Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

CARDIOVASCULAR DISEASE AND RELIEF AND LASSO FEATURE SELECTION TECHNIQUES USING ML

Dr.D.ANUSHA, Associate Professor **,**

M. DIVYA,M.NANDU,G.TIRUMALA VASU, J.YUGANDHAR,

Department of CSE, SRK Institute of Technology, Vijayawada, A.P., India.

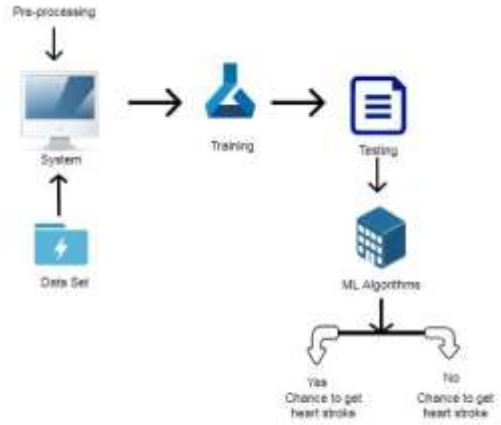
,

ABSTRACT: Cardiovascular disease more commonly known as heart disease is a class or type of illness that involves blood vessels such as the veins, arteries and capillaries, heart or all. The diseases that affect the cardiovascular system of the body are cardiac disease,vascular diseases of the brain and kidney, peripheral arterial disease. A number of diseases affect the heart and the blood vessels, they are as Angina, Arrhythmia, Congenital Heart Disease, Coronary Artery Disease CAD, Heart Attack, Heart Failure, Pulmonary Stenosis, Atherosclerosis, Renal Artery Disease, Stroke, Blood clots, Aneurism. CVDs may be prevented or mitigated by early diagnosis, and this may reduce mortality rates. Identifying risk factors using machine learning models is a promising approach. We‘rewould like to propose a model that incorporates different methods to achieve effective prediction of heart disease. For our proposed model to be successful, we have used efficient Data Collection, Data Pre-processing and Data Transformation methods to create accurate information for the training model. We have used a UCI Heart Disease dataset. The results are shown separately to provide comparisons. Based on the result analysis, we can conclude that our proposed model produced the highest accuracy while using RFBM and Relief feature selection methods.

INTRODUCTION

Cardiovascular diseases CVD are among the most common serious illnesses affecting human health. The increased rate of cardiovascular diseases with a high mortality rate is causing significant risk and burden to the healthcare systems worldwide. Cardiovascular diseases are more seen in men than in women particularly in middle or old age although there are also children with similar health issues According to data provided by the WHO, one-third of the deaths globally are caused by the heart disease. CVDs cause the death of approximately 17.9 million

people every year worldwide and have a higher prevalence in Asia. The European Cardiology Society (ESC) reported that 26 million adults worldwide have been diagnosed with heart disease, and 3.6 million are identified each year. Roughly half of all patients diagnosed with Heart Disease die within just 1-2 years and about 3% of the total budget for health care is deployed on treating heart disease. To predict heart disease multiple tests are required. Lack of expertise of medical staff may results in false predictions. Early diagnosis can be difficult. Surgical treatment of heart disease is challenging, particularly in developing countries which lack trained medical staff as of risk factors which meet the three criteria like the high prevalence in most populations; a significant impact on heart diseases independently; and they can be controlled or treated to reduce the risks.



Different researchers have included different risk factors or features while modelling the predictors for CVD. Features used in the development of CVD prediction models in different research works include age, sex, chest pain (cp), fasting blood sugar (FBS) – elevated FBS is linked to Diabetes, resting electrocardiographic results (Restecg), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope, number of major vessels coloured by fluoroscopy (ca), heart status (thal), maximum heart rate achieved (thalach), poor diet, family history, cholesterol (chol), high blood pressure, obesity, physical inactivity and alcohol intake.Recent studies reveal a need for a minimum of 14 attributes for making the prediction accurate and reliable .Current researchers are finding it difficult to combine these features with the appropriate machine learning techniques to make an accurate prediction of heart disease. Machine learning algorithms are most effective when they are trained on suitable datasets. Since the algorithms rely on the consistency of the training and test data, the use of feature selection techniques such as data mining, Relief selection, and LASSO can help to prepare the data in order to provide a more accurate prediction. Once the relevant features are selected, classifiers and hybrid models can be applied to predict the chances of disease occurrence. Researcher have applied different

techniques to develop classifiers and hybrid models. There are still a number of issues which may prevent accurate prediction of heart disease, like limited medical datasets, feature selection, ML algorithm applications, and a lack of in depth analysis.Our research aims

