Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **616:621***

Cardiovascular Disease Prediction using Machine Learning Ensemble Methods

Vandana Joshi[1*], Shruthi R[2], Varshitha B[3], Devarasetty Vivek Kumar[4], Mallikarjuna M[5]

[1,2,3,4,5]*School of Computer Science and Engineering, REVA University*
[1]*r17cs526@cit.reva.edu.in*
[2]*r17cs444@cit.reva.edu.in*
[3]*r17cs456@cit.reva.edu.in*
[4]*r17cs433@cit.reva.edu.in*
[5]*mallikarjuna.m@reva.edu.in*

**Corresponding author's E-mail: r17cs526@cit.reva.edu.in*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 29 Nov 2023

CC License
CC-BY-NC-SA 4.0

Abstract

Recently the main cause of death occurring due to cardiovascular disease has happened even in the past. Early diagnosis of the disease can be assessed to reduce the high risk and ensure healthiness. Data mining techniques have been significantly used as it helps in zero or less intervention of humans and it is seen as the best technique as it gives precise result with the best accuracy. The study is conducted on ensemble methods and built a model using boosting and bagging classifiers. The objective of this work is to design and implement a heart disease prediction system using machine learning ensemble methods namely, Random Forest, Adaptive Boosting, Gradient Boosting, and XGBoost. The effective performance of the applied ensemble techniques is analyzed, and a mobile application is developed for the same. The proposed mobile application is built as a user interface that accepts data based on clinical attributes concerning heart disease. This mainly helps in the medical field such as laboratories that incorporate the developed model. The outcome of the proposed model predicts the probability of a person suffering from heart disease. The accuracy of the models is evaluated.

Keywords: *Machine Learning, Ensemble Learning, Adaptive Boosting, Gradient Boosting, XGBoost, Mobile Application*

1. Introduction

Coronary heart disease is the most common type of heart disease. Millions of deaths occur among adults and old people. Cardiovascular disease prediction is being predicted in the proposed model. It is seen that various Machine Learning techniques would give accurate predictions compared to any other techniques. The ensemble technique has been used in the given model to enhance the accuracy to the best. A handheld device interface has been developed which takes the user input based on the clinical parameters such as name, age, sex, cholesterol, bp, etc. which are processed further. A literature survey is made on cardiovascular disease prediction by going through several technical papers and it is observed that the accuracy of the machine learning ensemble model has a higher standard compared to other models. In section III, a brief description of the proposed model is made by describing the flow of procedure involved and the algorithms used in the proposed model is given with an explanation. Section IV presents the result after a comparative study is performed on several ensemble classifiers such as Random Forest, Adaptive boosting, Gradient Boosting and, XGBoost. The one with the highest accuracy has been incorporated in the mobile application and the output is displayed on the same.

Literature Survey

O. Terrada et.al., mainly focused on predicting if a patient is suffering from atherosclerosis which is one type of cardiovascular disease by using two supervised ML algorithms. ANN and Adaptive Boosting techniques are used, and databases are collected from UCI and Z-Alizadeh datasets. The comparative study has been made among various systems Weighted Fuzzy, FDT, Neural Network and concluded that ANN stood among all with the accuracy of 91.41% and, with the Hungarian dataset the

proposed system outstands by giving 90% of accuracy and 94% in Z-Alizadeh Sani dataset for atherosclerosis disease. [1]

Jane Preetha Princy et.al., propounded a comparative study on various supervised machine learning classifiers. It has been asserted that the decision tree algorithm has given a better accuracy rate compared to other ML techniques including Naive Bayes, Logistic Regression, KNN, SVM, Random Forest. In this paper, the analysis is made on dimensionality reduction by inferring that the dimensionality of datasets impacts the algorithms and, the outcome can either be positive or negative. [2]

S. Mohan et.al., divide the proposed work into various sections which describe each phase involved in designing a machine learning model. Beginning from data pre-processing, feature selection, classification of various machine learning models and the performance results have been reported. It mainly focuses on the model which gives the best prediction. Hence, to achieve high accuracy, a Hybrid Machine Learning technique called HRFLM is introduced which is a combination of Random Forest and Linear Model. It has reduced the classification error and improved the accuracy up to 88.4%. [3]

