An Optimized Ensemble Approach For Heart Disease Detection

Ritika Kayal, Ushus Elizebeth Zachariah

Abstract: Today there are several machine learning models available to detect and diagnose the existence of various diseases. Each model has its own pros and cons and are appropriate for different datasets. This study aims to compare and contrast the use of various machine learning models on the heart disease dataset. Furthermore, the existing algorithms can be considered as weak classifiers and combined to create a novel optimized ensemble machine learning model with better performance for the given dataset. This project experiments and studies the effects of combining existing classification algorithms to detect heart disease. This study aims to assist diagnosis by utilizing machine learning models. It makes use of a novel ensemble learning approach for heart disease detection. The novel approach is studied in the UCI Heart Disease Dataset. The results of various models have been compared on the basis of several performance metrics such as accuracy, precision, recall, F1-score, computation time and the area under receiver operating curve. The proposed novel hybrid model is targeted to have a performance that is significantly greater than the existing classification algorithms. The optimized ensemble model created using Extreme Gradient Boost and K Nearest Neighbors have achieved an accuracy of 93.44%, recall (sensitivity) of 94%, precision of 93%, specificity of 92%, F1-Score of 87% and AUROC of 94%. Hence the performance of the proposed optimized ensemble is not only better than the existing individual classifiers but is also better than some of the models proposed in existing literature for the given dataset.

Keywords: Machine Learning, Ensemble Learning, Disease Detection, Classification Algorithms, Exploratory Data Analysis

I. INTRODUCTION

Data is an important asset for attaining valuable information, and is a vital component to gather information. It has aided in revolutionary advancements in science and innovation especially in the fields of engineering, medicine, etc. In the medical services area, there is an immense measure of raw medical information that needs to be mined. This information and medical reports of the patients have hidden patterns on the basis of various attributes which is an important component of data analysis. Studying these patterns can help uncover trends that can aid in disease detection.

Heart illness is a dangerous yet sudden cause of death bringing about an expanding death rate of victims on a yearly basis. Heart illnesses are scientifically called strokes or cardiovascular dand are the main source of death around the world. One in three deaths are a consequence of CVDs according to the World Heart Federation Report of 2016. This is in spite of the way that most untimely heart illnesses are preventable. CVDs occur because of restricted, obstructed, or solidified veins and arteries that forestall the necessary supply of blood to the heart, mind and other different organs of the body. The most common indications of coronary illness are weakness, chest pains, difficulty breathing and pain in the jaw, back, upper abdomen, neck and throat. There are however a few variables that can be controlled in order to assist us with lessening the odds of a coronary illness. These factors are things like regular exercise, stable blood pressure, quitting habits like smoking and maintaining a low level of cholesterol. Generally, a CVD probably won't be detected until an angina, coronary episode such as a heart attack or cardiovascular breakdown takes place. Hence, it is critical for specialists to monitor these cardiovascular parameters and attributes.

The breakthroughs in the field of computer science has permitted the medical sector to gather and organize routine clinical information that aids in important clinical diagnosis. The patient's medical data can be examined to make a better informed medical diagnosis. This includes disease prediction, medical decisions, illness diagnosis and treatment choices. The medical sector has an abundance of accessible information, and accordingly it is data rich yet sadly information poor. In the course of the most recent couple of years, ML models have been a significant part in settling complex, exceptionally non-linear prediction and classification use cases. Thus, it is feasible to create a prediction algorithm that would anticipate the presence or lack of coronary illness in view of different heart-related features (attributes). Generally, in illness detection use cases, it is a vital prerequisite that a detection model identifies a healthy patient as precisely as could really be expected. Since errors and misclassification might lead to subjecting a perfectly healthy patient to superfluous treatment. Various ML models, including näıve Bayes, random forest, decision trees, logistic regression, support vector machine (SVM), and K nearest neighbors (KNNs) and Extreme Gradient Boosting (XGB) have been broadly utilized in a large number of the human disease prediction and classification applications.

II. LITERATURE SURVEY

Raza, K. et. al [1] studied improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In their study Various machine learning models were combined using ensemble learning to optimize the testing and training accuracy. In this paper, ensemble learning was facilitated by making use of the majority voting technique to detect heart illnesses. Comparison of the outcomes with several existing works shows a greater accuracy of the proposed model over the existing classifiers. It shows an 88.88% accuracy.

