Webology (ISSN: 1735-188X)
Volume 19, Number 3, 2022

A Comparative Study Of Heart Disease Prediction Based On Principal Component Analysis And Classification Techniques

Dr. Muhammad Affan Alim[1] , Haris Ahmed[2] , Dr. Waleej Haider[3] , Ahsan Masroor[4]

[1,2,4]College of Computing and Information Sciences, PAF-Karachi Institute of Economics and Technology Karachi, Pakistan.

[3]Department of Computer Science & Information Technology, Sir Syed University of Engineering and Technology Karachi, Pakistan.

Abstract

One of the most difficult challenges today facing medical practitioners is determining whether or not a person may acquire heart disease. Heart disease is the greatest cause of death in the contemporary period, killing about one person every minute on average. One of the most important uses of data science is the analysis of massive amounts of data generated in the area of healthcare. Since anticipating cardiac sickness is difficult, there is an urgent need to automate the procedure. This will assist to limit the hazards of the treatment and provide patients with early warnings. By using the Principal Component Analysis (PCA) dimensionality reduction approach to datasets of cardiovascular disorders, this research analyses the overall performance of a variety of different machine learning models. Furthermore, the authors use the K Nearest Neighbor Model and the Random Forest Model to the datasets and compare the model's performance with and without PCA. This is done in order to identify which approach yields the greatest outcomes. In compared to the KNN approach, the Random Forest (RF) algorithm in combination with the Principal Component Analysis (PCA) achieved a score of 91.0 percent in the categorization of heart disease.

Keywords:

Random Forest, KNN Algorithm, PCA, Principal Component Analysis, Heart Disease Prediction

Introduction

To a significant extent, this article focuses on several data mining techniques that are used to diagnose heart illness (Kim et al., 2018). The research goes into further information about each of these strategies. A main organ in the human body, the heart performs a variety of essential activities. In a nutshell, it is in charge of maintaining a healthy circulation throughout our whole body. The heart's irregularities may induce discomfort in other places of the body (Lopaschuk et al., 2021). As a result, the heart serves as the brain's command center. If a person's heart is unable to pump blood throughout their body as it should, they are considered to have heart disease.

In today's culture, heart disease is a leading cause of mortality for the majority of those who have a deadly heart condition. Coronary heart disease and hypertension may both be exacerbated and even accelerated by a sedentary lifestyle that includes smoking, drinking alcohol, and eating a diet rich in fats (Chen, Gong, Wang & Guo, 2020). Every year, more than 10 million people die as a direct result of cardiovascular disease (Frieden, & Jaffe, 2018).

Only a healthy lifestyle and early detection of anomalies may lower one's chance of acquiring a heart-related ailment. Contemporary healthcare's biggest difficulty is in delivering high-quality services and correct diagnoses in an efficient manner. Despite the fact that heart illnesses have been recognized as the top cause of death worldwide in recent years, they are also the ones that are most easily treated and managed (Chen, Gong, Wang & Guo, 2020). In addition, obtaining a correct diagnosis of disease at the appropriate point in time is the single most important factor affecting the efficacy of disease treatment. Healthcare specialists have compiled vast volumes of data that may be examined and mined for valuable insights.

It is possible to find useful and previously unknown information by using data mining methods when applied to the enormous amounts of data that are now accessible (Amin, Chiam & Varathan, 2019). The vast bulk of the medical database's content is made up of discrete bits of data. It is because of this that the process of drawing inferences from discrete data is a difficult one. For large datasets that may be handled by a machine learning area called data mining, the proper organization is critical. Some of the medical applications that may benefit from machine learning include disease detection, sickness diagnosis, and disease forecasting.

