Heart Disease Prediction Using Machine Learning Techniques

BAISANI INDRAJA[1], SAI SHREYA POLA[2], NITISH JAIN[3], ULLAS REDDY CH[4], UMESH KUMAR M[5]

[1] *Assistant Professor, IT, VNR VJIET, Hyderabad , India*

[2, 3, 4, 5] *Under Graduate, IT, VNRVJIET, Hyderabad, India*

Abstract- ***Heart related diseases are one of the dominant causes of death in emerging, developing, and even wealthy countries, with millions of people dying each year. Heart Disease refers to the state when the blood supply to the body's organs is cut off, resulting in a blood clot. Usually, this disease affects elderly people but with the drastic changes in the environment and their lifestyles, we can observe minor heart attack occurrences in middle-aged persons as well. This is a situation of major concern. In most cases, heart disease diagnosis relies on a complicated combination of clinical and pathological data. Consequently, clinical professionals and researchers are interested in how to accurately and efficiently predict what is happening in the heart. Some of the most common types of heart diseases are Heart Valve Disease, Coronary Artery Disease (CAD), Pericardial Disease, Heart Arrhythmias etc. The data contains factors such as age, gender, BP, cholesterol and many more that need to be considered and analyzed. This process can consume a lot of time and delays the treatment procedure. To achieve brisk results of the data examination, technology can be used. This project aims to predict heart disease both accurately and quickly by applying machine learning algorithms. The dataset we have used is from two online sources named Kaggle and UCI Machine learning Repository.***

The proposed model uses the dataset from above mentioned sources. The Correlation-based feature selection method determines the best features that correlate with the target class significantly. And also, check for features that do not contribute to determining the target and thus remove them. By using the parameter tuning method, the best tuning parameters are applied and then machine learning algorithms are implemented to train the model. The Stacked Ensemble method algorithm is used to obtain precise results.

Indexed Terms- ***Data examination, heart disease, technology, machine learning, Stacked Ensemble Method***

I. INTRODUCTION

One of the primary diseases prevailing among youth to elderly people in present times is heart disease. According to a survey conducted in the United States, one person dies every 36 seconds. Around 659,000 people in the United States die from heart diseases every year which is 1 in every 4 deaths. Since the heart is responsible for supplying blood to all of the body's organs, human existence is largely dependent on its proper functioning. It is very important to distinguish between healthy people and the ones having a heart problem.

There are two types of factors responsible for heart disease. There can be Controllable and Uncontrollable factors.

- The factors that can be controlled by humans such as smoking, drinking, weight etc., are said as controllable factors.
- The factors that are not controllable such as hereditary factors, age, sex etc., are said to be uncontrollable factors.

Chest pain, having shortness of breath, weakness in the legs or limbs, and pain in the neck, jaw, or throat are all common signs of a cardiac arrest.

Machine learning techniques can be beneficial in diagnosing heart illness, that does not usually show obvious symptoms. A variety of tests like blood pressure, ECG, auscultation, cholesterol, and blood

sugar are performed to diagnose such illness. Because the amount of data is increased, it helps pathologists to reduce the timing of such tests, and obtain more accurate results. The dataset for such a problem contains various features such as age, chest pain type, cholesterol, fasting blood sugar, max heart rate achieved etc., The dataset can be huge and training on such data is a little complicated. The real-time data can contain missing values or repeated values for some features and preprocessing of such data will remove errors from it. By applying machine learning algorithms, we can train and test on such data.

