EFFICIENT AND PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHIM WITH RELIEF LASSO FEATURES SELECTION TECHNIQUE

A.NAGARAJU[1], SHAVUKARU SAI[2]

[1] **Assistant Professor MCA,DEPT,** Dantuluri Narayana Raju College, **Bhimavaram, Andhra Pradesh**
Email id:- nagaraju.dnr345@gmail.com
[2]**PG Student of MSc Computer Science,** Dantuluri Narayana Raju College, **Bhimavaram, Andhra Pradesh**
Email id :- shavukarusai23171@gmail.com

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 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%).

1 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-NearesNeighbors (KNN), and Random Forest (RF) together with hybrid classifiers are applied. Results are com-pared with existing studies

2. LITERATURE SURVEY AND RELATED WORK

Various available public data sets are applied. In the study of Latha and Jeeva [28] 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 identification of heart disease was considered satisfactory. They used the majority voting of Naïve Bayes, Bayes Net, C4.5, Multilayer Perceptron, PART and Random Forest (RF) classifiers in their study for the hybrid model development. Data are processed such thatthe K-Nearest Neighbors algorithm handlesthe 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.

3 EXISTING SYSTEM

Conditions related to the cardiovascular system (also known as CVDs) are the biggest cause of mortality according to WHO standards, accounting for 17.9 million deaths each year. The term "cardiovascular diseases" aka CVDs refers to a set of illness that affect both the heart and blood arteries. We collected, pre-processed, and transformed correct data for the training model. Relief and LASSO are used to select features. The highest accuracy value was 63.92% when feature selection was not used, however this value could be raised to 88.52% by employing backward feature selection in conjunction with a decision tree classifier. 78% accuracy has been achieved by the Relief-based selection technique. Top of Form Bottom of Form. The highest accuracy value was 63.92% when feature selection was not used, however this value could be raised to 88.52% by employing backward feature selection in conjunction with a decision tree classifier. According to the results of the experiments, using feature selection method is likely capable of accurately categorizing the condition using only a limited amount of features. Keywords Heart disease Cardiovascular disease dataset Machine learning Data mining LASSO feature selection Relief feature selection Classification algorithms AdaBoost Support vector machines RFKNN Decision tree Gradient boosting.

PROPOSED WORK AND ALGORITHM

The aim of this research is to develop an effective method to predict heart disease, in particular Coronary Artery Disease or Coronary Heart Disease, as accurately as possible. Required steps can be summarized as follows:

1) Five datasets are combined to develop a larger and more reliable dataset.

2) Two selection techniques, Relief and LASSO, are utilised to extract the most relevant features based on rank values in medical references. This also helps to deal with over fitting and under fitting problems of machine learning.

3) Additionally, various hybrid approaches, including Bagging and Boosting, are implemented to improve the testing rate and reduce the execution time. 4) The performance of the different models is evaluated based on the overall results with All, Relief, and LASSO selected features.

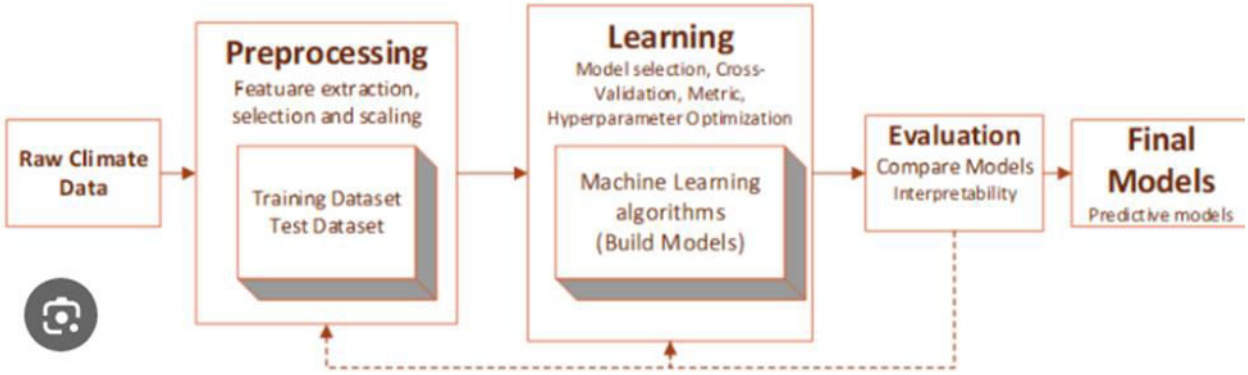

FIG 1: SYSTEM ARCHITECTURE

4 METHODOLOGIES

MODULES:

Data Collection:

Gather a comprehensive dataset with relevant features related to cardiovascular health. This data may include patient demographics, medical history, lifestyle factors, and diagnostic test results.