Medical practitioners in the early stages of heart disease might benefit from the development of a diagnostic tool. Consequently, providing the right therapy to patients will be easier while avoiding the negative consequences of doing so. Machine learning makes it much simpler to uncover previously concealed discrete patterns and, as a result, undertake an analysis of the data that is supplied (Babu et al., 2017). Heart disease may be accurately predicted and diagnosed using machine learning approaches when data analysis is completed. In this concern, the current research has employed two machine learning techniques, such as Random Forest and KNN, with and without PCA. These two machine learning algorithms were analyzed first with PCA and then, without PCA in this work in order to predict cardiac disease in its earliest stages.

Related Work

Using the machine learning dataset by UCL has focused on the use of machine learning in the prediction of heart illness. A multitude of data mining approaches, which may be further divided into the following categories, have allowed researchers to achieve varying degrees of accuracy. To classify cardiac disease, Golande & Pavan Kumar (2019)'s research shed light on many potential machine-learning (ML) algorithms that may be used. Decision Tree and K-Means algorithms, which may be used to classify data for classification purposes, were studied in-depth, and their accuracy was assessed. According to this study, the Decision Tree is the most accurate method. It is also feasible to increase its effectiveness by merging various study methodologies.

In addition, a paper by Ramalingam, Dandapath & Raja (2018) has demonstrated the usage of K-Nearest Neighbor, Decision Trees (DT), Support Vector Machines (SVM), NaïveBayes and Random Forest (RF), with PCA for feature extraction. In this study, it was found that Random Forest and Naïve Bayes classifier performed well with PCA and provided effective results.

Using the MapReduce algorithm and data mining methods, the research by Nagamani, Logeswari & Gomathy (2019) has proposed a system implementation. The study claimed that the 45 test cases used by them, achieved a higher level of accuracy than a normal fuzzy artificial neural network. This conclusion was based on the comparison of the two sets of results. The combination of dynamic schema and linear scaling increased the accuracy of the technique utilized in this situation.

For the purpose of comparing and contrasting the efficacy of various techniques to machine learning, Alotaibi (2019) developed a model. The study found that the Rapid Miner tool outperformed the Matlab and Weka tools when it came to mining results. Decision Tree, Logistic Regression, Random Forest, Naive Bayes and SVM were some of the classification methods examined and compared in this study. Alotaibi (2019) concluded that decision tree was the most accurate.

Repaka, Ravikanti & Franklin (2019) conceived the notion of employing Naive Bayesian (NB) techniques to classify datasets and the Advanced Encryption Standard (AES) algorithm to protect the transfer of data, which he subsequently put into practice. In addition, the primary objective of the study by T Thomas & Princy (2016) was to use a range of classification algorithms to predict cardiovascular disease. The data were classified using Naive Bayes, KNN, Decision trees, and neural networks, and the accuracy of the classifiers was tested using a range of criteria.

Using a Naive Bayes classifier and a support vector machine, Gavhane et al (2018) were able to effectively predict cardiovascular disease. Using Mean Absolute Error (MAE), Sum Squared Error (SSE), and Root Mean Squared Error (RMSE) analysis, it was shown that the SVM technique was more accurate than the Naive Bayes technique. A heart disease prediction system based on the inputs of Table 1 was the primary inspiration for the system that was suggested after reading the aforementioned papers. The study compared the Accuracy, Precision, Recall, and F-measure scores of the algorithms Decision Tree, Random Forest, Logistic Regression, and Naive Bayes to see which algorithm was most suited for use in the prediction of heart disease.