II. LITERATURE SURVEY

Various researches have been done by understanding and applying different machine learning techniques to predict whether a person is a probable subject of having a heart disease. M A Jabbar et al. [1] implemented the Random Forest machine learning algorithm and used the chi square technique for the feature selection process, and obtained the accuracy of 83.70%. Shekharesh Barik et al. [2] proposed three different models trained with three different algorithms which are K-Nearest Neighbours (KNN) algorithm, Random Forest algorithm and Decision Tree algorithm. The model trained with KNN algorithm obtained the highest accuracy of 85.06% among the three models. Pronab Ghosh et al. [3] also designed three models using K-Nearest Neighbours, Decision Tree and Random Forest algorithms. However, they pre-processed the dataset by applying various feature selection techniques. They applied Relief feature selection technique and were able to obtain 6 highly correlated features out of the total 13 features. Then these 6 features were used to build a Random Forest Classifier from which the maximum accuracy of 90% was achieved. Fajr I Alarson et al. [4] applied Gradient-Boosted Trees model and Random Forest classifier and achieved the accuracies of 96.75% and 97% respectively. M Chau Tu et al. [5] implemented Decision Tree algorithm and the Bagging algorithm with Decision Tree and reduce-error pruning option which resulted in greater accuracy of 81.41%. A Relative Study using Feature Selection Techniques was done by Kaushalya Dissanayake et al. [6] to predict heart disease. They researched various existing models that used multiple feature selection techniques and built a graphical representation summarizing the performance evaluation of all the models. V V Ramalingam et al. [7] also conducted a survey on different machine learning models for prediction of cardiovascular diseases. They mainly studied papers that included one or more of these five models: Support Vector Machine (SVM), Naive Bayes Classifier, Decision Tree, K-Nearest Neighbours (KNN) and Random Forest Classifier. P Nancy et al. [8] implemented a tuned Random Forest algorithm over three datasets, and achieved the best accuracy of 86% among all the models implemented. Yamala Sandhya [9] proposed a model built using the Support Vector Machine (SVM) algorithm. She pre-processed the data and used a split ratio of 7:3 on the dataset to build the model. The accuracy achieved in this case is 85.97%. Dr Dilbag Singh et al. [10] proposed an extensive analysis after thoroughly studying various research works. A tabular form containing information of the research work, the machine learning algorithms used in each of them and their performance status, the accuracy, was constructed. Indu Yekkala et al. [11] developed a model that combines Rough Set and Random Forest classifier. The preprocessed data is reduced using the Rough Set feature selection technique and then the model is trained using Random Forest classifier. They achieved an accuracy of 84% through this model.

III. APPROACH USED

The proposed outcome has led us to achieve maximum possible accuracy on the dataset. After collecting and merging the data, we checked for missing values of any feature and corrected them with proper values. Once the data was properly set, feature selection algorithms were applied to identify the features that do not relate with the target class significantly. It is very crucial to differentiate such features from valuable ones because they can deviate the model from an accurate prediction by creating noise. Thus, the feature selection algorithm was applied and unnecessary features were removed from the dataset. This procedure of refining data is called Preprocessing data. Subsequently, we applied parameter tuning methods to figure out the right set of parameters to pass them while using algorithms. The task of selecting a collection of ideal hyperparameters for a learning algorithm is known as hyperparameter optimization or hyperparameter tuning in machine learning. In

simpler words, the value of a parameter that enables the learning algorithm to perform better and give more accurate results is a hyperparameter. We used these parameter values on different machine learning algorithms such as decision trees, KNN, XGB classifier, Random Forest and trained the model and stored their respective accuracies. Finally, we applied the Stacked Ensemble technique to train the model to achieve the maximum accuracy.

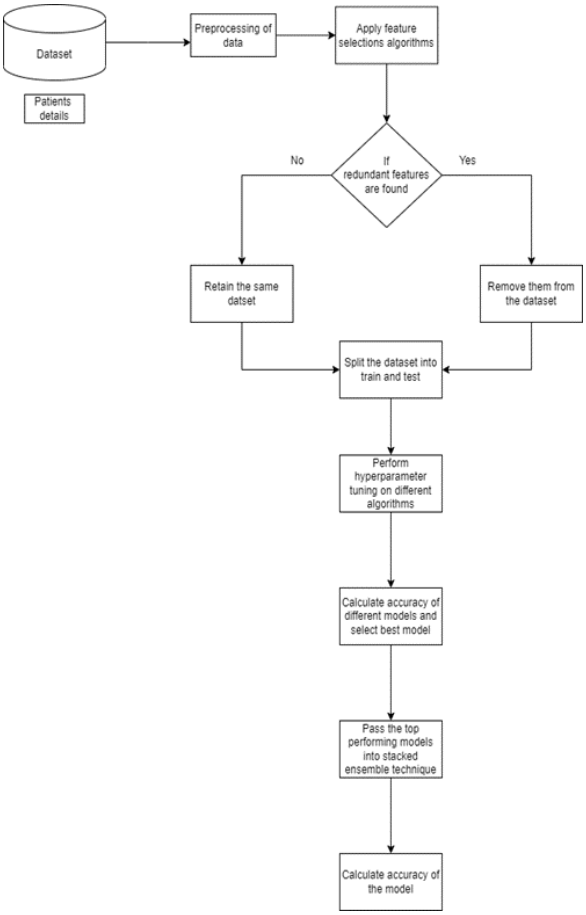

Fig 1 – Flowchart of our work model

A. DATASET DESCRIPTION

We have used the dataset from two online sources named Kaggle and UCI Machine learning Repository. The UCI KDD source possesses databases for four places which are Hungary, Cleveland, the VA Long Beach and Switzerland. And in this project, we have used the Hungarian dataset. The original database from UCI KDD has 76 raw attributes. However, only 13 of these are mentioned in all of the published experiments. These 13 features are predicting variables such as age, sex, cholesterol etc., and the target variable is an independent variable. Since it's a binary classification case, the target variable contains two values (0,1). The value 0 indicates no presence of a heart problem while the value 1 indicates a presence of a heart problem. The dataset from Kaggle online source also contains the same 13 features and the combined dataset consists of 597 records.

