Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithem with Relief and Lasso Feature Selection Techniques

CH.Murali Krishna Yadav [1], M.Akhila[2], K.Manjusha[3], K.Niharika[4], B.Balasri[5],

[1]Asssistant Professor, [2,3,4,5]Student, Department of ECE.
N.B.K.R Insttitute of Science And Technology.Tirupati(Dist),Andhra Pradesh.

Abstract: Cardiovascular diseases (CVD) are among the most common serious illnesses affecting human health. 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 would 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 combined dataset (Cleveland, Long Beach VA, Switzerland, Hungarian and Stat log). Suitable features are selected by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. New hybrid classifiers like Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM) are developed by integrating the traditional classifiers with bagging and boosting methods, which are used in the training process. We have also instrumented some machine learning algorithms to calculate the Accuracy (ACC), Sensitivity (SEN), Error Rate, Precision (PRE) and F1 Score (F1) of our model, along with the Negative Predictive Value (NPR), False Positive Rate (FPR), and False Negative Rate (FNR). 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 (99.05%).

Keywords: Heart disease, machine learning, CVD, relief feature selection, LASSO feature selection, decision tree, random forest, K-nearest neighbors, Ada Boost, and gradient boosting learning.

INTRODUCTION:

Cardiovascular disease has been regarded as the most severe and lethal disease in humans. 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 well as testing equipment and other resources required for proper diagnosis and care of patients with heart problems. An accurate evaluation of the risk of cardiac failure would help to prevent severe heart attacks and improve the safety of patients. Machine learning algorithms can be effective in identifying the diseases, when trained on proper data. Heart disease datasets are publicly available for the comparison of prediction models. The introduction of machine learning and artificial intelligence helps the researchers to design the best prediction model using the large databases which are available. Recent studies which focus on the heart-related issues in adults and children emphasized the need of reducing mortality related to CVDs. Since the available clinical datasets are inconsistent and redundant, proper pre-processing is a crucial step. Selecting the significant features that can be used as the risk factors in prediction models is essential. Care should be taken to select the right combination of the features and the appropriate machine learning algorithms to develop accurate prediction mode. It is important to evaluate the effect 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 result (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. In this study, various supervised models such as AdaBoost(AB), Decision Tree (DT), Gradient Boosting (GB), K-Nearest Neighbors (KNN), and Random Forest (RF) together with hybrid classifiers are applied. Results are com-pared with existing studies.

LITERATURE SURVEY:

Various available public data sets are applied. In the study of Latha and Jeeva [28] ensemble techniquewas 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 Net, C 4.5, MultilayerPerceptron, PART and Random Forest (RF) classifiers in their study for the hybrid model development.

Data are processed such thatthe K-Nearest Neighbors algorithm handles the missing data.The feature selection process is done following the Relief and LASSO. Various machine learning algorithms are implanted using the Bagging and Boosting approaches. The brain-heart connection is characterized by sex- and gender-related differences that tend to modify over an individual's lifetime, thus in relation to age. However, since the need for a gender-specific approach has had growing attention only in the latest years, this issue has not yet been fully elucidated. The knowledge gap is especially marked for pathologies that have historically been considered pertaining mostly to men, e.g., cardiovascular diseases, or to women, e.g., neuropsychiatric conditions, and it is even more pronounced with regard to the relationship that exists between these dysfunctions. This chapter will present an overview of the current evidence on the sex- and gender-related aspects that could influence the brain-heart connection and the possible effect of aging on such features. Sex- and gender-related aspects will, in particular, be evaluated in regard to individual vulnerability and the risk factor patterns associated with the development and co-occurrence of cardiovascular and neuropsychiatric pathologies; the mechanisms by which the nervous and cardiovascular systems interact with one another; the bidirectional connection between neuropsychiatric disorders and cardiovascular diseases; and the disparities in how cardiovascular and neuropsychiatric conditions are recognized and treated that can affect the course and the co-occurrence of these diseases.