Obasi, T. et. al [2] proposed a novel approach to heart disease prediction in their paper titled "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases". The given method works according to the working of classification models such as Logistic Regression, Random Forest and Bayesian Classification. It offers a scientific backing for medical decisions made by professionals to identify and diagnose heart illnesses and strokes in people using parameters considered as heart disease risk factors. The outcome portrayed a 92.44% accuracy in heart disease classification using Random Forest. Naïve Bayes Classifier and Logistic Regression Classifier showed an accuracy of 61.96%, and 59.7% respectively.

Mienye et. al. [3] suggested an improved ensemble learning approach for the prediction of heart disease risk. The technique consists of using a mean-based division technique to divide the dataset randomly into small sections. Each of the subsets are fed into the model with the help of classification and regression trees (CART). The various CART models are then used to create a homogeneous ensemble with the help of a weighted aging classifier ensemble on the basis of accuracy. The model makes sure to achieve optimal performance. The outcome of the experiments on the Cleveland dataset attained an accuracy of 93% and 91% on the Framingham dataset. This performance was better than the individual classification models and similar research models.

Dissanayake et. al [4] performed a comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms. The studies in this paper performed an experimental assessment of the working of models made with the help of classification algorithms and appropriate attributes extracted through several feature selection techniques. EDA outcome is done on 10 feature selection approaches namely, recursive feature elimination, MI (mutual information), exhaustive feature selection, ANOVA, Ridge regression, ReliefF, Lasso regression, Chi-squared, etc. To the Cleveland heart disease dataset 6 of the standard classification approaches were applied. The highest accuracy was 63.92% in the absence of feature selection. Using a decision tree along with a backward feature selection, the accuracy was raised to 89% approximately. Its precision, sensitivity and F-score is approximately 91%, 81.76%, and 86%.

Spencer et. al. [5] studied heart disease detection in their paper named "Exploring feature selection and classification methods for predicting heart disease". PCA (principal component analysis), Chi-square testing, symmetrical uncertainty and ReliefF were used to analyze 4 popular heart disease datasets to build important feature sets. Furthermore, a collection of classification models were used to build algorithms which were compared. The model with the most optimized performance was built using an ensemble of Chi-square feature selection with the BayesNet approach. It attained a 85% accuracy on the selected dataset.

III. PROBLEM STATEMENT

In order to help medical professionals make better informed diagnoses, there needs to be an optimized model that will detect diagnosis through novel enhanced machine learning models. This study will also compare and contrast the efficiency of the enhanced model with the existing classification algorithms such as Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Extreme Gradient Boosting, Decision Trees and Random Forest. The primary analysis of the enhanced model will be done on the UCI heart disease dataset.

In order to create an optimized novel hybrid approach to disease classification, the following objectives must be fulfilled:

- Exploratory data analysis of the selected dataset that is, the heart disease detection dataset.
- In-depth understanding of existing classification algorithms namely Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Extreme Gradient Boosting, Decision Trees and Random Forest on the selected dataset.
- Comparing the results of each of the 7 selected classification models on the basis of appropriate performance metrics to identify optimal weak classifiers.
- Creating a novel hybrid ensemble learning model by experimenting with various combinations of weak classifiers.
- Analyzing the performance of the hybrid model to identify application specific advantages and disadvantages.

IV. PROPOSED METHODOLOGY

A. Classification

Detecting whether a patient is at risk of heart disease or not is a classification problem. Classification is the technique of identifying, analyzing, and categorizing data into definite classes. Machine learning models make use of several programs to classify data into preset classes by making use of previously categorized data for training. This study analyzes all existing classification models such as Logistic Regression, Naive Bayes, K Nearest Neighbors, Decision Tree, Random Forest, Extreme Gradient Boost and Support Vector Machine. Logistic Regression is a machine learning technique that is used to detect a binary output, that is, if an event occurs or not. The output always fits in one of the two classes by analyzing the independent variables which can be either numeric or categorical. However, it is always required for a dependent variable that is of type categorical. It is generally represented as, P (Y=1 | X) or P (Y=0 | X). It evaluates the probability of Y, the dependent variable when X, the independent variable, is given. The probability of a value being a part of a specific class is determined with the help of Naive Bayes. The formula is given by (1).