Proposed Model

The suggested study evaluated the performance of the aforementioned two classification algorithms (Random Forest and KNN with and without PCA Algorithm) to produce a heart disease prognosis. This study aimed to assess with sufficient precision whether or not the patient in issue has a heart problem. A qualified medical practitioner use the patient's health report to enter data into the system. This information is included into a model that is used to estimate the risk of heart disease for individuals. Figure 1 illustrated the whole method for convenience:

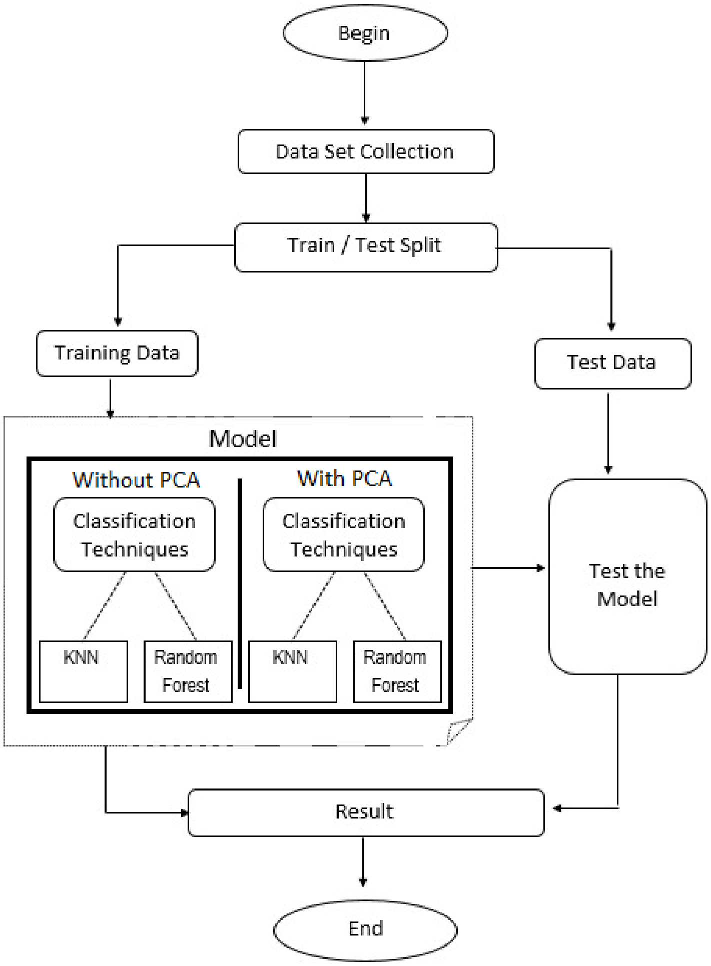
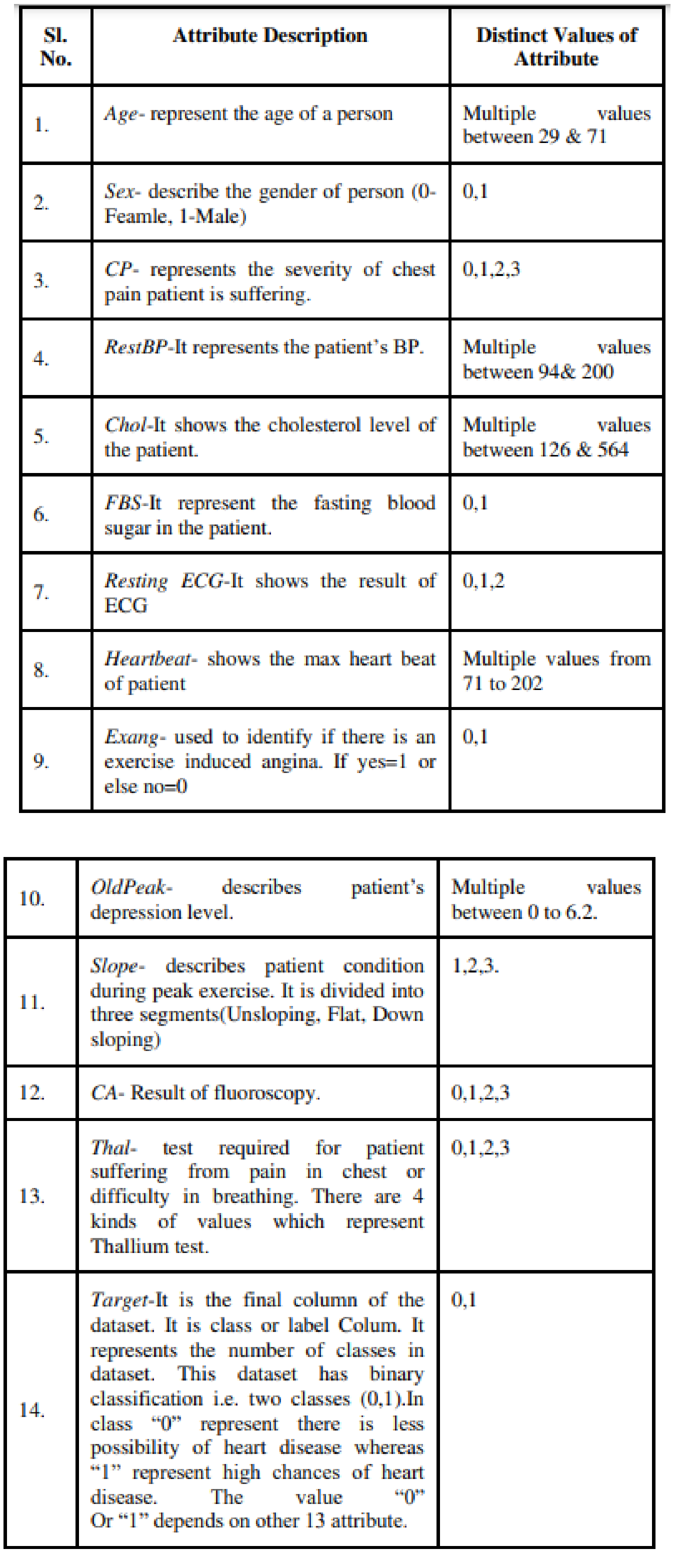