Prediction. Various available public data sets are applied. In the study of Latha and Jeeva .ensemble technique was applied for improved prediction accuracy. Using bagging and boosting techniques, the accuracy of weak classifiers was increased, and the performance for risk identificationof heart disease was considered satisfactory. They used the majority voting of Naïve Bayes, Bayes NetMultilayer Perceptron, PART and Random Forest (RF) classifiers in their study for the hybrid model development. An accuracy of 85.48% was achieved with the designed model. More recently machine learning and conventional techniques like RF, Support Vector Machine (SVM), and learning models were tested on the UCI Heart Disease dataset. The accuracy

LITERATURE SURVEY

M. S. Oh and M. H. Jeong, ‘‘Sex differences in cardiovascular disease risk factors among Korean adults,’’ Korean J. Med., vol. 95, no. 4, pp. 266–275, Aug. 2020

The socioeconomic status (SES) and health behaviors of workers are associated with the risks of developing obesity, diabetes, hypertension, hyperlipidemia, and other cardiovascular diseases. Herein, we investigated the factors influencing cardiovascular disease (CVD) risk based on the SES of male and female workers. This cross-sectional analysis used the National Health Information Database to assess the associations between gender, SES (income level, residential area), health behaviors, and CVD-related health status of workers, through multinomial logistic regression. Upon analysis of a large volume of data on workers during 2016, the smoking and drinking trends of male and female workers were found to differ, causing different odds ratio (OR) tendencies of the CVD risk. Also, while for male workers, higher ORs of obesity or abdominal obesity were associated with higher incomes or residence in metropolitan cities, for female workers, they were associated with lower incomes or residence in rural areas. Additionally, among the factors influencing CVD risk, lower income and residence in rural areas were associated with higher CVD risk for male and female workers. The study findings imply the importance of developing gender-customized intervention programs to prevent CVD, due to gender-specific associations between CVD-related

health status and health behaviors according to SES.

In the present study, we investigated the associations between SES and CVD-related health status for male and female workers. When comparing the associations for male and female workers, both differences and similarities were observed. The main observations follow. First, the smoking and drinking tendencies varied between male and female workers. Second, while the OR of obesity and abdominal obesity was higher for men with higher incomes or men residing in metropolitan cities, conversely, it was higher for lower income women or women residing in rural areas. Third, the prevalence of belonging to the prehypertension group and the prediabetes group was higher for both male and female workers than all adults (the entire population over 30 years old). In addition, for both male and female workers, as age increased or income decreased, the OR of CVD risk increased for obesity, abdominal obesity, hypertension, high FG, hypercholesterolemia,
hypertriglyceridemia, low HDL-C, high LDL-C, when subjects resided in a rural area, were current smokers, engaged in walking 34 times per week or less, did moderate exercise 12 times per week or less, or did not engage in vigorous exercise.Previous studies based on large volumes of data, which could be compared to the findings of this study, focused on all adults alone rather than on workers. However, some studies have been conducted among workers in a particular region or at a particular place of business. Herein, the trends of health behavior and CVD-related health status were found to vary in male workers and female workers, and this was the major finding of the present study. In a study that analyzed the smoking and drinking behavior of workers using the Korean Working Conditions Survey (KWCS), the current smoking rate of men was found to be high among teens and middle-aged men; however, a constant decrease was found in women with increase in age, supporting the findings of this study. The period when an individual quits smoking and the amount of smoking performed before an individual quits are the major factors affecting the decrease in CVD risk 32%. The presence of a disease, such as diabetes or hypertension, also affects the decrease in the CVD risk for men 32%; however, overall smoking causes a higher CVD risk for women than men 33%. Based on such results, unlike for women, the OR of CVD risk in male workers who were previous smokers decreased in this study, which may be due to a combination of the period when an individual quits smoking, the amount of smoking performed before

quitting, and their health condition. Therefore, besides encouraging smoking cessation programs for current smokers, establishing smoking cessation continuing education programs for men and smoking prevention programs for women is also worthwhile.Lee and Jeon reported that excessive drinking (16 days per month) is highest among male workers in their 40s and 30s and female workers in their 20s. However, this tendency decreased with age for females, similar to the findings of this study. Such a finding also corresponds to the results of a study reporting that the lowest occurrence of cardiovascular events is at a light-to-moderate drinking level. On the other hand, Corral et al. reported that the maximum cardio protective effect of alcohol consumption is 72 g/day, and that an alcohol consumption of 89 g/day increased the risk of developing CVD

