PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS WITH RELIEF FEATURE TECHNIQUES

Mrs.K.Karpagavalli Assistant Professor in Department of CSE, Annamacharya Institute Of Technology And Sciences(Autonomous), Rajampet, Andhra Pradesh, India-516126.

D.Sudha[2] , C.Ajay Kumar[3] , C.Sushmitha[4] , K.Schithra [5] , J.Naveen Kumar [6] Department of Computer Science and Engineering, Annamacharya Institute Of Technology and Sciences (Autonomous), Rajampet, Andhra Pradesh, India, 516126.

ABSTRACT:

One of the most prevalent and deadly diseases affecting human health is cardiovascular disease (CVD). Early diagnosis may allow for CVD mitigation or prevention, which may lower mortality rates. A viable strategy is to locate risk indicators using machine learning algorithms. To obtain accurate cardiac disease prediction, we would want to suggest a model that combines various techniques. We have successfully created accurate data for the training model using effective approaches for data collection, pre-processing, and data transformation. In order to allow for comparisons, the findings are presented separately. Using the RFBM and Relief feature selection approaches, we can infer from the outcome analysis that our suggested model provided the maximum accuracy (99.05%).

KEYWORDS: CVD, heart disease, machine learning, K-nearest neighbours, gradient boosting, decision trees, random forests, and relief feature selection techniques.

INTRODUCTION:

The most serious and fatal disease affecting people has been described as cardiovascular disease. A significant danger and burden is being placed on the healthcare systems around the world by the rise in cardiovascular illnesses with high death rates. Although children can also have similar health problems, men are more likely than women to develop cardiovascular disorders, especially in middle or late life.

One-third of deaths worldwide are attributable to heart disease, according to data supplied by the WHO. About 3% of the overall health care budget is spent on treating heart disease, and roughly half of all patients with heart disease pass away within just 1-2 years of diagnosis. Multiple tests are necessary to predict cardiac disease. False projections could be caused by a lack of medical staff experience. It can be challenging to make an early diagnosis. Surgery for heart disease is difficult, especially in underdeveloped nations where there is a dearth of skilled medical personnel, diagnostic equipment, and other resources needed for accurate diagnosis and treatment of heart disease patients.

When taught on relevant data, machine learning algorithms are capable of accurately recognising the disorders. The comparison of prediction models may be done with publicly available datasets on heart disease. Using the vast resources that are accessible, researchers may create the best prediction model with the use of machine learning and artificial intelligence. It has been stressed in recent research that there is a need to lower CVD-related mortality in both adults and children. Proper pre-processing is an important step since the available clinical datasets are inconsistent and redundant. It is crucial to choose the key aspects that can be included as risk variables in prediction models.

Hybrid classifiers are used along with a variety of supervised models, including AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB),
K-Nearest Neighbours (KNN), and Random Forest (RF). Results are compared with those from earlier research.

LITERATURE REVIEW:

The predictions can now be made with greater accuracy and efficiency, the use of artificial intelligence and machine learning algorithms has grown significantly in recent years. In order to create and choose models with the maximum accuracy and efficiency, research in this field is crucial. Hybrid models, which combine several machine learning models with information systems (important elements), are a potential method for illness prediction. Different publicly accessible data sources are utilised. The ensemble approach was used in the study to increase prediction accuracy. The accuracy of weak classifiers was improved by the use of bagging and boosting approaches, and the performance for heart disease risk detection was deemed good. In the research for the construction of the hybrid model, The created model had an accuracy of 85.48%. Recently, the UCI Heart Disease dataset was used to evaluate machine learning as well as more traditional methods like RF, Support Vector Machine (SVM), and learning models. The voting- based approach and many classifiers helped to increase accuracy.

Several machine learning classifiers have been used in studies that predict the survival of patients. A comparison between the conventional biostatistics’ tests and the offered machine learning methods was done after features pertaining to the important risk variables were ranked. The conclusion was that the two most important characteristics for precise predictions were found to be serum creatinine and ejection fraction. The AL Algorithm was used to create a CVD detection model. Four methods were used for dataset preparation and analysis. The accuracy was 85.32% and 84.49%, respectively, for SVM and KNN, and 99.83% for Decision Tree, Random Forest, and other approaches. Another study that used the ensemble approach to analyse the Heart Rate Variability successfully predicted congestive heart failure.