Data Preprocessing:

Clean the data by handling missing values, encoding categorical variables, and normalizing or scaling numerical features.

Feature Selection:

Relief Feature Selection: Implement the Relief algorithm to rank and select relevant features based on their importance in predicting cardiovascular disease.

Lasso Feature Selection: Utilize Lasso regression to further refine feature selection by penalizing and shrinking less important features to zero.

Data Splitting: Divide the dataset into training and testing sets to evaluate the model's performance

Machine Learning Algorithms:

Choose appropriate machine learning algorithms for classification tasks. Common choices include logistic regression, decision trees, random forests, support vector machines, and gradient boosting methods like XGBoost or LightGBM.

Model Training:

Train the selected machine learning models on the training dataset using the reduced feature set obtained from Relief and Lasso.

Hyperparameter Tuning:

Fine-tune the hyperparameters of the chosen algorithms to optimize their performance. Techniques like grid search or random search can be used.

Model Evaluation:

Evaluate the models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC on the testing dataset to measure their predictive performance.

Cross-Validation:

Apply cross-validation techniques (e.g., k-fold cross-validation) to ensure the model's robustness and reduce overfitting.

Ensemble Methods:

Consider using ensemble methods like stacking or bagging to further enhance model performance.

Interpretability:

Interpret and visualize the model's results to understand the factors contributing to cardiovascular disease prediction.

Deployment:

Once satisfied with the model's performance, deploy it in a healthcare setting, ensuring compliance with regulatory and ethical standards.

Continuous Monitoring:

Continuously monitor and update the model to adapt to changing data patterns and improve accuracy.

User Interface:

Develop a user-friendly interface for healthcare professionals to input patient data and receive predictions.

Documentation:

Document the entire process, including data sources, preprocessing steps, model selection, and performance metrics, for transparency and reproducibility.

Ethical Considerations:

Ensure the responsible use of the model, considering privacy, bias, and fairness concerns.

Remember that the effectiveness of the model depends on the quality and representativeness of the data, as well as the choice of algorithms and feature selection techniques. Additionally, it's crucial to involve domain experts and healthcare professionals throughout the development process to ensure clinical relevance and accuracy