Figure 1: Generic Model Predicting Heart Disease

Data Collection and Pre-processing

The UCI dataset was the only one used in the study, despite the Heart Disease Dataset also being utilized. Despite the fact that this database comprises 76 different attributes, this research has used just 14 of them. These attributes were used in the study of Ramalingam, Dandapath & Raja (2018) and we adopted those attributes to use in our study. The following table (Table 1) provides a detailed description of each of the fourteen attributes adopted from Ramalingam, Dandapath & Raja (2018) that will be used in the proposed research.



Sl. No.	Attribute Description	Distinct Values of Attribute
1.	*Age*- represent the age of a person	Multiple values between 29 & 71
2.	*Sex*- describe the gender of person (0-Feamle, 1-Male)	0,1
3.	*CP*- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	*RestBP*-It represents the patient's BP.	Multiple values between 94& 200
5.	*Chol*-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	*FBS*-It represent the fasting blood sugar in the patient.	0,1
7.	*Resting ECG*-It shows the result of ECG	0,1,2
8.	*Heartbeat*- shows the max heart beat of patient	Multiple values from 71 to 202
9.	*Exang*- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	*OldPeak*- describes patient's depression level.	Multiple values between 0 to 6.2.
11.	*Slope*- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3.
12.	*CA*- Result of fluoroscopy.	0,1,2,3
13.	*Thal*- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	*Target*-It is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "0" Or "1" depends on other 13 attribute.	0,1

Table 1: Features Selected From Dataset (Source: Ramalingam, Dandapath & Raja, 2018)

Classification

Methods of machine learning, such as the Random Forest Algorithm and the KNN Algorithm, employ the Table 1 attributes as input. The input dataset is then separated into a training dataset and a test dataset. In order to characterize the process of "teaching" an algorithm, this dataset is referred to as the "training dataset." Testing data are used to evaluate the performance of the trained model. Algorithms are assessed using many metrics, including as accuracy, precision, recall, and F-measure scores, which will be discussed in further depth in the following sections. In the course of doing this research, a variety of computational techniques were exhaustively explored.