RESEARCH METHODOLOGY:

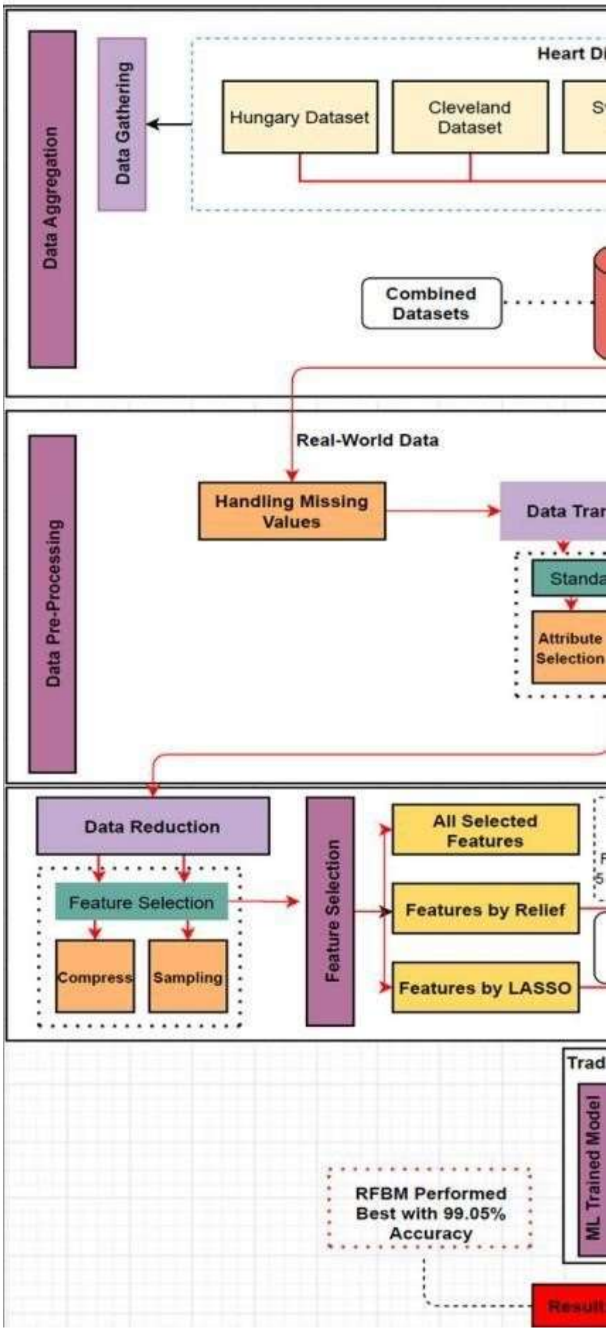
On the dataset of chronic heart disease, a general explanation is provided on how to construct an intelligent machine learning system.

- **Overview of the purposed model:**

In order to create a dataset, many datasets are combined. Here, the workflow of suggested models is provided in the framework. The merged dataset is examined for missing values during data preparation, and those that are found are filled in using the K-Nearest Neighbours imputation method. Two distinct feature selection strategies are employed to address overt difficulties and prevent lengthy execution times: methods for reassurance. In doing so, the best characteristics are extracted. Analysis is done on how well classifiers perform using both the original features and the features chosen by these approaches. The dataset is separated into training and testing after the feature selection process. The remaining 20% of the data are assigned to the testing phase, and the remaining 80% of the data are assigned for the training phase based on model learning rates.

PERFORMANCE MEASURE INDICES

TP = True Positive (when the model correctly Identiedas having HD). **TN** = True Negative (when the model correctly identified the opposite class, such as patients truly having no heart issues).