Jesada Kajornrit et.al., introduces a cardiovascular disease prediction system using comparative analysis on ensemble back-propagation neural netwok. It includes linear regression, k-nearest neighbor, ensemble voting, support vector machine, bagging techniques, and a back-propagation neural network. Initially, the 10-fold cross-validation method is used for evaluation by varying k values. Ensemble methods, voting gave extremely accurate results showing best performance which showed a variation of 10 percent from BPNN. [4]

H. A. Esfahani et.al., presents data mining techniques and performed a comparative study among several classifiers such as Decision tree, rough set, Naive Bayes, SVM. As an outcome of the comparison rough set, Naive Bayes and neural networks give greater accuracy when compared to other classifiers. Hence combining these an ensemble classifier is built. The fusion of these classifiers employing a weighted majority vote yielded an accuracy of 89 percent. [5]

Dinesh Kumar G et.al., recommends a prediction model for a person suffering from cardiovascular disease. The comparative study has been made based on the accuracy rates of several classifiers such as Support Vector Machine, Naïve Bayes, Random Forest, Logistic Regression, and Gradient boosting by making use of R software and further, it aims towards ensemble method as it yields the better performance by aggregating the classifiers and produce better accuracy rate. [6]

Shan Xu et.al., have worked on the real-time dataset provided by Peking University People's Hospital. Text analysis has been made by using several methods such as text segmentation and synonym identification. To enhance the accuracy to the maximum, instead of using individual classifiers ensemble method has been adopted. Sub- classifiers are chosen based on accuracy and a voting mechanism is established to aggregate the sub-classifiers. As a result, it is seen that 79.3% of accuracy has been achieved which is higher than the individual classifiers such as Naive Bayes and SVM. [7].

2. Materials And Methods

Data Source

The healthcare sectors have possessed substantial patient records. The Cleveland dataset for the cardiovascular disease prediction has been obtained from the UCI Machine Learning Repository. The dataset comprises 303 instances and 14 clinical traits. The clinical attributes are analogous to heart disease like age, sex, resting blood pressure, serum cholesterol etc.

Proposed System

The proposed work involves the probability prediction of how likely a person is diagnosed with heart disease. The bagging and boosting techniques of ensemble learning are availed. The applied algorithms are relatively studied and the finest is employed for the prediction. The model is then deployed to the Heroku cloud platform and further, a mobile application is developed. The patient's data can be collected either through a mobile application or website. The data is processed using XGBoost algorithm and the probability prediction is displayed on the user interface [8-15].

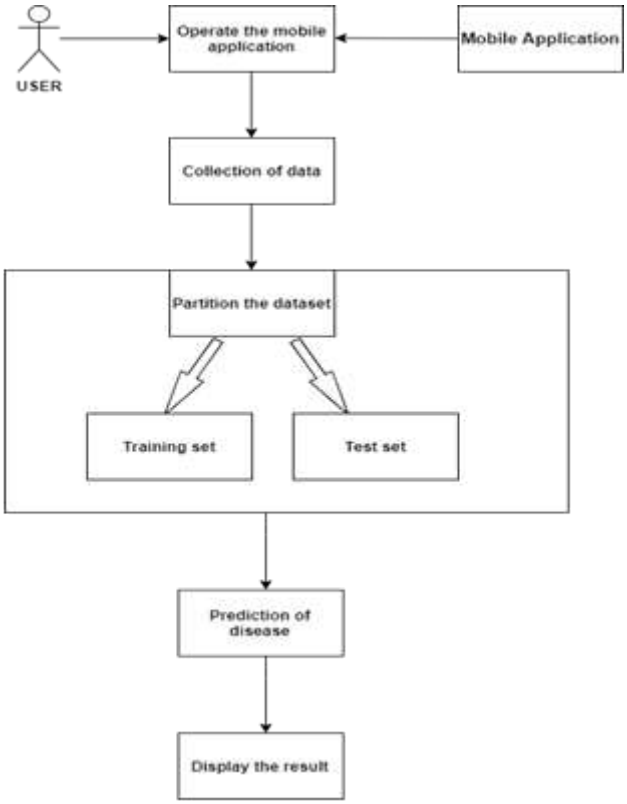

Fig. 1. Proposed Architecture

Figure 1 illustrates the proposed architecture. The mobile application renders an interface through which the user furnishes medical data. The probability of the occurrence of heart disease is reckoned with and is displayed on the application.