• Random Forest Random

In addition to its usage in regression, Random Forest Algorithm may also be used to classification. It is feasible to derive conclusions about the future from the data using a tree structure (Yadav & Pal, 2020). When applied to a large dataset, the Random Forest method yields the same results regardless of the amount of missing record values (Javeed et al., 2019). It is possible to save the samples generated by the decision tree so that they may be applied to more data sets in the future. This increases the adaptability of the samples. Random forest consists of two stages: first, you produce a random forest classifier, and then you use it to make a prediction.

• KNN Algorithm

K-Nearest Neighbor is one of the most basic applications of machine learning that use the Supervised Learning paradigm. Using this method, a new case or piece of data is assigned to the category that most closely resembles the existing categories (Yadav & Pal, 2020). Existing instances are placed in the category that is most equivalent to the new category using the new category. As long as the K-NN algorithm has access to the data, it is able to classify new information based on the degree to which it resembles the data that has been maintained. In other words, if new data is acquired, it may be rapidly allocated to a more relevant group. This may be performed using a more expedient strategy.

This non-parametric technique makes no assumptions about the data, which is an additional advantage. This is the result of its non-parametric nature. During the learning process, the KNN algorithm does nothing other than store the collected data in its internal memory (Anggoro & Kurnia, 2020). When a new piece of information is received by the system, it will be put in a category with other information that is comparable.

• Principal Component Analysis

Principal Component Analysis (PCA) is a kind of unsupervised learning technique used in the field of machine learning to reduce the number of dimensions included in a dataset.

This statistical method use orthogonal transformation to turn observations with correlated attributes into a set of linearly uncorrelated data (Karamizadeh et al., 2020). To achieve this objective, the connected attributes are translated into orthogonal coordinates. Principal Components is the name given to these freshly rebuilt characteristics and attributes. It is one of the most often used tools for exploratory data analysis and predictive modelling. It is a method that seeks to decrease the amount of variation in a dataset in order to extract the greatest number of consistent patterns from the data.

In most circumstances, principal component analysis will seek a lower-dimensional surface on which to project higher-dimensional data. For PCA, to be successful, the variance of each characteristic must be taken into account. This is done because a high attribute demonstrates a clear separation between classes, and as a result, the dimensionality is reduced to its minimum feasible value (Kherif & Latypova, 2020). Real-world applications of PCA include image processing, movie recommendation systems, and power distribution optimization across many communication channels. Due to the fact that it is a strategy for extracting characteristics, it prioritizes the most relevant aspects while disregarding the less significant ones.

- KNN and Random Forest with PCA

The historical data set that is used as an input in a model that combines PCA and KNN is first generated with the help of a sliding window, then it is subjected to PCA in order to obtain principal components that are abundant in information, and finally, it is fed into KNN in order to generate predictions (Yadav & Pal, 2020). In this way, the model is able to generate accurate predictions. In order to build the model, each and every one of these phases needs to be completed. PCA, on the other hand, is able to accomplish dimensionality reduction, which may lead to a reduction in the number of features that the Random Forest model is needed to analyze (Yadav & Pal, 2020). This may be the case if the number of features is decreased. Due to this fact, principal component analysis could be able to assist in accelerating the process of training your Random Forest model. It is important to be aware that one of the most significant drawbacks of using Random Forests is the high processing cost that is linked with them (it can take a long time to run the model).

Results and Findings

Analysis of Machine Learning Algorithms with and without PCA

These results are obtained by running KNN and Random Forest algorithm with and without PCA. The metrics used to carry out the performance analysis is the precision, recall, accuracy score and F-measure. In the table presented below, the Precision score (eq 1) demonstrates the measure of the positive analysis, whereas the Recall score (eq 2) demonstrates the actual positives, lastly the F-measure (eq 3) is used to test the accuracy of the dataset.