FP = False Positive (when the model incorrectly identi_ed HD patients i.e., identifying non-HD patients as HD patients)

FN = False Negative (when the model incorrectly identified the opposite class, such as HD patients as normal patients). Accuracy (Acc) = (TP C TN)
(TP +TN+ FP+ FN)
Precision = (TP)
(TP+ FP)
Recall or Sensitivity (Sen) = (TP)
(TP+FN)
F1-score = 2(Precision X Recall)

Fig. working of proposed model

(Precision X Recall)

False Positive Rate = FP
FP+TN

False Negative Rate = FN
(TP+FN)

Negative predictive value = TN
(TN+FN) **IMPLEMENTATION:**

```
BEGIN
1. Let D = {d1, d2, d3, . . . dn} be the given dataset
2. E = {}, the set of ensemble classifiers
3. C = {c1, c2, c3, . . . cn}, the set of classifiers
4. X = the training set, X D
5. Y = the test set, Y D
6. L = n(D)
7. for i = 1 to L do
8. S(i) = {Bootstrap sample I with replacement}
I X
9.   M(i) = Model trained using C(i) on S(i)
10. E = E C(i)
11. next I
12. for i = 1 to L
13. R(i) = Y classified by E(i)
14. next i
15. Result = max(R (i): i = 1, 2, ........... , n)
END
```

DIFFERENT MACHINE LEARNING LIBRARIES

When utilising machine learning approaches to produce reliable outcomes, data is thought of as the first and most fundamental component. A well- known data source, the "UCI machine learning repository," provided the applicable dataset. In addition to 14 unique features, more than 1190 examples from their database are compiled into a text file. The 'num' attribute from these merged datasets is chosen as the output, while 13 attributes are used as the inputs. Age in years (age), sex (sex), resting blood pressure (trestbps), fasting blood sugar (fbs), chest pain type (cp), and resting electrocardiographic findings were the only six characteristics contained in all or the majority of the records, according to the medical literatures.

AN OVERVIEW OF DATA PREPROCESSING AND CLEANING TECHNIQUES:

In the current world, a lot of information is obtained through surveys, tests, and other means, including the internet. But frequently, the needed data may include noise, distortions, and missing values. Missing or null values can also be found in the pooled dataset utilised in this study. To deal with missing values, a number of well-liked strategies may be applied, including imputation and deletion.

Standardization: *X* =(*X*-)/

THE RELIEF FEATURE SELECTION TECHNIQUES:

Relief is a selection attribute approach that weighs all of the dataset's attributes. This allows for progressive weight modifications [64]. The vital features should have a significant weight, while the other features should have a low weight. For feature weighting, Relief employs methods like to those used in KNN.

BAGGING TECHNIQUE:

When it comes to reducing the variance of Decision Tree classifiers, bagging is applied. The goal is to separate the training samples' data into several subgroups. They train their decision tree using sets of subset data that were randomly selected. We end up with an ensemble of many models as a consequence.

Then, the mean of every forecast made by every tree is applied. Compared to only one Decision Tree classifier, this is more reliable. It aids in both the correct handling of larger multidimensional data as well as the reduction of the overt issue. It fixes problems with missing data and upholds correctness**.**

BOOSTING TECHNIQUE:

The recurrent process of "boosting" modifies the weight and is dependent on the most recent forecast. Instances that are improperly classified have their weight enhanced. In most cases, Boosting creates effective prediction models [69]. It works by merging the weak models to improve their performance and creates various loss functions. In order to build our hybrid models for this research, the Boosting approach was used to the two classification algorithms, AB and GB.

DECISION TREES:

One of the most potent and well- known prediction tools is the Decision Tree method, which only supports two num Classes [70]. Every branch of a decision tree's structure relates to a test result, and each leaf node represents a different class. Each internal node refers to testing a property. A approach known as "learning" based on decision trees (DT) sometimes uses an upside- down tree-based process. Both regression and classification issues can be solved using the approach. The best feature or characteristic from the collection of possible attributes is chosen as the root node's starting point, and "splitting" is then used to expand the tree from that point.

RANDOM FOREST:

For the optimal outcome, the Random Forest ensemble classier constructs and combines a variety of decision trees. Assembling bootstraps is mostly used to understand trees.

Let's say that the inputted data are X D x1, x2, x3;::::::: ; xn) and the outputs are Y D x1, x2, x3;::::: ; xn) with a lower limit of b D 1 and an upper limit of B: By averaging the forecasts, the forecast for sample x0 is generated.