(1) P (A | B) = (P (B | A) x P (A)) / P (B)

k-NN (K Nearest Neighbors) is an algorithm that can identify patterns by making use of datasets for training in order to detect the k nearest data points. When using k-NN in classification problems, the data is evaluated to determine if it should be a part of the same class as its nearest neighbor. For example, if k is 1 then it would belong to the category that is closest to 1. A technique that is a great fit for classification use cases is decision tree. This is because of its ability to categorize data with precision. This allows a more organic classification approach in the absence of human intervention since it makes classes within classes. An extension of the decision tree method is the random forest classification approach. The random forest classifier builds several decision trees by making use of the training dataset. The user data is then fitted within a decision tree created as part of the random forest. This process basically performs an average function on the data to categorize it into the closest tree on the scale of data. The advantage of Random forest models over decision trees is that they avoid the unnecessary forcing of data values into a given class. The SVM (support vector machine) classifier makes use of algorithms for training and classification of data values on the basis of polarity degrees which takes it a step ahead of standard X/Y prediction. XGB (Extreme Gradient Boosting) is an efficient model for classification problems. It works on the principle of gradient boosted decision trees.

B. Ensemble Learning

Once the existing classifier models have been run, we select the best performing models to act as weak classifiers for the novel ensemble approach. The method of strategically generating several models which are then concatenated for a specific machine learning use case is known as ensemble learning. Ensemble learning is essentially a technique that enhances machine learning model performance, or reduces the chances of picking a poor performance model for the particular use case. The two of the most common ensemble approaches are voting and averaging. Both of them are easily understandable and implementable. Classification problems rely on voting whereas regression problems make use of averaging. Hence, the approach in this project uses voting. Each model that is a part of the ensemble votes (performs a prediction) for every test case and the result is the class that has gained the maximum number of votes. Some studies call this method plurality voting.

C. Performance Metrics

After the ensemble model is created we evaluate its performance in comparison with all the existing classification models. The performance metrics used in this project include accuracy, precision, recall, F1-score, AUROC (Area Under Receiver Operating Curve) and training time. Accuracy, precision, recall, F1-score and AUROC all use the values of TP (True Positives), TN (True Negatives), FP (False Positives) and FN (False Negatives). TP is the count of positive instances whose prediction labels are positive as well. TN is the count of negative instances whose prediction labels are negative as well. FP is the count of negative instances whose prediction label is positive. FN is the count of positive instances whose prediction label is negative. Accuracy given by (2) is the ratio of the correctly labeled data to the rest of the data. It is the best metric to make use of since it is the most intuitive. It also works great for symmetric datasets.

(2) Accuracy = (TP+TN)/(TP+FP+FN+TN)

Precision is the ratio of positively labeled data that is correct to all positively labeled data. Precision is generally used when we want a higher degree of confidence in true positives. It is given by (3).

(3) Precision = TP/(TP+FP)

Recall, also known as sensitivity, is the ratio of the positively labeled data that is correct to all the data values that are observed to be positive. This metric is good if having a higher value of false positives is preferred over the idea of having a greater number of false negatives. Alternatively, if recall is used when the idea of having false negatives is dangerous and false positives are preferred instead. It is given by (4).

(4) Recall = TP/(TP+FN)

F1-score, also known as F-Score or F-Measure, is a combination of precision and recall. It is essentially the harmonic average of precision and recall. Hence, F1 Score turns out to be optimal if the dataset has some sort of equal distribution between precision and recall. It is given by (5).

(5) F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Another performance metric used to analyze classification models is the area under the receiver operating curve (AUROC). It tells us the connection between sensitivity (true positive rate) and specificity (false positive rate). In the case of a binary classification application, the AUROC represents the chances of a randomly selected positive instance having a greater chance of being predicted as positive than a randomly selected negative instance. Thus, it is a great metric to evaluate discrimination since it determines the technique's capability to diffrentiate between positive and negative instances. Greater the area under the receiver operating curve curve, the better the model.

Training time is the amount of time taken to train the model on the training dataset. Prediction time is the amount of time taken for the model to run in real-time when given user-defined input. For heart disease detection, it is beneficial for the prediction time to be almost instantaneous to allow real-time prediction.