Precision = (TP) / (TP +FP)………………………………………………………… (eq 1)

Recall = (TP) / (TP+FN) …………………………………………………………… (eq 2)

F– Measure = (2 * Precision * Recall) / (Precision +Recall) ………………………. (eq 3)

- TP True positive: the patient has the disease and the test is positive.
- FP False positive: the patient does not have the disease but the test is positive
- TN True negative: the patient does not have the disease and the test is negative.
- FN False negative: the patient has the disease but the test is negative

In the analysis, the pre-processed dataset is utilized to carry out the tests, and the techniques that were mentioned above are examined and employed in the process of carrying out the experiment. It is necessary to make use of the confusion matrix in order to get the performance metrics that have been discussed up to this point in the research. The accuracy score that was achieved for each of the 2 distinct classification techniques, Random Forest, and KNN, is shown in the table 2 (without PCA) and 3 (with PCA) that follows:

Algorithm	Precision	Recall	F-measure	Accuracy
Random Forest	0.8545	0.9216	0.8868	88.00%
KNN	0.7895	0.8654	0.8257	81.00%

Table 2: Analysis of Machine Learning Algorithms without PCA

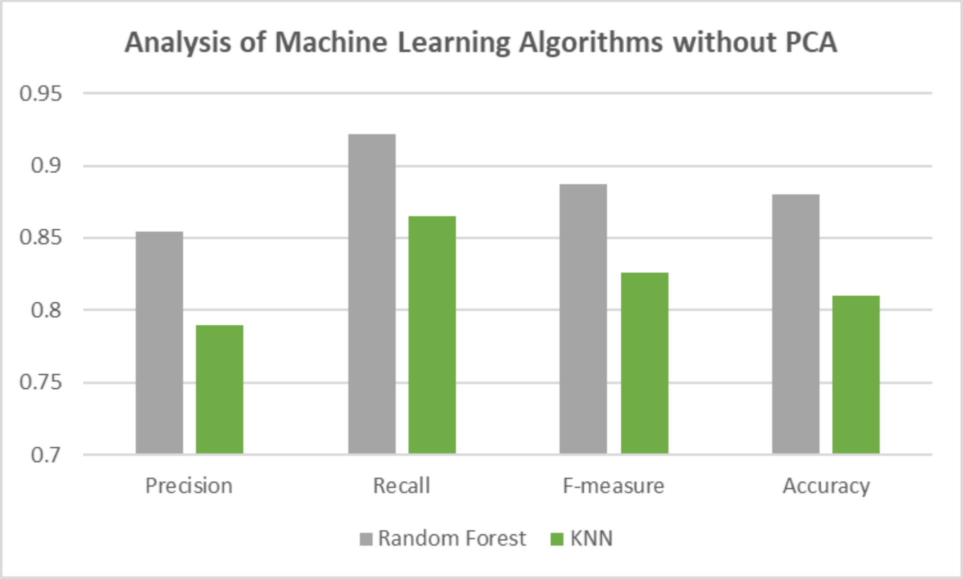


Figure 2: Graph of Analysis of Machine Learning Algorithms without PCA

Algorithm	Precision	Recall	F-measure	Accuracy
Random Forest	0.9107	0.9273	0.9189	91.00%
KNN	0.8448	0.9245	0.8829	87.00%