K-NEAREST NEIGHBORS:

The most widely used classification method in the field of machine learning is K-Nearest Neighbours. For coronary artery disease, it has previously been used. Since KNN makes no assumptions about how data will be distributed, it is regarded as nonparametric. The new data is placed in the class that is closest to the existing classes by KNN after taking into account their similarity. Both regression and recognition issues can be solved using KNN. Given that it takes some time to process a set of training data, it is also referred to as the lazy learner algorithm [80]. KNN determines how far apart new A (x1, y1) and previously available B(x2, y2) data are from one another.

ADABOOST:

In order to create a more reliable classifier, the AdaBoost or Adaptive Boosting method combines a number of weak classifiers. Based on 1000 samples, this method generates the anticipated accuracy. N is the frequency of training instances, and training instance and training dataset instances are weighted with a starting weight. Each input variable receives an output from the decision stump. After that, an equation is used to determine the misclassification rate.

GRADIENT BOOSTING:

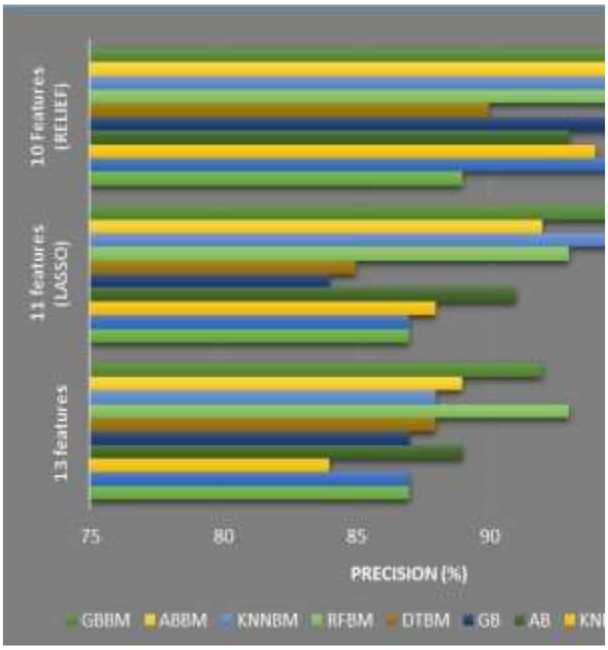
For classification and regression issues, gradient boosting is a Boosting approach that just needed 100 samples. Three components make up the core of gradient boosting: an improved loss function, a weak learner to generate predictions, and an additive model to join weak learners to reduce the loss function. By removing overfitting, the approach of gradient boosting can improve the effectiveness of the algorithm.

When there is an imbalance in the numbers in each class, the use of gradient tree boosting, often known as the "Grabit" model, to the Tobit model aids in improving accuracy.

COMPARISON TABLE BETWEEN THE ACCURACY OF THE PROPOSED MODELS AND EXISTING TECHNIQUES

Significant improvements have been observed after modifying the amount of characteristics that are picked by using selection algorithms. The RFBM hybrid model had the highest accuracy score when data from all features were combined (92.65%), whereas KNN had the lowest accuracy score (83.61%). A few significant modifications result from the use of the LASSO selection technique. The GBBM model

produced the best accuracy (97.85%), whereas the RF model produced the lowest results. With the Relief feature selection approach, the best results were attained With RFBM, it is 99.05%accuracy.



CONCLUSION

Regardless of socioeconomic or cultural background, identifying the risk of heart disease with a reasonable degree of precision has the potential to significantly impact the long-term death rate of people. A crucial element in reaching that aim is early diagnosis. With the aid of machine learning, several research have previously tried to forecast cardiac disease. This work follows a similar path, but uses an enhanced and unique methodology and a larger dataset to train the model. According to this study, the Relief feature selection method may provide a highly correlated feature set that can be applied to a variety of machine learning techniques. The study also found that the high impact features and RFBM function exceptionally effectively together. RFBM has achieved with ten features, accuracy is 99.05%. Future generalisation efforts will focus on making the model more resistant to datasets with high levels of missing data and compatible with other feature selection approaches. One such potential strategy is the use of Deep Learning algorithms. The main goal of this research was to advance previous work by developing the model in a fresh and inventive manner while also making it practical and simple to apply to real-world situations.