Ensemble methods are the classifier combination methods that combine multiple machine learning techniques into one powerful predictive model. The two categories of ensemble methods used are: Bagging and Boosting.

Bagging Algorithms

Bagging method creates the classifiers parallelly. The base learners in bagging are independent of each other and results in the decrease of variance. It bootstraps the subsamples and then aggregates the results over the weak learners, hence the name Bootstrap Aggregation.

1) Random Forest: Random Forest is a bagging ensemble method that utilizes decision tree as the base learner. The sample of records is created using row sampling with replacement and feature sampling. Each of the decision trees is provided with the set of records for training. For a given test data, majority vote is performed on the predicted results and the cardiovascular disease prediction is done.

Boosting Algorithms

In boosting technique, the base learners are created sequentially and combined to form a strong learner to increase the accuracy. The objective here is to train the weak learners successively, each trying to rectify the formerly misclassified records.

1) Adaptive Boosting: Adaptive boosting can be used to boost the performance of any machine learning model. The trees in AdaBoost consist of one root and two leaf nodes called stumps. This can be contemplated as a forest of stumps that technically are weak learners. In the proposed model, the accuracy of the model improved by 5 percent when logistic regression is used as a weak learner, an alternative to decision tree.

2) Gradient Boosting: Gradient boosting is an ensemble technique in which shallow decision trees are used as weak learners and learning takes place by optimizing the loss. In the cardiovascular disease classification problem, the logarithmic loss function is employed, and a tree is adjoined to the model that diminishes the loss since it follows gradient descent approach.

3) Extreme Gradient Boosting: XGBoost provides the gradient boosting framework that braces many languages like C++, R, Python, Java. The proposed work is developed using python programming.

XGBoost is a good candidate for cloud integration. The pros of using XGBoost are cache optimization, speed, portability, regularization, and auto pruning.

3. Results and Discussion

This project intends to design and implement a mobile application for the prediction of cardiovascular disease and to analyze the effective performance of the applied ensemble techniques. Both gradient boosting and XGBoost attained an accuracy of 91 percent. The XGBoost model is integrated with the Heroku cloud application platform. The accuracy has been computed using the confusion matrix metric.

The proposed system is developed using the Python Flask web framework. The model developed using the XGBoost algorithm is deployed to the Heroku cloud application platform. The URL of the website is created for the users to access. The website can be accessed by the users in two ways: in a web browser or a mobile application. The mobile application is in .apk file format which can be installed on an Android operating system. The mobile application connects to the website hosted on the Heroku cloud platform.

Fig. 2. Front-end view

Figure 2 shows the front-end view of the system. The first webpage prompts the user to enter the 13 clinical parameters which are then processed to predict cardiovascular disease.

Fig. 3. Webpage for confirming details

Figure 3 presents the second web page of the website. It displays the values entered by the user in a table for confirmation.

Fig. 4. Webpage displaying result

Figure 4 presents the result section of the second web page. The probability of a patient vulnerable to heart disease is predicted and displayed on the webpage.

TABLE I COMPARISON of CLASSIFICATION TECHNIQUES

Algorithm Name	**Accuracy**
Random Forest	0.8688524590163934
AdaBoost with Decision Tree as base estimator	0.8360655737704918
AdaBoost with Logistic Regression as base estimator	0.8852459016393442
Gradient Boosting	0.9180327868852459
XGBoost	0.9180327868852459

Table 1 depicts the ensemble classification methods that are applied for cardiovascular disease prediction and their corresponding accuracies. It is inferred that the efficiency of AdaBoost algorithm with logistic regression as base estimator is high when compared to decision trees.

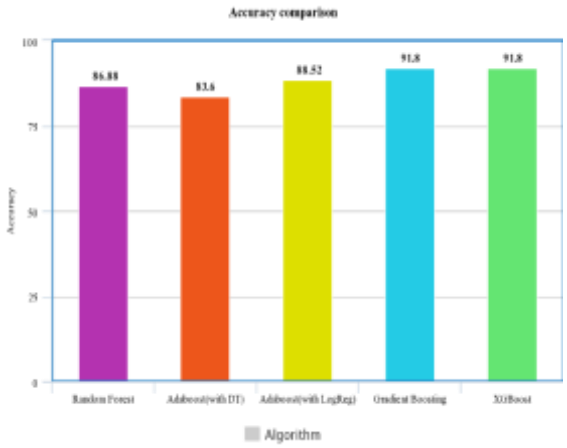

Fig. 5. Comparison of test set accuracy

Figure 5 presents the bar plot which provides the information on the efficiency of algorithms on the split of test set data. It is deduced that XGBoost and gradient boosting algorithms have procured the highest accuracy amidst the applied ensemble techniques.