List of attributes and their corresponding description:

- Age is mentioned in years
- Sex refers to gender of the person.
  1-> indicates “male gender”,
  0-> indicates female gender
- Cp stands for “chest pain type”. The value ranges from [1,4].
- Value 1 - this value indicates “typical angina”.
- Value 2 - this value indicates “atypical angina”.
- Value 3 - this value indicates “non-anginal pain”.
- Value 4 - this value indicates “asymptomatic”.

- Trestbps indicates “resting blood pressure”.
- Chol indicates “serum cholesterol in mg/dl units”.
- Fbs stands for “fasting blood sugar”. (fbs>120 mg/dl) (1=true; 0=false)
- Restecg stands for “resting electrocardiographic results”.
- Value 0 - specifies “Normal Restecg”.
- Value 1 - having “ST-T wave abnormality”.
- Value 2 - specifies “left ventricular hypertrophy by Estes' criteria”.
- Thalach max heart rate achieved
- Exang specifies “exercise induced angina”.
- Value 1 is for “yes”
- Value 0 is for “no”
- Oldpeak defines “ST depression induced by exercise relative to rest”.

- Slope is the slope of the peak exercise ST segment
- Value 1 - “upsloping”
- Value 2 - “flat”
- Value 3 - “downsloping”
- Ca defines “ Number of major vessels(0-3) coloured by fluoroscopy”.
- Thal shows the “Thalassemia Value”
- Value 0 - it indicates “NULL”
- Value 2 - it indicates “fixed defect”
- Value 6 - it indicates “normal blood flow”

- Value 7 - it indicates “reversable defect”
- Target shows “Diagnosis classes”.
- 0-> indicates “healthy”
- 1-> indicates “presence of a heart problem”

B. FEATURE SELECTION

Feature selection plays a very crucial role in the preprocessing stage. It can be plainly defined as reducing the number of features from a dataset such that the efficiency of the model is enhanced along with the deduction of computing cost. The objective of feature selection procedures is to identify those features that have a strong association with the target variable. It denotes that a change in the values of those features will result in a considerable change in the value of the target variable. So, such features are to be retained in the dataset while those features that do not bring a difference in the target variable can be removed.

a. Correlation Coefficient

The Correlation between two or more variables is the measure of linear relationship between them. In simple terms, it can be defined as the prediction of one variable from the other. It can be observed that variables that are strongly linked to the target variable have major influence on it. In addition, the target and predictor variables should be closely correlated whereas the correlation among the predictor variables themselves should not be significant. One variable can be predicted from other if a strong correlation value is present between them and hence only one variable is needed to predict the target class. Only either of them is required by the model as the other one doesn’t provide any additional information. We can use Pearson Correlation here.

C. HYPERPARAMETER TUNING

Hyper-parameters are various parameter values that are used to influence the learning process and have a substantial impact on machine learning model performance. Multiple trials are done in a single training job to perform hyper-parameter optimization. In every trial, the training application is completely run with values of selected hyperparameters, set within the boundaries defined. It keeps track of each trial's outcomes, makes modifications for subsequent trials, and finally provides the optimum hyper-parameters to attain the highest level of accuracy.

The two best strategies for hyperparameter tuning [14]:

i. GridSearchCV
ii. RandomizedSearchCV

D. CLASSIFIERS

a. Decision Tree Classifier

Decision trees are used to classify both regression and classification problem statements. They are tree-like structures created in a top-down fashion where each path from the root node to the leaf node represents a sequence of rules. The decision-making process starts from the root node and continues until leaf nodes have appeared.

b. KNN Classifier

The KNN model checks for the similarity between new data points and the existing data points, and assigns the former a category that is most related to the existing categories. The model initially prepares all the known data and stores the data. Once the model encounters a new data point, it classifies the data point based on similarity.

c. Naive Bayes Classifier

In Naïve Bayes classifier, it is assumed that presence of all variables is independent of one another, given the target variable. It follows the Bayes Theorem approach and calculates prediction based on the probability of an object.

d. Support Vector Classifier

This approach generates a decision boundary that divides n-dimensional space into classes, allowing us to simply place additional data points in the appropriate category in the future.

e. Random Forest Classifier

Random Forest is an ensembling technique that makes use of a number of decision tree models to build a final model. It aggregates these models where each model is built using subsets of a single dataset, so that the anomalies are reduced. This improves in managing the outliers and reduces deviations caused by them.

f. Extreme gradient boosting classifier

The boosting model is one of the ensemble machine learning models. The working of the model states that

trees are inserted one at a time to fit the ensemble model by correcting the predictions made by previous models.