V. IMPLEMENTATION

The UCI Heart Disease Data Set has been used for this project. This dataset gives information related to heart disease. Dataset contains 13 columns, target is the class variable which is affected by other 12 columns. Exploratory data analysis has been performed to facilitate efficient data cleaning and scaling. The processed data is then fed into each of the seven individual classification models namely, Logistic Regression, Naive Bayes, K Nearest Neighbors, Decision Tree, Random

Forest, Extreme Gradient Boost and Support Vector Machine. The performance of each of the models have been compared on the basis of various metrics such as accuracy, precision, recall, F1-score, Area Under the Receiver Operating Curve (AUROC), etc to determine the best performing models that can act as weak classifiers for our ensemble model.

The best performing models turned out to be Extreme Gradient Boost (XGB) and K-Nearest Neighbors (KNN) with accuracies of 90.16% and 88.52% respectively. KNN is chosen over SVM despite having the same accuracy because both KNN and XGB are non linear classification models. Hence, KNN would be a better fit to ensemble with XGB. The outputs of both KNN and XGB are generalized to represent whether the data point is positive (at risk of heart disease) or negative (not at risk). An ensemble model is created with 5 XGB components and 5 KNN components of varying tuning parameters determined through trial and error to achieve the best performance. A voting classifier is used to determine the output of the ensemble model created through plurality voting. The class with the highest number of votes is the final output.

VI. RESULTS AND DISCUSSIONS

The project aims to create a web portal that makes use of a novel optimized ensemble model for the detection of patients with the risk of heart disease. It is expected that the proposed model is an enhanced and optimized model that can detect heart disease efficiently and in real-time. The metrics used to evaluate the optimized model are accuracy, precision, recall, F1-score, AUROC and computation time. Accuracy has been taken as the primary metric since it is most intuitive and gives an idea of the overall performance of the model, that is, for both positive and negative instances. The accuracy of the proposed Optimized Ensemble (OE) model is significantly greater than the existing classifiers as can be seen in Fig. 3. This shows that the proposed model is better at making correct predictions of heart disease.

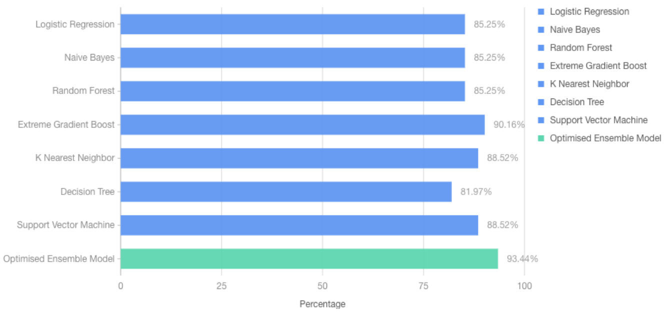

Fig 3. Accuracy Comparison

The proposed model has the greatest Area Under the Receiver Operating Curve (AUROC) which tells us that it is better at discriminating between positive and negative results and has a higher probability of being accurate. This is shown in Fig 4. where the proposed Optimized Ensemble (OE) is marked as orange.

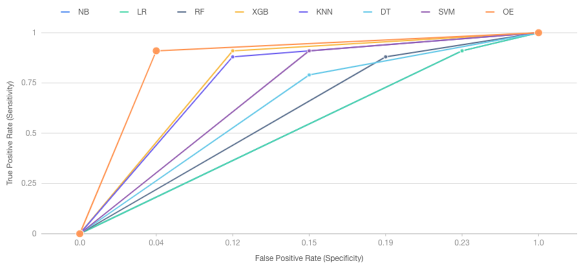

Fig 4. ROC Curves

The proposed Optimized Ensemble (OE) model has a significantly greater value for precision, recall and F1-score. From this we can infer that the proposed model has a higher probability of identifying positive instances and works to reduce the number of false negatives. This is extremely desirable for our use case since it is important to correctly identify patients at risk of heart disease and minimize the chances of missing a patient who is at risk by labeling them incorrectly through prediction.

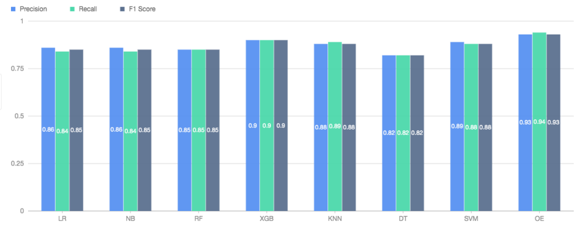

Fig 5. Precision, Recall and F1-Score Comparison

The proposed Optimized Ensemble (OE) model is expected to have an instantaneous prediction time to facilitate real-time predictions using user input data. From Fig 6. we can infer that the results of the prediction time taken by each of the models is in the order of milliseconds. Hence, the required speed has been achieved. The training time is slightly greater but still quite fast since it is also in the order of milliseconds.

