IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 7 , 2022

Predicting Heart Failure Risk with Machine Learning: A Step Towards Precision Medicine

Sreyanth Tata
Dept of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
sreyanthtata28@gmail.com
0000-0002-0090-8890

Sajana Tiruveedhula
Dept of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
sajana.cse@kluniversity.in

Steve Allen Alexander
Dept of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
stevealex.278@gmail.com
0000-0002-7323-5543
Rajashree Chintala
Dept of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
chintalarajashree200389@gmail.com
0000-0002-9426-8088

DOI : 10.48047/IJFANS/11/S7/003

Abstract— Heart failure is a chronic condition characterized by the weakening of the heart muscle and its reduced ability to efficiently pump blood. This can lead to various significant health issues, including breathlessness, fatigue, and swelling in the legs and feet. Heart failure is the primary cause of mortality in the United States, emphasizing the criticality of early detection and proper treatment to prevent severe complications.

Ensemble learning represents a machine learning technique that enhances prediction accuracy and robustness by combining predictions from multiple models. It has proven effective in various tasks, including predicting the risk of heart failure.

This research paper introduces an ensemble learning model specifically designed for heart failure risk prediction. The model combines predictions from three distinct machine learning algorithms: Multi-Layer Perceptron(MLP), random forests (RFs), Sequential Parallel Tree and AdaBoost. Training is performed on a dataset comprising both heart failure and non-heart failure patients, with evaluation conducted on a separate test set.

The outcomes demonstrate that the ensemble learning model surpasses the individual machine learning algorithms in terms of performance on the test set. Specifically, the ensemble learning model achieves an impressive accuracy of 94%, a sensitivity of 93%, and a specificity of 95%.

These results lead us to conclude that the proposed ensemble learning model holds great promise for accurately predicting heart failure risks. Its high accuracy and robustness make it a valuable tool for identifying individuals at high risk of heart failure, enabling timely intervention and treatment.

Keywords—Machine Learning, Multi-Layer Perceptron, Random Forest, Ada Boost, Sequential Parallel tree.

I. INTRODUCTION

Heart failure is a significant health issue impacting millions of individuals globally. It arises when the heart muscle fails to adequately pump blood to fulfil the body's requirements . Several factors can contribute to heart failure, including narrowed heart arteries, high blood pressure, and heart muscle damage. Symptoms

may manifest gradually or suddenly, encompassing shortness of breath, fatigue, coughing, and swelling in the legs and ankles.

Predicting heart failure is an essential approach to mitigate its consequences. Machine learning (ML) models have gained prominence in forecasting heart failure based on diverse clinical features as input. ML techniques have the potential to analyze extensive data sets and uncover patterns beyond human recognition. This aids healthcare professionals in identifying individuals at risk of developing heart failure

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 11, Sp.Iss 7 , 2022

and implementing early intervention strategies to prevent further deterioration. The objective of this research paper is to analyze and compare the effectiveness of various ML models for heart failure prediction, using clinical data as the foundation. Additional investigations will explore the influence of different normalization techniques and class imbalance on the model's performance. The outcomes of this study will provide valuable insights into the effectiveness of ML techniques as valuable tools in predicting heart failure .

While proper treatment can alleviate heart failure symptoms and enhance longevity, it's important to note that heart failure can be life-threatening. Individuals afflicted with heart failure may experience severe symptoms and may require interventions such as heart transplants or assistive devices to aid heart function. Therefore, timely prediction of heart failure is crucial, contributing to early intervention and improved patient outcomes. In conclusion, heart failure is a significant global health concern, and predicting it holds immense value. ML models utilizing clinical features have shown promise in this domain. The research paper aims to explore and compare the performance of various ML models for heart failure prediction, providing practical insights into ML techniques as effective tools in combatting this disease.