The tight crosstalk between heart and brain is becoming increasingly recognized as the underlying mutual mechanisms are better identified, having a potential impact for clinical approach. Cardiac control is achieved by means of a three-level hierarchical neuronal network (central nervous system neurons, extracardiac-intrathoracic neurons, and intrinsic cardiac nervous system), where all the components work together to fulfil the physiological demands. However, each component of this network can undergo pathologic-mediated changes due to the transduction of altered sensory inputs originating from a deteriorating heart. A key role in the maintenance of cardiovascular homeostasis is played by the autonomic nervous system with its sympathetic and parasympathetic branches, which operate in a reciprocal manner. Heart rate best mirrors the relative balance between these two systems, and especially heart rate variability has emerged as a key parameter that reflects the health status of a given individual. Neural reflexes (i.e., the baroreceptor reflex) and several neuromodulators released from the heart itself or coming from other sites, as well as neurotrophins, also contribute to cardiovascular homeostasis and will be considered in the present chapter. A deeper understanding of heart-brain interactions will facilitate the prompt recognition and management of cardiac diseases, as well as of neurologic disorders associated to heart dysfunction, and, at the same time, will help in optimizing the therapeutic approach.The understanding of cardiac neuronal control has dramatically evolved in the last 50 years, both from an anatomical and a functional point of view. Cardiac neuronal control is mediated via a series of reflex control networks involving somata in the intrinsic cardiac ganglia (heart), intrathoracic extracardiac ganglia (stellate, middle cervical), superior cervical ganglia, spinal cord, brainstem, and higher centers. Each of these processing centers contains afferent, efferent, and local circuit neurons, which interact locally and in an interdependent fashion with the other levels to coordinate regional cardiac electrical and mechanical indices on a beat-to-beat basis. This neuronal control system shows plasticity and memory capacity, allowing it to maintain an adequate cardiac function in response to normal physiological stressors such as standing and exercise.This neuronal control system shows plasticity and memory capacity, allowing it to maintain an adequate cardiac function in response to normal physiological stressors such as standing and exercise.Yet, pathological events such as myocardial ischemia as well as any other type of cardiac stressor may overcome the homeostatic capability of the system, leading to excessive sympathoexcitation coupled with withdrawal of central parasympathetic drive. In turn, autonomic dysregulation is central to the evolution of heart failure and the development of life-threatening arrhythmias. As such, understanding the anatomical and physiological basis for cardiac neuronal control is crucial to implement effectively novel neuromodulator therapies to mitigate the progression of cardiac disease.

METHODOLOGY:

An overall explanation is explained to build an intelligent machine learning system over the dataset of chronic heart diseas Dataset is constructed by combining five different datasets (Cleveland, Hungary, Switzerland, and VA Long Beach and Statlog). This is included in the framework. Fig. 1 illustrates the workflow of recommended models. During data preprocessing, the combined dataset is analyzed to check for missing values which are then dealt with by the K-Nearest Neighbors imputation technique. To overcome overfitting issues and avoid long execution times, two different feature selection techniques are utilized: Relief and LASSO. This assists in extracting the best features. Performance of classifiers with the features selected by these techniques as well as with the original features is analyzed. After feature selection, the dataset is split into two parts: training and testing. Based on model learning rates, 80% of data is assigned for the training phase, and the remaining 20% d for the testing phase. All ensemble models with classifiers are implemented to make a comparison over the combined dataset; however, the generated outcome of our model is gained within a short period. Different training model has been given for testing the dataset so that we can pick the best model for our reliable dataset. The process resulted in RFBM being the most useful with 99.05% of accuracy. Furthermore, the most suitable features of a patient having affected by heart disease have been suggested in this diagnosis system. PERFORMANCE MEASURE INDICES The effectiveness and accuracy of the machine learning method can be evaluated using performance indicators. Positive classification occurs when a person is classified as having HD. When a person is not classified as having HD, he has a negative classification.Having a suitable application of the proposed model is key to the development of this unique system and will also help to deal with the real world challenges. The process has been illustrated in this section. how a community health center can put the system to use, the following steps describes the procedures.

• Step 1: Reports are uploaded into the database.

• Step 2: Attributes are selected from the uploaded data to create input for the trained RFBM model.

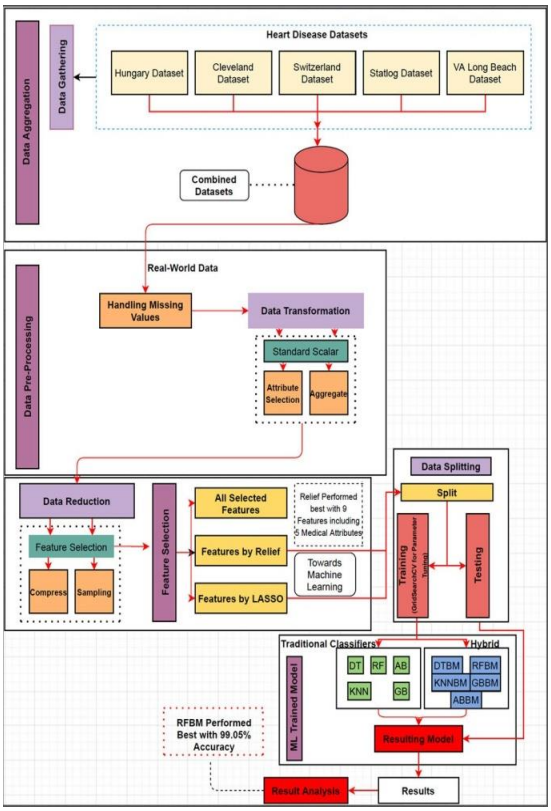


Figure: Working diagram of proposed model.