5 RESULTS AND DISCUSSION SCREENSHOTS

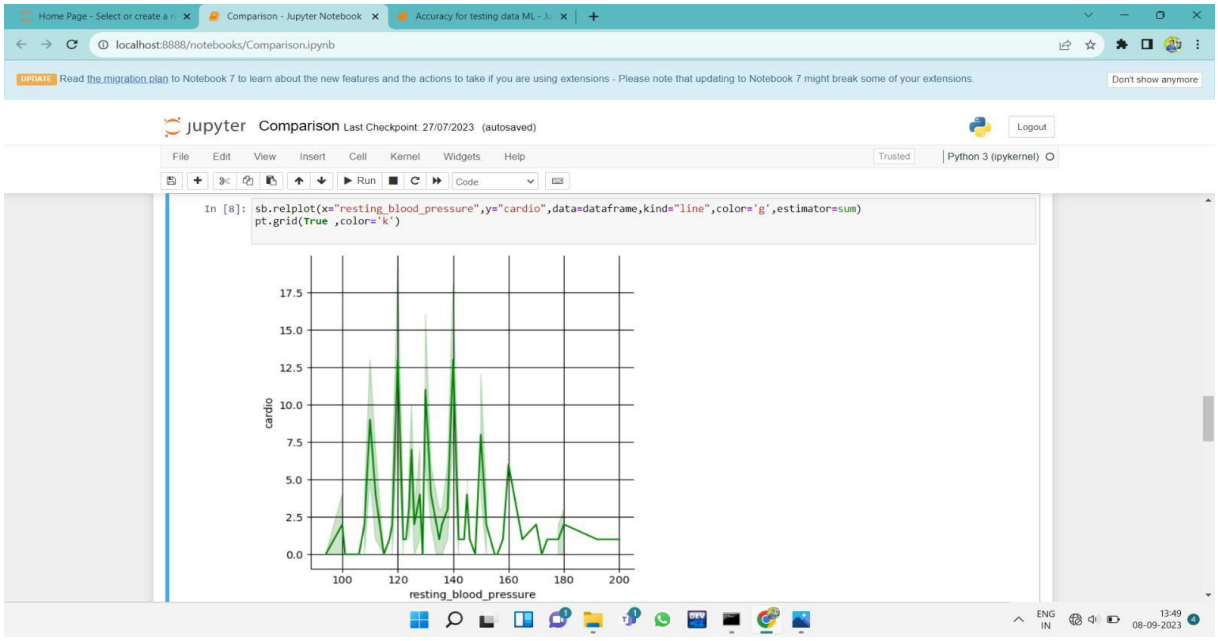

Fig 2 :- graph representing about cardio and blood pressure

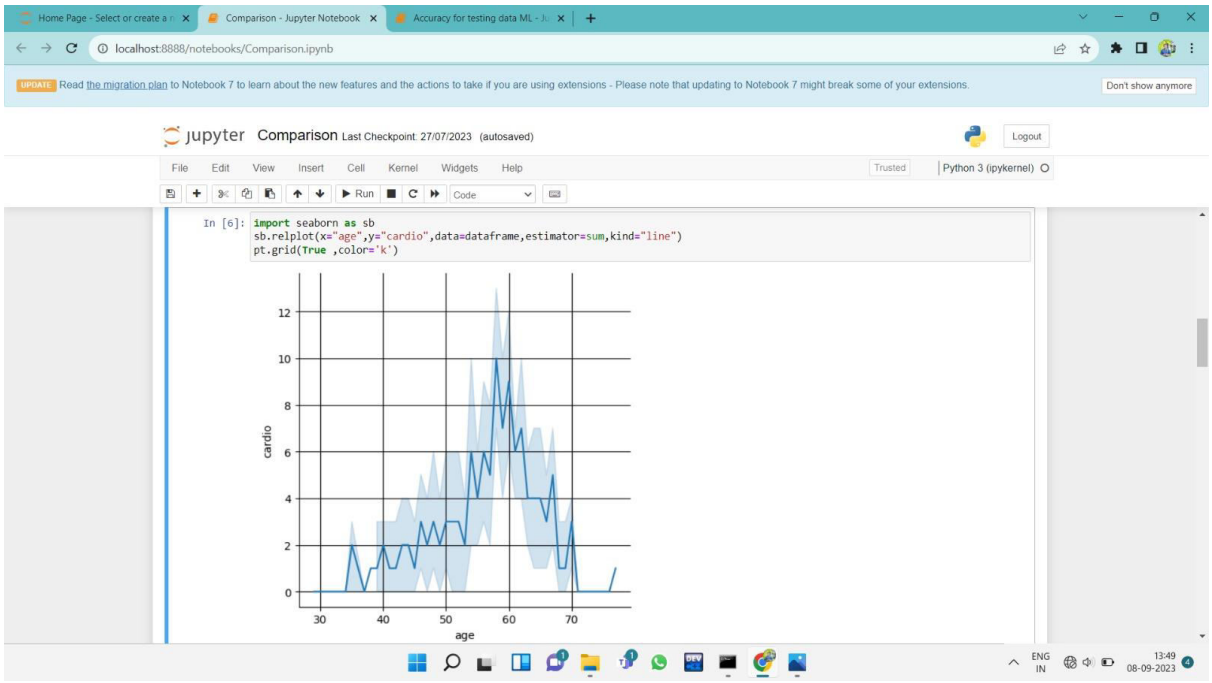

Fig 3 :- graph representing about age and cardio

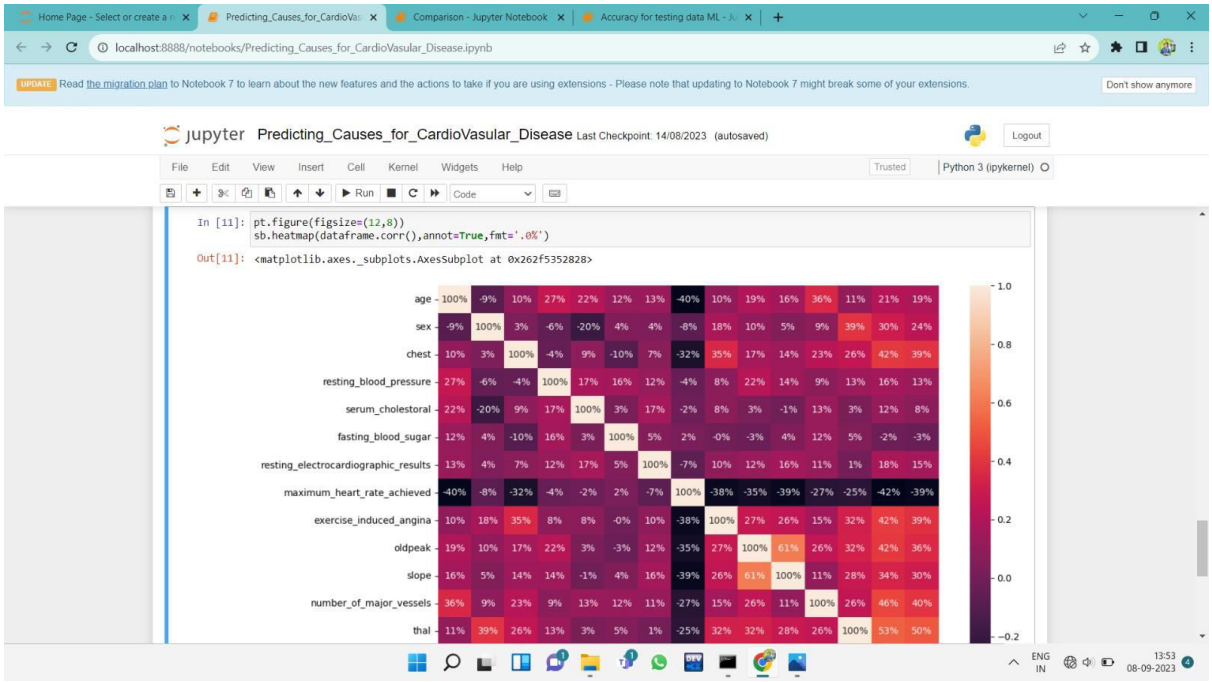

Fig 4 :- corleation matrix

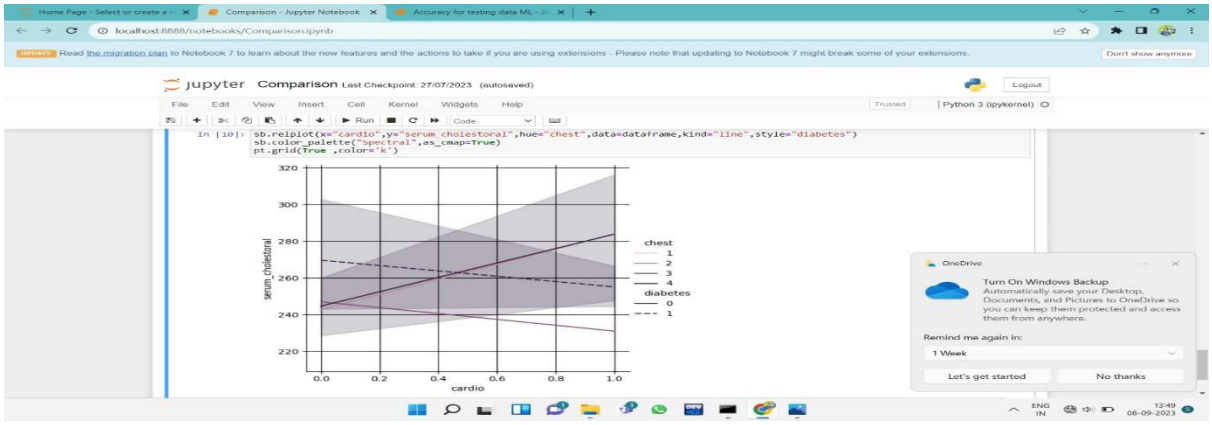

Fig 5 :- graph representing chest and diabties

6.CONCLUSION AND FUTURE SCOPE

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 and cultural background. Early diagnosis is a key step in achieving that goal. Several studies have already attempted to predict 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. 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 (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. 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. The application of Deep Learning algorithms is another future approach. The primary aim of this research was to improve upon the existing work with an innovative and novel way of building the model, as well as to make the model useful and easily implementable to practical settings.

Future scope:

The overall discussion has shown that the performance of different classifiers were good enough in comparison to previous studies, however, there are indeed few limitations, such as, the dependency on a specific Feature Selection technique, for instance more reliance on Relief in this case to produce highly accurate results. Additionally, high level of missing values in the dataset can have an adverse effect.

We have demonstrated how to address the issue through the proper methods, and therefore other dataset when used with this model, must also take care of this issue if the missing value is quite significant.Furthermore, though our training The overall discussion has shown that the performance of different classifiers were good enough in comparison to previous studies, however, there are indeed few limitations, such as, the dependency on a specific Feature Selection technique, for instance more reliance on Relief in this case to produce highly accurate results. Additionally, high level of missing values in the dataset can have an adverse effect. We have demonstrated how to address the issue through the proper methods, and therefore other dataset when used with this model, must also take care of this issue if the missing value is quite significant. Furthermore, though our training dataset is reasonably extensive, larger dataset would make the model more presence.

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.150–155.