Summary:To the best of our knowledge, this study is the first global attempt to use a large volume of data from a nationwide database to determine the differences and similarities in CVD-related health status between male and female workers owing to SES.

D. C. Yadav and S. Pal, ‘‘Prediction of heart disease using feature selection and random forest ensemble method,’’ Int. J. Pharmaceutical Res., vol. 12, no. 4, 2020

The heart is very soft and sensitive part of body by which brain handles blood related system in body. The heart disease that greatly affects in body as like: pulmonary artery, atalata, enzaina and birth defects included. Heart disease is mainly related to contraction or blocked blood vessels in the heart. The symptoms of heart disease depend on the type of disease. Heart disease occurs not only in adults but also in children. The infection affecting the tissues is known as percarditis. In this, the tissues closest to the heart are affected. Infections affecting the lining of the heart muscle are known as myocardium .The study of medical datasets is made very intuitive by machine learning algorithms. The machine learning algorithms provide techniques to identify dataset attributes and the relationship between them. In this research work, we used heart disease related information from UCI repository. The dataset contained 1025 Instances with 14 attributes, sick and nonstick patients in target variable. In this paper, we proposed and analyzed classification accuracy, precision and sensitivity by four tree based classification algorithms: M5P, random Tree and Reduced Error Pruning with Random forest ensemble method. All the prediction based algorithms have applied after the features selection of heart patient’s dataset. In this paper, we

used three features based algorithms: Pearson Correlation, Recursive Features Elimination and Lasso Regularization. The data table analyzed by different feature selection methods for better prediction. All the analysis is done by three experimental setup; First experiment applied Pearson Correlation on M5P, random Tree, Reduced Error Pruning and Random forest ensemble method. In the second experiment we used Recursive Features Elimination and applied on above four tree based algorithms. In the third experiment we used Lasso Regularization and applied on as above tree based algorithms. After all the performance we analyzedand calculated classification accuracy, precision and sensitivity. With the results, we finally concluded that feature selection methods Pearson correlation and Lasso Regularization with random forest ensemble method provide better results 99% accuracy. We analyzed and find the random forest ensemble method predicted better result compare to other algorithms in the previous year's works.

Data is generated by the medical industry. Often this data is of very complex nature electronic records, handwritten scripts, etc. Since it is generated from multiple sources. Due to the Complexity and sheer volume of this data necessitates techniques that can extract insight from this data in a quick and efficient way. These insights not only diagnose the diseases but also predict and can prevent disease. One such use of these techniques is cardiovascular diseases. Heart disease or coronary artery disease (CAD) is one of the major causes of death all over the world. Comprehensive research using single data mining techniques have not resulted in an acceptable accuracy. Further research is being carried out on the effectiveness of hybridizing more than one technique for increasing accuracy in the diagnosis of heart disease. In this article, the authors worked on heart stalog dataset collected from the UCI repository, used the Random Forest algorithm and Feature Selection using rough sets to accurately predict the occurrence of heart disease. Advancement and emergence of newer technologies such as analytics, artificial intelligence, machine learning have impacted many sectors such as health care, automotive etc. In the healthcare sector, these technologies resulted in various benefits such as clinical decision support, better care coordination, improving patient wellness etc. The World over, Coronary Heart Disease (CHD) is affecting millions of people. Various machine learning techniques include ensemble classifiers can be used in healthcare for improving prediction accuracy. This paper analyses various ensemble methods (Bagged Tree,

Random Forest, and AdaBoost) along with Feature subset selection method - Particle Swarm Optimization (PSO), to accurately predict the occurrence of heart disease for a particular patient. Experimental results show that Bagged Tree and PSO achieved the highest accuracy.Heart disease is the deadliest disease and one of leading causes of death worldwide. Machine learning is playing an essential role in the medical side. In this paper, ensemble learning methods are used to enhance the performance of predicting heart disease. Two features of extraction methods: linear discriminant analysis (LDA) and principal component analysis (PCA), are used to select essential features from the dataset.