E. STACKED ENSEMBLE TECHNIQUE

Stacked Ensemble techniques combine predictions of multiple heterogeneous models. These models are considered to be base models and a prediction is made using these models. The model follows a meta-learning algorithm in which the model learns to combine the projections from two or base machine learning models such that the accuracy is improved.

The advantage of this technique is that it can merge the capabilities of many weak learners or base models to make predictions that are more accurate than the predictions made by any single model in the ensemble on a classification or regression challenges.

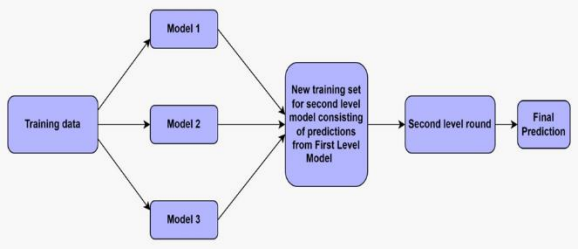

Fig 2- Architecture of Stacked Ensemble Model

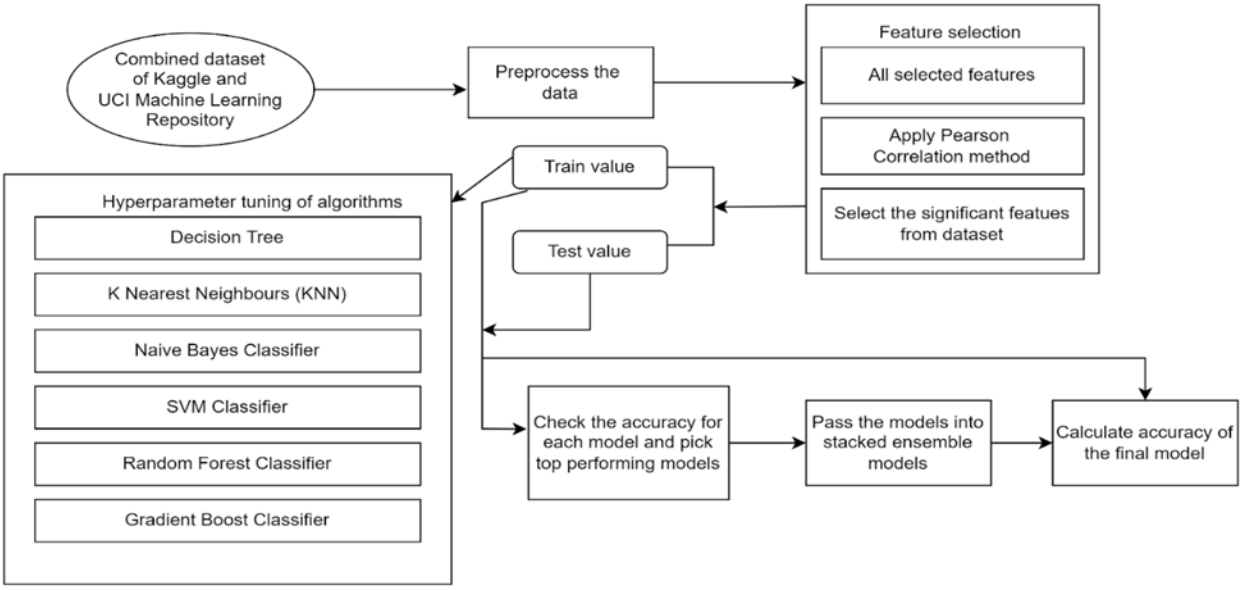

Fig 3- Detailed workflow of our system

IV. RESULT

The following table depicts the accuracies achieved with different machine learning models as well as with the Stacked Ensemble Model. It also shows the specificity and sensitivity values obtained. The maximum accuracy obtained was 88.89% using stack ensembled model among all the models.