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
arti_cial
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, ``Arti_cial neural networks in medical diagnosis," *Int.
J. Comput. Sci.*, vol. 8, no. 2, pp. 150_154, 2011.
[11] F. M. J. M. Shamrat, M. A. Raihan, A. K. M. S. Rahman, I. Mahmud, and

R. Akter, ``An analysis on breast disease prediction using machine learning approaches,'' *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 2450_2455, Feb. 2020.

[12] M. S. Amin, Y. K. Chiam, and K. D. Varathan, ``Identi_cation of significant features and data mining techniques in predicting heart disease,'' *Telematics Informat.*, vol. 36, pp. 82_93, Mar. 2019.

[13] N. Kausar, S. Palaniappan, B. B. Samir, A. Abdullah, and N. Dey, ``Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classi_cation for diagnosis of cardiac patients,'' in *Applications of Intelligent Optimization in Biology and Medicine*. Cham, Switzerland: Springer, 2016, pp. 217_231.

[14] J. Mackay and G. A. Mensah, ``The atlas of heart disease and stroke,'' World Health Org., Geneva, Switzerland, Tech. Rep., 2004.

[15] M. Ashraf, S. M. Ahmad, N. A. Ganai, R. A. Shah, M. Zaman, S. A. Khan, and A. A. Shah, *Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS*. Singapore: Springer, 2021, pp. 239_255.

[16] F. Andreotti, F. S. Heldt, B. Abu-Jamous, M. Li, A. Javer, O. Carr, S. Jovanovic, N. Lipunova, B. Irving, R. T. Khan, R. Dürichen, ``Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units,'' 2020, *arXiv:2007.08491*. [Online]. Available: https://arxiv.org/abs/2007.08491 [17] W. Wiharto, H. Kusnanto, and H. Herianto, ``Hybrid system of tiered multivariate analysis and arti_cial neural network for coronary heart disease diagnosis,'' *Int. J. Electr. Comput. Eng.*, vol. 7, no. 2, p. 1023, Apr. 2017.

[18] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, ``Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease,'' in *Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV)*, May 2016, pp. 145_150.

[19] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ``A hybrid classi_cation system for heart disease diagnosis based on the RFRS method,'' *Comput. Math. Med.*, vol. 2017, pp. 1_11, Jan. 2017. [20] D. Singh and J. S. Samagh, ``A comprehensive review of heart

disease prediction using machine learning," *J. Crit. Rev.*, vol. 7, no. 12, p. 2020, 2020.

[21] M. Shouman, T. Turner, and R. Stocker, ``Integrating clustering with different data mining techniques in the diagnosis of heart disease," *J. Comput. Sci. Eng.*, vol. 20, no. 1, pp. 1_10, 2013. [22] I. D. Mienye, Y. Sun, and Z. Wang, ``An improved ensemble learning approach for the prediction of heart disease risk," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100402.

[23] H. Wang, Z. Huang, D. Zhang, J. Arief, T. Lyu, and J. Tian, ``Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease," *IEEE Access*, vol. 8, pp. 97064_97071, 2020.

[24] B. A. Tama, S. Im, and S. Lee, ``Improving an intelligent detection system for coronary heart disease using a two-tier classi_er ensemble," *BioMed Res. Int.*, vol. 2020, Apr. 2020, Art. no. 9816142.

[25] J. Mishra and S. Tarar, *Chronic Disease Prediction Using Deep Learning*. Singapore: Springer, 2020, pp. 201_211.

[26] F. Z. Abdeldjouad, M. Brahami, and N. Matta, *A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques*. Cham, Switzerland: Springer, 2020, pp. 299_306.

[27] M. Tarawneh and O. Embarak, ``Hybrid approach for heart disease prediction using data mining techniques," *Acta Sci. Nutritional Health*, vol. 3, no. 7, pp. 147_151, Jul. 2019. [28] C. B. C. Latha and S. C. Jeeva, ``Improving the accuracy of prediction of heart disease risk based on ensemble classi_cation techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.