Table 3: Analysis of Machine Learning Algorithms with PCA

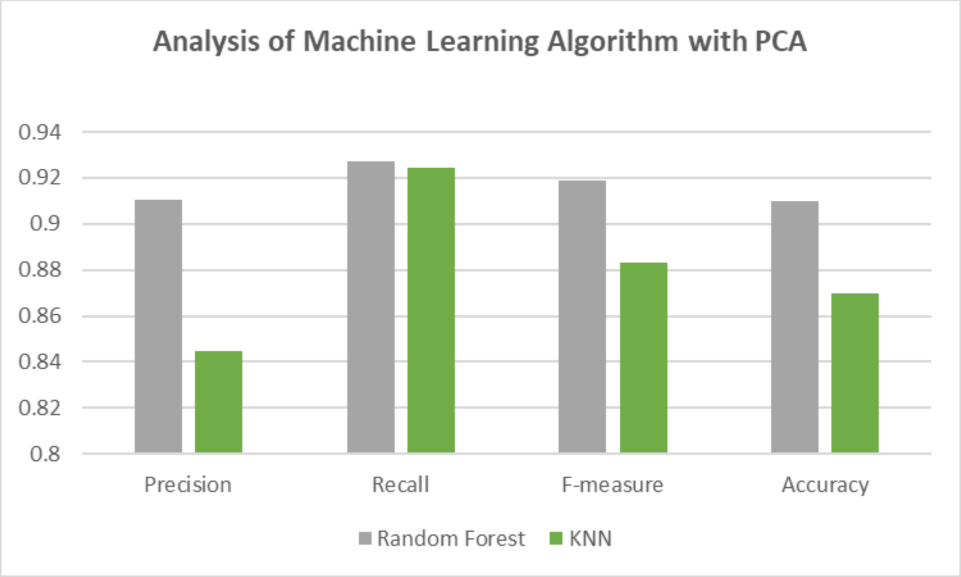


Figure 3: Graph of Analysis of Machine Learning Algorithms with PCA

Conclusion

It is crucial to build a system that can reliably and efficiently forecast the occurrence of cardiac problems due to the increasing number of deaths caused by these conditions. The objective of the study was to determine the best algorithm for diagnosing cardiac problems using machine learning. The accuracy of four machine learning algorithms, Random Forest, and KNN in predicting cardiovascular disease is evaluated with and without PCA using a dataset from the UCI machine learning repository. According to the findings of this research, the Random Forest algorithm along with PCA, with an accuracy score of 90.00 percent, is the best technique for predicting heart disease. This approach may be utilized to analyze a far larger dataset than the one used in this research. This will enhance the prognosis of cardiac disease by medical professionals and lead to better results.

References

Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6), 261-268.

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

Anggoro, D. A., & Kurnia, N. D. (2020). Comparison of accuracy level of support vector machine (SVM) and K-nearest neighbors (KNN) algorithms in predicting heart disease. International Journal, 8(5).

Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., & Hena, M. (2017, April). Heart disease diagnosis using data mining technique. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 1, pp. 750-753). IEEE.

Chen, Y., Gong, X., Wang, L., & Guo, J. (2020). Effects of hypertension, diabetes and coronary heart disease on COVID-19 diseases severity: a systematic review and meta-analysis. MedRxiv.

Frieden, T. R., & Jaffe, M. G. (2018). Saving 100 million lives by improving global treatment of hypertension and reducing cardiovascular disease risk factors. The Journal of Clinical Hypertension, 20(2), 208.

Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1275-1278). IEEE.

Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. International Journal of Recent Technology and Engineering, 8(1), 944-950.

Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., & Nour, R. (2019). An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access, 7, 180235-180243.

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2020). An overview of principal component analysis. Journal of Signal and Information Processing, 4.

Kherif, F., & Latypova, A. (2020). Principal component analysis. In Machine Learning (pp. 209-225). Academic Press.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry investigation, 15(3), 235.

Lopaschuk, G. D., Karwi, Q. G., Tian, R., Wende, A. R., & Abel, E. D. (2021). Cardiac energy metabolism in heart failure. Circulation research, 128(10), 1487-1513.

Nagamani, T., Logeswari, S., & Gomathy, B. (2019). Heart disease prediction using data mining with mapreduce algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI) (pp. 292-297). IEEE.

Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.

Yadav, D. C., & Pal, S. (2020). Prediction of heart disease using feature selection and random forest ensemble method. International Journal of Pharmaceutical Research, 12(4), 56-66.