial pressure (MAP Class), it was observed that cholesterol and age were highly correlated factors. Intra-feature dependency can also be explored with the help of this correlation matrix. Next, a training dataset comprising 80% of the data and a testing dataset containing the remaining 20% were created from the dataset.

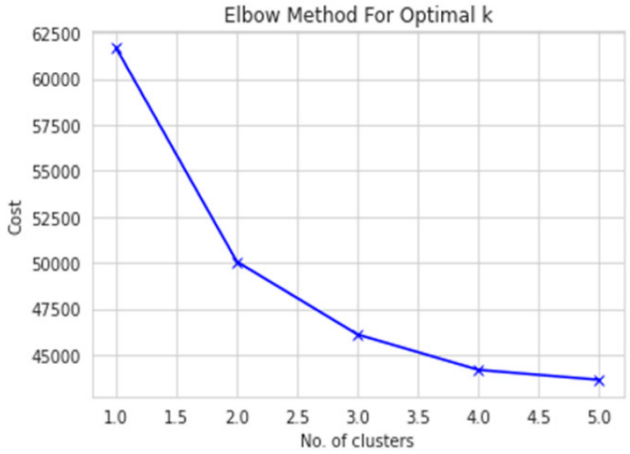

Figure 2. Described female dataset with elbow method

A model was trained using the training dataset, and its performance was assessed using the testing dataset. Various classifiers, including decision tree classifier, random forest classifier, multilayer perceptron, and Boost, were applied to the clustered dataset to evaluate their performance. The performance of each classifier was then evaluated using metrics such as accuracy, precision, recall, and F-measure scores. Decision trees, which are treelike structures used to manage large datasets, were among the classifiers applied. They are often depicted as flowcharts, with outer branches representing the results and inner nodes representing the dataset's properties. Decision trees are popular due to their efficiency, reliability, and ease of interpretation. The projected class label for a decision tree originates from the tree's root, and subsequent steps in the tree are decided by comparing the value of the root attribute with the information in the record. Entropy changes when training examples are divided into smaller groups using a decision tree node, and the measurement of this change in entropy is known as information gain.

Random Forest

The random forest [13] algorithm belongs to a supervised classification technique category consisting of multiple decision trees working together as a group. The class with the most votes become the prediction made by our model. Each tree in the random forest makes a class prediction, which eliminates the limitations of the decision tree algorithm. This improves accuracy and reduces the overfitting of the dataset. When used on large datasets, the random forest approach may still provide the same results even if a significant portion of record values are missing. The samples produced by the decision tree may be saved and used with various data types [31]. In the research in [7], random forest achieved a test accuracy of 73% and a validation accuracy of 72% with 500 estimators, four maximum depths, and one random state.

Multilayer Perceptron

The multilayer perceptron (MLP) is an artificial neural network of multiple layers. Single perceptron can only solve linear problems, but MLP is better suited for nonlinear examples. MLP is used to tackle complex issues. A feed-forward neural network with many layers is an example of an MLP [32]. MLP usually uses other

activation functions beyond the step function. The buried layer neurons often perform sigmoid functions. As with step functions, smooth transitions rather than rigid decision limits are produced using sigmoid functions [33]. In MLPs, learning also comprises adjusting the perceptions' weights to obtain the lowest possible error. This is accomplished via the back propagation technique, which reduces the MSE. Then, rigid decision limits are produced using sigmoid functions [33]. In MLPs, learning also comprises adjusting the perception's weights to obtain the lowest possible error. This is accomplished via the back propagation technique, which reduces the MSE.

XGBoost

XGBoost [14] is a version of gradient-boosted decision trees. This algorithm involves sequentially creating decision trees. All the independent variables are allocated weights, subsequently used to produce predictions by the decision tree. If the tree makes a wrong prediction, the importance of the relevant variables is increased and used in the next decision tree. The output of these classifiers/predictors is then merged to produce a more robust and accurate model. In a study by [34], the XGBoost model achieved 73% accuracy with the parameters 'learning rate': 0.1, 'max_depth': 4, 'n_estimators': 100, 'cross-validation': 10 folds including 49,000 training and 21,000 testing data instances on 70,000 CVD dataset.

3. CONCLUSIONS AND FUTURE WORKS

This study aimed to classify heart disease using various models and a real-world dataset. The k-mode clustering algorithm was applied to a dataset of patients with heart disease. The preprocessing steps included converting age attributes into years, dividing them into 5-year intervals and segmenting diastolic and systolic blood pressure data into ten intervals. Gender-based dataset splitting was also employed to consider unique characteristics and progression of heart disease in men and women. The elbow curve method determined the optimal number of clusters for both male and female datasets, with the MLP model exhibiting the highest accuracy at 87.23%. These results underscored the potential of k-mode clustering for precise heart disease prediction, suggesting its utility in targeted diagnostic and treatment strategies. Despite promising outcomes, limitations should be acknowledged. The study's reliance on a single dataset may hinder generalizability to other populations or patient groups.