4. Conclusion

In the proposed work, a website and a mobile application are developed to predict the threat of cardiovascular disease. It is perceived that the boosting algorithms like gradient boost and XGBoost render highest accuracy when contrasted to other algorithms. XGBoost is prone to overfitting and its performance is remarkable, hence used to predict through the user interface.

Future Work

Various mobile applications can be developed to diagnose other chronic diseases like cancer, diabetes, arthritis, etc. The work can be blended with IoT for real-time prediction. The heart disease prediction can be done using artificial neural networks or through machine learning cloud platforms for improved accuracy.

References:

AH Omar Baabood, Prajoona Valsalan, Tariq Ahmed Barham Baomar, IoT Based Health Monitoring System, Journal of Critical Reviews , Vol. 7, Issue. 4, pp. 739-743, 2020.

Bindhia K Francis, Suvanam Sasidhar Babu, Predicting academic performance of students using a hybrid data mining approach, Journal of Medical Systems, 43:162, 2019. https://doi.org/10.1007/s10916-019-1295-4

H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, Iran, 2017, pp. 1011-1014, doi: 10.1109/KBEI.2017.8324946.

Jesada Kajornrit, Piyanuch Chaipornkaew, "A Comparative Study of Ensemble Back-propagation Neural Network for the Regression Problems" 2017 2nd International Conference on Information Technology (INCIT), Bangkok,Thailand, 2017 978-1-5386-1431-0/17.

K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.

O. Terrada, B. Cherradi, S. Hamida, A. Raihani, H. Moujahid and O. Bouattane, "Prediction of Patients with Heart Disease using Artificial Neural Network and Adaptive Boosting techniques," 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, 2020, pp. 1-6, doi: 10.1109/CommNet49926.2020.9199620.

PK Sadineni, Comparative Study on Query Processing and Indexing Techniques in Big Data, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 933-939, 2020.

PM Surendra, S Manimurugan, A New Modified Recurrent Extreme Learning with PSO Machine Based on Feature Fusion with CNN Deep Features for Breast Cancer Detection, Journal of Computational Science and Intelligent Technologies, Vol. 1, Issue. 3, Pp. 15-21, 2020.

R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 570-575, doi: 10.1109/ICICCS48265.2020.9121169.

S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

S. Xu, H. Shi, X. Duan, T. Zhu, P. Wu, and D. Liu, "Cardiovascular risk prediction method based on test analysis and data mining ensemble system," 2016 IEEE International Conference on Big Data Analysis (ICBDA), Hangzhou, China, 2016, pp. 1-5, doi: 10.1109/ICBDA.2016.7509809.

Sajay KR, Suvanam Sasidhar Babu, Vijayalakshmi Yellepeddi, Enhancing The Security Of Cloud Data Using Hybrid Encryption Algorithm, Journal of Ambient Intelligence and Humanized Computing, 2019. https://doi.org/10.1007/s12652-019-01403-1

Sudhan Murugan Bhagavathi, Anitha Thavasimuthu, Aruna Murugesan, Charlyn Pushpa Latha George Rajendran, A Vijay, Raja Laxmi, Rajendran Thavasimuthu, Weather forecasting and prediction using hybrid C5.0 machine learning algorithm International Journal of Communication Systems, Vol. 34, Issue. 10, Pp. e4805, 2021.

Teena Jose, Dr. Suvanam Sasidhar Babu – Detecting Spammers on Social Network Through Clustering Technique, Journal of Ambient Intelligence and Humanized Computing, pp.1-15, 2019. https://doi.org/10.1007/s12652-019-01541-6

Vijayalakshmi Y, Manimegalai P, Suvanam Sasidhar Babu, Accurate Approach towards Efficiency of Searching Agents in Digital Libraries using Keywords, Journal of Medical Systems, 43:164, 2019. https://doi.org/10.1007/s10916-019-1294-5