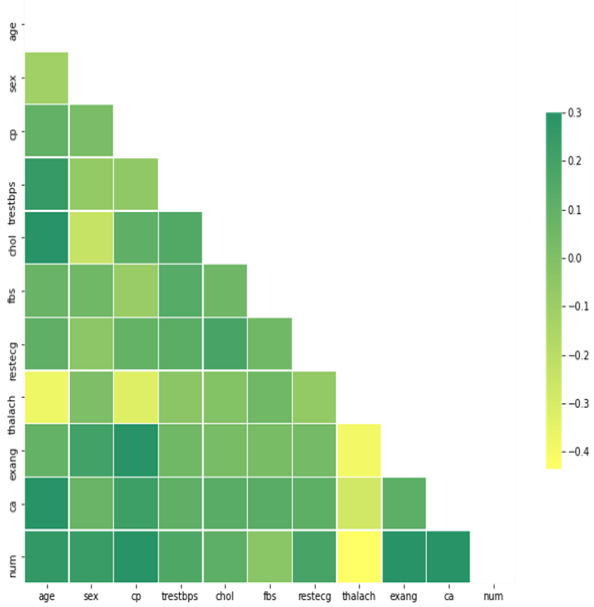
As has been discussed before, previous works that are somewhat related to this study and deal with the datasets used here are available, however, the performance of those systems were not as expected in most cases.

We believe one reason for the lack of performance of some systems is the inability of those systems to identify the mostimportant and highly correlated features. We want to developa method that will first identify the optimal group of featuresand then identify the algorithms that works best with those features.

In our understanding, algorithms that performed well ben-efitted from the tightly correlated feature-set, mainly derivedfrom the use of Relief, whereas the algorithms that did not show strong performance, could not properly evaluate the correlative structure among the features used.

It is clearly seen that ca, chol and trestbps features have strong relationship with age where the value was approx- imately 0.3, on the other hand, the lowest correlation was observed for thalach that was about 0.4. Similarly, cp shows a significant correlation with exang.

Basically, we felt the need to improve the current studies in this field and analyzed previous models to determine what might be lacking, after which we took the initiative to devisea solution that might reshape the current ideas and provide anacceptable level of results that makes the system suitable forpractical implementation.



IMPLIMENTATION:

The implemented model is written in Jupiter notebook's Python programming language using simple libraries like Panda [56], Pyplot [57] and Scikit-learn [58].

The value of the 'num' attribute can be 0, 1, 2, 3 or 4. Thepredicted value '0' represents that a patient does not have heart disease and the values from 1 to 4 reflect the various stages of chronic heart disease.

There is a large amount of collected data in the modern world that can be gathered via the internet, surveys, and experiments, etc. Often the data to be used contain missing values, noise, and distortions, however. The combined dataset used for this research also contains missing or null values.

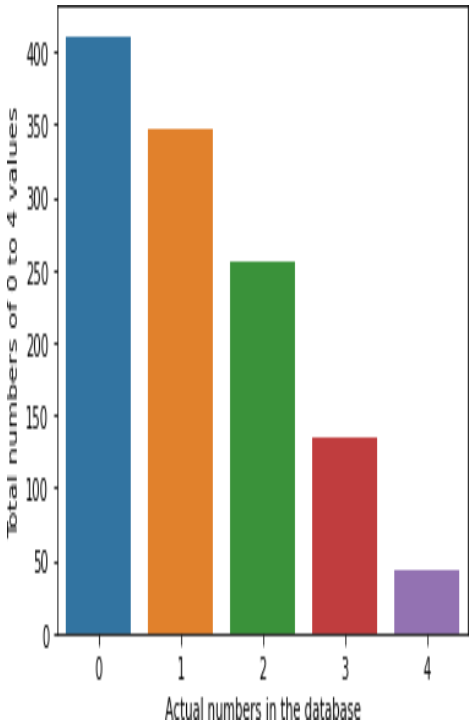


Figure:Actual data points in the datasets

There are some popular techniques, such as imputation and deletion that can be used to deal with missing values. Feature selection techniques are important for the machine learning procedure as the best attributes for classification need to be extracted. This also helps to reduce the execu- tion time. We have selected two algorithms: Relief feature selection and the Least Absolute Shrinkage and Selection Operator. . These weights can then be modified gradually [64]. The aim is to ensure that the impor-tant features have a large and that the remaining features have low weights.