Algorithm	Accuracy (in %)	Specificity (in %)	Sensitivity (in %)
Naïve Bayes	76.67	78.70	73.61
Support Vector Machine	85.86	84.26	87.50
KNN	86.67	86.11	87.50
XGB	84.44	84.26	84.72
Decision Tree	85.00	80.56	91.67
Random Forest	87.22	86.11	88.89
Stacked Ensemble Method	88.89	87.96	90.28

Table 1- Comparing Various Algorithms

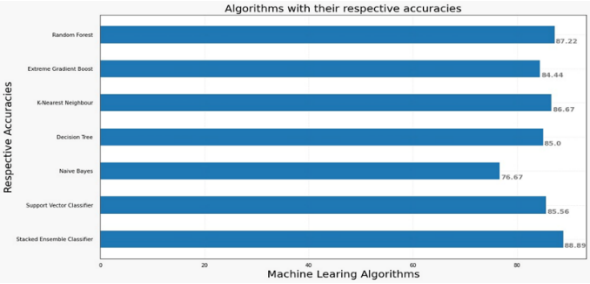

Fig 4- Graphical representation of accuracies with different algorithms

V. CONCLUSION AND FUTURE SCOPE

Six different machine learning algorithms were implemented among which the Random Forest Classifier gave the maximum accuracy. The models were implemented after performing hyperparameter tuning methods on the training dataset. To obtain further greater accuracy, the top five performing models from these six were given as the input to the stacked ensembled model. The top five performers were Random Forest, KNN, Support Vector Classifier, Decision Tree and Extreme Gradient Boost models. However, Stacked Ensemble model of those resulted in maximum accuracy of almost 89%.

Today, there is a dire need of diagnosing if a person is a probable subject of having a heart disease. Early prediction and treatment may avert the risk of fatality. There are many models that are capable of determining if a person is having heart disease or not, but still higher accuracy can be obtained. Hence, for future work, we may work on larger datasets, preprocess the data using more and more methods, and further increase the accuracy and scalability of the system.

REFERENCES

[1] Jabbar M.A., Deekshatulu B.L., Chandra P. (2016) Prediction of Heart Disease Using Random Forest and Feature Subset Selection. In: Snášel V., Abraham A., Krömer P., Pant M., Muda A. (eds) Innovations in Bio-Inspired Computing and Applications. Advances in Intelligent Systems and Computing, vol 424. Springer, Cham. https://doi.org/10.1007/978-3-319-28031-8_16

[2] Barik S., Mohanty S., Rout D., Mohanty S., Patra A.K., Mishra A.K. (2020) Heart Disease Prediction Using Machine Learning Techniques. In: Pradhan G., Morris S., Nayak N. (eds) Advances in Electrical Control and Signal Systems. Lecture Notes in Electrical Engineering, vol 665. Springer, Singapore. https://doi.org/10.1007/978-981-15-5262-5_67

[3] Pronab Ghosh, Sami Azam, Asif Karim, Mirjam Jonkman, and MD. Zahid Hasan. 2021. Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases. 2021 the 5th International Conference on Information System and Data Mining. Association for Computing Machinery, New York, NY, USA, 14–20. DOI: https://doi.org/10.1145/3471287.3471297

[4] Alarsan, F.I., Younes, M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. J Big Data 6, 81 (2019). https://doi.org/10.1186/s40537-019-0244-x

[5] M. C. Tu, D. Shin and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," 2009 2nd International Conference on Biomedical Engineering and Informatics, 2009, pp. 1-4, doi: 10.1109/BMEI.2009.5301650.

[6] Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021.https://doi.org/10.1155/2021/5581806

[7] Ramalingam, V V & Dandapath, Ayantan & Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

[8] P.Nancy, B.Swaminathan, K.Navina, B.Nandhine, P.Lokesh. (2020). Tuned Random Forest Algorithm for Improved Prediction of Cardiovascular Disease. International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, Vol 9. 10.35940/ijrte.A1599.059120.

[9] Sandhya, Yamala. (2020). Prediction of Heart Diseases using Support Vector Machine. International Journal for Research in Applied Science and Engineering Technology. 8. 126-135. 10.22214/ijraset.2020.2021.

[10] Singh, Dilbag & Samagh, Jasjit. (2020). A COMPREHENSIVE REVIEW OF HEART DISEASE PREDICTION USING MACHINE LEARNING. Journal of Critical Reviews. 7. 281-285. 10.31838/jcr.07.12.54.

[11] Yekkala, Indu & Dixit, Sunanda. (2019). "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection: Breakthroughs in Research and Practice". 10.4018/978-1-5225-8185-7.ch011.

[12] https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview#:~:text=Hyperparameter%20tuning%20takes%20advantage%20of,maximizes%20your%20model's%20predictive%20accuracy

[13] https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

[14] https://www.freecodecamp.org/news/hyperparameter-optimization-techniques-machine-learning/

[15] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/