Summary: In this research work, the important features were identified by Pearson correlation, Recursive Features Elimination Relief feature selection with the selected important features we examine with improved Random Forest.

EXISTING SYSTEM

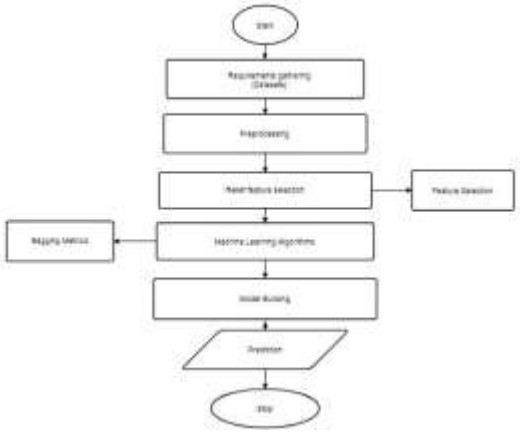
In the existing system, implementation of machine learning algorithms is bit complex to build due to the lack of information about the data visualization and Relief feature selection. Mathematical calculations are used in existing system for model building this may takes the lot of time and complexity. To overcome all this, we use machine learning packages available in the scikit-learn library**.**

Disadvantages:

- High complexity.
- Time consuming.

PROPOSEDSYSTEM

In our proposed model, ten features have been evaluated to make this comparison more unique. The introduced algorithms were conducted based on the all features, Relief selected features the obtained outcomes were compared to other works to show the percentage of improvement, while decrease in performance also noted in one occasion (RFBM, DTBM, KNNBM, ABBM, GBBM). The highest increment was noticed for AB approach as opposed to previous works which was about percentage improvement were calculated for 13 attributes.Cardiovascular disease is used to determine whether or not a patient is at risk of having a heart attack.



Advantages:

- Highest accuracy
- Reduces time complexity.
- Better information on Relief features Selection.

CONCLUSION

In this project an improved and novel method and with a larger dataset for training the model. This research demonstrates that the Relief feature selection algorithm can provide a tightly correlated feature set which then can be used with several machine learning algorithms. The study has also identified that RFBM works particularly well with the high impact features and produces an accuracy, substantially higher than related work. RFBM achieved abestaccuracy with 13 features. Cardiovascular disease is used to determine whether or not a patient is at risk of having a heart attack

FUTURE SCOPE: In the future we aim to generalize the model even further so that it can work with other feature selection algorithms and be robust against datasets where the level of missing data is high.

REFERENCES

[1] C. Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, ‘‘Gender differences in brain-heart connection,’’ in Brain and Heart Dynamics. Cham, Switzerland: Springer, 2020, p. 937.

[2] M. S. Oh and M. H. Jeong, ‘‘Sex differences in cardiovascular disease risk factors among Korean adults,’’ Korean J. Med., vol. 95, no. 4, pp. 266–275, Aug. 2020.

[3] D. C. Yadav and S. Pal, ‘‘Prediction of heart disease using feature selection and random forest ensemble method,’’ Int. J. Pharmaceutical Res., vol. 12, no. 4, 2020.

[4] World Health Organization and J. Dostupno, ‘‘Cardiovascular diseases: Key facts,’’ vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https:// www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[5] K. Uyar and A. Ilhan, ‘‘Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,’’ Procedia Comput. Sci., vol. 120, pp. 588–593, Jan. 2017.

[6] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, ‘‘A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,’’ Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, ‘‘A comprehensive investigation and comparison of machine learning techniques

in the domain of heart disease,'' in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2017, pp. 204–207.

[8] J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, ''Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data,'' NeuroImage, vol. 28, no. 4, pp. 980–995, Dec. 2005.

[9] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ''Innovative artificial neural networks-based decision support system for heart diseases diagnosis,'' J. Intell. Learn. Syst. Appl., vol. 5, no. 3, pp. 176–183, 2013.