Relief uses the similar techniques as in KNNto determine feature weights. This well – known algorithm of feature selection approaches has been shown by Kira andRendell [65]. $\boldsymbol{R_i}$ is for a randomly selected instance. Relief searches for its two nearest neighbours: one from the same class, called closest hit $\boldsymbol{H}$, and one from the opposite class, called closest miss $\boldsymbol{M}$. It adjusts the consistency calculation $\boldsymbol{W\ [A]}$ for feature $\boldsymbol{A}$ according to the $\boldsymbol{R_i}$, $\boldsymbol{M}$, and $\boldsymbol{H}$ values.If there is a large difference between $\boldsymbol{R_i}$ and $\boldsymbol{H}$ occur this is not desirable, so the performance value $\boldsymbol{W\ [A]}$ is reduced.

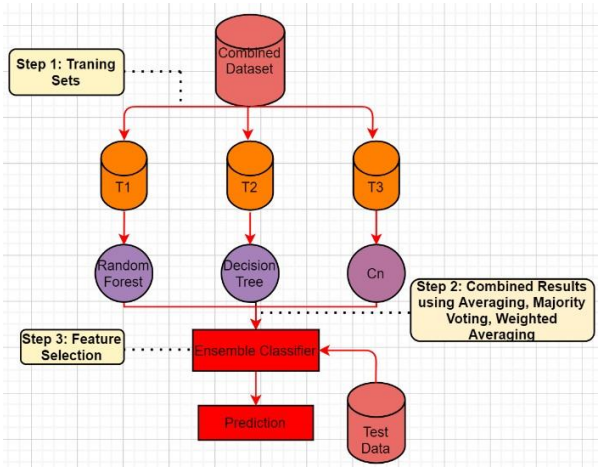


Figure:The working techniques of ensemble process.

RESULT:

The LASSO treats closely related features as true, and therest as false. After applying the LASSO, chest pain (cp) hadthe

highest rank score (0.0796), whereas maximum heart rate(thalach) had a very low score.

Accuracy is usually considered to be the most impor- tant techniques to evaluate machine learning algorithms.As mentioned above, we use five classifiers and five hybrid classifiers. We applied the ten different methods on the origi- nal 13 input features then on the eleven input features selected by the LASSO approach, and on the 10 features selected withthe Relief method. Figure shows the accuracy of the differenttypes of classifiers, including the five hybrid classifiers.

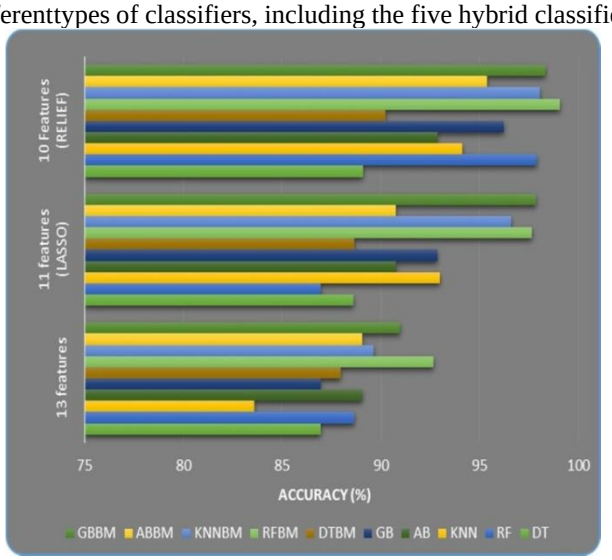


Figure:Accuracy

Looking at the accuracy of these ten strategies with the Relief features, the Random Forest Bagging method(RFBM), which is a hybrid classifier, demonstrated an excel-lent accuracy of 99.05%.

The results of the hybrid modelsof DT, AB, and GB were similar to the previous results.

CONCLUSION:

Identifying the risk of heart disease with reasonably high accuracy could potentially have a profound effect on the long-term mortality rate of humans, regardless of social andcultural background. Early diagnosis is a key step in achiev-ing that goal. Several studies have already attempted to pre-dict heart disease with the help of machine learning. This study takes similar route, but with an improved and novel method and with a larger dataset for training the model. Thisresearch demonstrates that the Relief feature selection algo-rithm can provide a tightly correlated feature set which thencan be used with several machine learning algorithms. The study has also identified that RFBM works particularly wellwith the high impact features (obtained by feature selection algorithms or medical literature) and produces an accuracy, substantially higher than related work. RFBM achieved an accuracy of 99.05% with 10 features.

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.
10. Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J. Comput. Sci., vol. 8, no. 2, pp. 150–154, 2011.