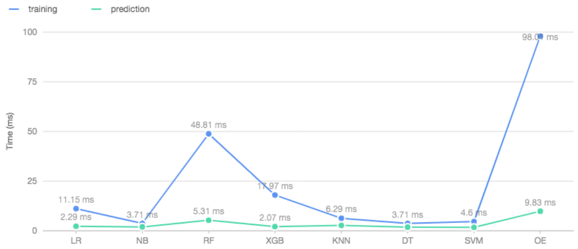

Fig 6. Computation Time Comparison

It is also expected that the performance of the proposed optimized ensemble be better than the models proposed by existing literature that have carried out studies for the same use case using the same dataset. We can evaluate the performance of the proposed Optimized Ensemble against other models with the help of Table 1 and Table 2.

Table 1. Performance of proposed Optimized Ensemble

Characteristics of proposed Optimized Ensemble	Accuracy: 93.44% Recall (Sensitivity): 94% Precision: 93% Specificity: 92% F1-Score: 87% AUROC: 94%

Table 2. Evaluating models in existing literature

Reference	**Characteristics**	**Explanation**
Raza et al. [1].	Accuracy: 89% Recall (Sensitivity): 85% Precision: 89% F1-Score: 87% AUROC: 88%	The performance of the proposed optimized ensemble model is significantly better in terms of almost all metrics.
Obasi et al. [2].	Accuracy: 92% Recall (Sensitivity): 93% Specificity: 92% Precision: 92%	The performance of the proposed optimized ensemble model is better in terms of all metrics.
Mienye et al. [3].	Accuracy: 93% Recall (Sensitivity): 91% Precision: 96% F1-Score: 93%	The proposed optimized ensemble model has a greater recall score and a lower precision score. For this use case, it is critical to avoid false negatives over concentrating on perfect true positives. Hence, a greater recall score is better and the proposed model is preferred.
Robinson et al. [4].	Accuracy: 85% Precision: 84.73% Recall: 85.56%	The performance of the proposed optimized ensemble model is better in terms of all metrics.
Dissanayake et al. [5].	Accuracy: 88.52% Precision: 91.30% Recall (Sensitivity): 80.76% F1-Score: 85.71%	The performance of the proposed optimized ensemble model is significantly better in terms of all metrics.

VII. CONCLUSION

This project achieves an in-depth understanding of the working of various existing classification algorithms namely Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Extreme Gradient Boosting, Decision Trees and Random Forest on the basis of several performance metrics such as accuracy, precision, recall and F1 score. Accordingly, the existing models are treated as weak classifiers and ensembled using a majority voting algorithm to create a novel hybrid approach for heart disease detection.

The proposed novel hybrid model is targeted to have a performance that is significantly greater than the existing classification algorithms. The optimized ensemble model created using Extreme Gradient Boost and K Nearest Neighbors have achieved an accuracy of 93.44%, recall (sensitivity) of 94%, precision of 93%, specificity of 92%, F1-Score of 87% and AUROC of 94%. Hence the performance of the proposed optimized ensemble is not only better than the existing individual classifiers but is also better than some of the models proposed in existing literature for the given dataset. Comparative analysis has also been performed to highlight the advantages and disadvantages of the proposed model in contrast with the existing models. Therefore, the project aids medical professionals to make better informed diagnosis in real time with the power of machine learning.

REFERENCES

1. Raza, Khalid. "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule." In U-Healthcare Monitoring Systems, pp. 179-196. Academic Press, 2019.
2. Obasi, Thankgod, and M. Omair Shafiq. "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases." In 2019 IEEE international conference on big data (big data), pp. 2393-2402. IEEE, 2019.
3. Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang. "An improved ensemble learning approach for the prediction of heart disease risk." Informatics in Medicine Unlocked 20 (2020): 100402.
4. Spencer, Robinson, Fadi Thabtah, Neda Abdelhamid, and Michael Thompson. "Exploring feature selection and classification methods for predicting heart disease." Digital health 6 (2020): 2055207620914777.
5. Dissanayake, Kaushalya, and Md Gapar Md Johar. "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms." Applied Computational Intelligence and Soft Computing 2021 (2021).