Prediction of CardioVascular Disease (CVD) using Ensemble Learning Algorithms

C. Oswald
Postdoctoral Fellow, CSE
IIT Kanpur
oswaldc@cse.iitk.ac.in

Gadi Jaya Sathwika
Undergraduate CSE
IIITD&M Kancheepuram
coe18b019@iiitdm.ac.in

Arnab Bhattacharya
CSE, IIT Kanpur
arnabb@cse.iitk.ac.in

ABSTRACT

Over the last decade, CardioVascular Diseases (CVD) and allied heart disorders have been the leading cause of death world wide. Early prediction of CVD can help high-risk patients make lifestyle changes and as a result can reduce complications. Researchers in the past have worked on developing computational models to aid health care professionals in the prediction of CVD. Most of the existing techniques lack precise feature sets and suffer from high overfitting and low accuracy. To present more accurate model of predicting CVD along with employing better feature set, ensemble learning models along with individual classification techniques are proposed. Extensive performance analyses on Kaggle Cleveland Heart Disease dataset clearly show our model can significantly improve on the accuracy and F1-score than some of the existing competitors.

KEYWORDS

CardioVascular Disease, Accuracy, Ensemble learning, Feature Importance

ACM Reference Format:
C. Oswald, Gadi Jaya Sathwika, and Arnab Bhattacharya. 2022. Prediction of CardioVascular Disease (CVD) using Ensemble Learning Algorithms. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD 2022), January 8–10, 2022, Bangalore, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3493700.3493747

1 INTRODUCTION AND RELATED WORK

An early prediction of the CVD can help us to overcome many health issues and save human lives. Predicting the outcome of a disease using a computational model is one of the most interesting and complex task which experts in the healthcare domain can use it for. In the healthcare industry, Data Mining and Machine Learning have been used to define appropriate treatment approaches, anticipate illness risk factors, and determine effective patient care cost structures. Existing research work focussing on predicting CVD can be found from [2, 3, 5] and prediction of heart related diseases can be seen in [1, 6–8]. To the best of our knowledge, most of the existing methods lack better feature selection models, testing with multiple train-test splits and also have not incorporated cross validations for better performance through a high accurate model, which lead to the motivation of this work. Also they focused on a standalone Machine Learning model to predict heart diseases. In this work, an attempt is made to improve accuracy of the prediction model using various ensemble learning methods. The methodology aims to predict CVD and compare the efficiency of combining the results of hybrid models along standalone models. Our novel approach differs from others in that, we use efficient feature selection and ensemble methods for designing a high accurate prediction model for CVD, to reduce the possibility of overfitting.

CODS-COMAD 2022, January 8–10, 2022, Bangalore, India

ACM ISBN 978-1-4503-8582-4/22/01.
https://doi.org/10.1145/3493700.3493747

2 PROPOSED ARCHITECTURE

The purpose of this study is to predict Cardiovascular Disease using various ensemble methods in Machine Learning. The benchmark dataset we considered contains 303 records, 14 attributes and 2 class labels and is preprocessed accordingly. As given in Figure 1, we used the Univariate feature selection method to choose the relevant features. We use feature importance method as it helps in better data comprehension, better understanding of the model by reducing the number of input features to output the features that are more relevant to the target attribute. The data is split into

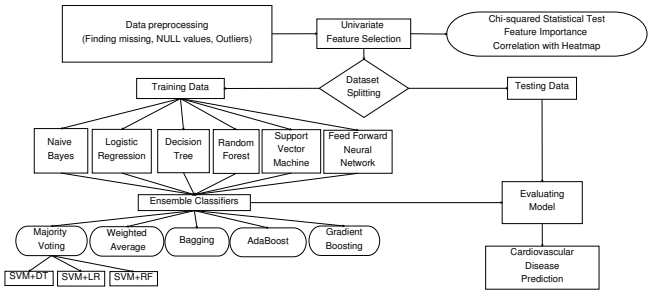
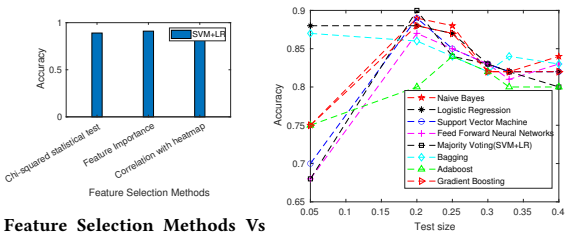


Figure 1: Our proposed Architecture for CVD prediction

training and testing sets using 5-fold and 10-fold cross validation along 5, 20, 25, 30, 33, 40 and 50 percent test splits on which we ran six individual classification models as given in Figure 1. Since, ensemble learning helps in reducing the variance component of prediction errors and increases the predictive power of the model, it is decided to use various ensemble learning methods such as: Majority Voting, Weighted Average, Bagging, and Boosting Methods. The predictions from multiple models are combined in a voting ensemble which can be used to classify or predict data. Majority Voting helps to improve model performance, Bagging helps in lowering the variance and Gradient Boosting improves the efficiency of an algorithm by removing overfitting.