Additionally, it overlooked potential heart disease risk factors such as lifestyle choices or genetic predispositions. Evaluation on a held-out test dataset was omitted, limiting insights into model generalizability, and the interpretability of cluster formation was not assessed. Future research could address these limitations by comparing k-mode clustering performance with other algorithms like k-means or hierarchical clustering. Evaluating the impact of missing data and outliers on model accuracy, developing strategies for handling these cases, and assessing model performance on unseen data would be beneficial. Moreover, efforts should establish the robustness and generalizability of results and interpretability of formed clusters, thus providing insights for informed decision-making in healthcare settings.

REFERENCES

[1]. Patel B, Sengupta P. Machine learning for predicting cardiac events: what does the future hold? Expert Rev Cardiovasc Ther. 2020; 18(2):77–84.

[2]. Baharvand-Ahmadi B, Bahmani M, Zurbaran A. A brief report of Rhazes manuscripts in cardiology and cardiovascular diseases. Int J Cardio. 2016; 207:190–1.

[3]. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944.

[4]. An H, Ye Q, Zhang T, Yu D-J, Yuan X, Xu Y, et al. Least squares twin bounded support vector machines based on L1-norm distance metric for classification. Pattern Recogn. 2018; 74:434–47.

[5]. Jaworski M, Duda P, Rutkowski L. New splitting criteria for decision trees in stationary data streams. IEEE Trans Neural Newt Learn Syst. 2018;29:2516–29.

[6]. Zhang S, Cheng D, Deng Z, Zong M, Deng X. A novel K-NN algorithm with data-driven k parameter computation. Pattern Recogn Lett. 2018;109:44–54.

[7]. Abdar M, Zomorodi-Moghadam M, Das R, Ting IH, Performance analysis of classification algorithms on early detection of liver disease. Expert Syst Appl. 2017; 67:239–51.

[8]. Abdar M, Yen NY, Hung JC-S. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng. 2017; 10:1–13.

[9]. Pławiak P. Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals, Swarm. Evol Computer. 2018; 39:192–208.

[10]. Pławiak P, Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. Expert Syst Appl. 2018;92:334–49.

[11]. Khozeimeh F, Alizadehsani R, Roshanzamir M, Khosravi A, Layegh P, Nahavandi S. An expert system for selecting wart treatment method. Computer Biol Med. 2017; 81:167–75.

[12]. Khozeimeh F, Azad FJ, Oskouei YM, Jafari M, Tehranian S, Alizadehsani R, et al. Intralesional immunotherapy compared to cryotherapy in the treatment of warts. Int J Dermatology. 2017; 56:474–8.

[13]. Alizadehsani R, Abdar M, Jalali SMJ, Roshanzamir M, Khosravi A, Nahavandi S. Comparing the performance of feature selection algorithms for wart treatment selection. Proc. Int. Workshop Future Technol; 2018. p. 6–18.

[14]. Wu C, Yeh W, Hsu WD, Islam M, Nguyen P, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Computer Methods Prog Biomed. 2019; 170:23–9.

[15]. Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl. 2019; 78:19905–16.

[16]. Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge-

driven approach. Expert Syst Appl. 2013;40(1):96–104.

[17]. Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst. 2016;40(7):1–7.

[18]. EI-Bialy R, Salamay MA, Karam OH, Khalifa ME. Feature analysis of coronary artery heart disease datasets. Proc Computer Sci. 2015; 65:459–68.

[19]. Alizadehsani R, Abdar M, Roshanzamir M, Khosravi A, Kebria PM, Khozeimeh F, et al. Machine learning-based coronary artery disease diagnosis: A comprehensive review. Computers Biol Med. 2019; 111:103346.

[20]. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. 2019; 16:100203.

[21]. Ahmed H, Younis EMG, Hendawi A, Ali AA. Heart disease identification from patients' social posts, machine learning solution on spark. Future Gener Computer Syst. 2020; 111:714–22. 10.1016/j.future.2019.09.056.

[22]. Beunza J-J, Puertas E, García-Ovejero E, Villalba G, Condes E, Koleva G, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inform. 2019; 97:103257.

[23]. Kim D, You S, So S, Lee J, Yook S, Jang DP. A data-driven artificial intelligence model for remote triage in the prehospital environment. Plops ONE. 2018;13(10): e0206006.

[24]. Shah D, Patel S, Bharti SK. Heart Disease Prediction using Machine Learning Techniques. SN Computer Sci. 2020; 1:345–6.

[25]. Pal M, Parija S. Prediction of Heart Diseases using Random Forest. J Physics: Conf Ser. 2021; 1817:012009.10.1088/1742-6596/1817/1/012009.

[26]. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, GYH Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240. [Google Scholar] [CrossRef]

[27]. Hassan, Ch Anwar ul, et al. "Effectively predicting the presence of coronary heart disease using machine learning classifiers." Sensors 22.19 (2022): 7227.

[28]. Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. Sustainability 2022, 14, 14208.