3 RESULTS AND DISCUSSIONS

We used the Cleveland heart disease dataset [4] from Kaggle repository to train and test our model using various ensemble learning methods. Chest pain type, Thalach (maximum heart rate achieved), Exang (exercise induced angina), Oldpeak (ST depression induced by exercise relative to rest), Number of major vessels (0-4) colored by fluoroscopy and Thalassemia are the 6 best features selected by our model. We assessed using Accuracy, Precision, Recall, F1-score, and AUC where Fig. 2(a) compares the best accuracies obtained with the various feature selection methods. 5- fold and 10-fold cross-validation are used to avoid overfitting and the results are compared using various test sizes and found that 5-fold is doing better. We splitted the dataset for testing into 5, 20, 25, 30, 33, 40 percents and compared the efficiency of our model as shown in Fig. 2(b). The time taken for the model in 5, 20, 25, 30, 33 and 40 percentage test splits are 0.06, 0.08, 0.09, 0.09, 0.09 and 0.1 sec. respectively. As per Fig. 2(b), we observe that most of our methods perform well in the 80:20 ratio with SVM+LR in Majority Voting giving highest accuracy of 0.9123. This is due to the fact that SVM works well with unstructured, semi-structured data and Logistic Regression is easier to implement, interpret and train. Fig. 3 depicts a compari-



(a) Feature Selection Methods Vs Accuracy

(b) Test size Vs Accuracy

Figure 2: Accuracy of our proposed approach

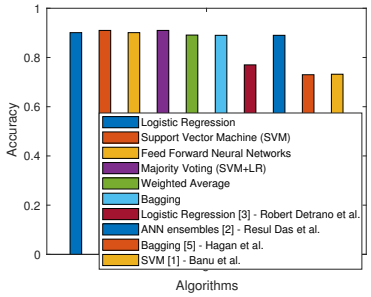


Figure 3: Performance of Proposed approach (5-fold cross validation) with existing algorithms

son of the performance of our approach after cross-validation with the existing competitors. SVM+LR gave the highest accuracy and this is due to the fact of using feature importance in our feature selection methods and ensemble learning. Our proposed LR, SVM, FFNN and Bagging with novel feature importance has achieved significant accuracy of 0.901, 0.91, 0.901 and 0.89 which is higher than [3], [2], [1] and [5].

4 CONCLUSIONS AND FUTURE WORK

To predict CVD, we observe that ensemble learning methods with efficient feature set creation performs better than the existing models. A possible extension to this could be the use of various deep learning models with hyper parameter tuning. Since heart diseases are affected significantly by factors such as sleep disorder conditions, stress mismanagement conditions and pollution factors, we could potentially include these factors as features.

REFERENCES

[1] NK Salma Banu and Suma Swamy. 2016. Prediction of heart disease at early stage using data mining and big data analytics: A survey. In *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE, 256–261.

[2] Resul Das, Ibrahim Turkoglu, and Abdulkadir Sengur. 2009. Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications* 36, 4 (2009), 7675–7680.

[3] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology* 64, 5 (1989), 304–310.

[4] Heart Disease. 1988. Dataset. https://www.kaggle.com/ronitf/heart-disease-uci.

[5] Rachael Hagan, Charles J Gillan, and Fiona Mallett. 2021. Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked* 24 (2021), 100606.

[6] Mai Shouman, Tim Turner, and Rob Stocker. 2011. Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*. 23–30.

[7] Evanthia E Tripoliti, Theofilos G Papadopoulos, Georgia S Karanasiou, Katerina K Naka, and Dimitrios I Fotiadis. 2017. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Computational and structural biotechnology journal* 15 (2017), 26–47.

[8] Yanwei Xing, Jie Wang, Zhihong Zhao, et al. 2007. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)*. IEEE, 868–872.