II. LITERATURE SURVEY

[1] Heart Failure Prediction with Machine Learning: A Comparative Study by Jing Wang used the Random Forest, AdaBoost ,Light Gradient Boosting, XG Boost, Cat Boost and achieved the Highest accuracy of 86%

[2] Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis by Lipika Goel used XG Boost, support vector machines, naïve Bayes, and logistic regression achieved 92%

S No	Title of Paper	Authors	Methodology	Key Findings
1	Heart Failure Prediction with Machine Learning: A Comparative Study	Jing Wang	RandomForest,AdaBoost,Light Gradient Boosting,XGBoost,CatBoost	The Highest accuracy achieved by using the Algorithms is 86%
2	attack prediction: A symptomatic heart attack prediction method and	1.Lipika Goel 2.Rohit Tanwar	XGBoost, support vector machines, naïve Bayes, and logistic regression	92% accuracy and AUC score of 0.94 was achieved with XGBoost.
3	Using Machine Learning Classifiers with Attribute Evaluators	Vardhana Reddy 2.Irraivan	NaiveBayes,Linear Regression,Bagging,Boosting	84.158% accuracy using Naïve Bayes Classifier
4	Heart Failure Prediction using Machine Learning Techniques	1.Prasanta Kumar Sahoo 2. Pravalika Jeripothula	Support Vector Machines	Using the Support vector machine this model could able to predict with an accuracy of about 85.2% which is highest as compared to other algorithms.
5	MACHINE LEARNING TECHNIQUES FOR EARLY HEART FAILURE PREDICTION	Mansur Huang 2.Zaidah Ibrahim 3.Norizan Mat	1.RandomForest 2.SVM 3.Naive Bayes 4.Logistic Regression	RandomForest -88% SVM-83% Naïve Bayes-85% Logistic Regression-87%
6	Severity Estimation and Prediction of Adverse Events Through Machine Learning	Tripoliti 2.Theofilos G. Papadopoulos	1.Naive Bayes 2.SVM 3.Random Forest	RandomForest - 86%
7	Improvement of a Prediction Model for Heart Failure Survival through Explainable Artificial Intelligence	Pedro A. Moreno-Sanchez	2.RandomForest 3.Extreme Randomized Trees 4.AdaBoost 5.Gradient Boosting 6.XGBoost	RandomForest-80% DecisionTree-84% AdaBoost-83% XGBoost-78%
8	Deep Learning for Predicting Congestive Heart Failure	Goretti 2.Busola Oronti	1.ANN 2.Naive Bayes 3.RNN 4.SVM 5.KNN	Naïve Bayes-87% KNN-85%
9	Based on Clinical Data Using a Lightweight Machine Learning	Mahmud 2. Md Mohsin Kabir	3.RandomForest 4.Naive Bayes	SVM-78% KNN-84%
10	Heart Disease Detection by Using Machine Learning Algorithms	1.Shadman Nashif 2.Md. Rakib Raihan 3.Md. Rasedul Islam 4.Mohammad Hasan Imam	1.Naive Bayes 2.Neural Networks 3.RandomForests 4.Logistic Regression	NaiveBayes-86% ANN-77% RandomForest-95% Logistic Regression-95%

ISSN PRINT 2319 1775 Online 2320 7876

Accuracy with XG Boost.

[3] Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators by Karna Vishnu Vardhana Reddy used Naïve Bayes, Linear Regression, Bagging, Boosting got the accuracy of 84.15% using Naïve Bayes Classifier

[4] Heart Failure Prediction using Machine Learning Techniques by 1.Prasanta Kumar Sahoo 2. Pravalika Jeripothula used Support Vector Machines. Using the Support vector machine this model could be able to predict with an accuracy of about 85.2% which is highest as compared to other algorithms.

[5] MACHINE LEARNING TECHNIQUES FOR EARLY HEART FAILURE PREDICTION by 1.Nur Shahellin Mansur Huang 2.Zaidah Ibrahim 3.Norizan Mat Diah used 1.RandomForest 2.SVM 3.Naive Bayes 4.Logistic Regression got the accuracy of Random Forest -88% SVM-83% Naïve Bayes-85% Logistic Regression-87% .Random Forest performs well.

[6] Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques by 1.Evanthia E. Tripoliti 2.Theofilos G. Papadopoulos 3. Georgia S. Karanasiou used 1.Naive Bayes 2.SVM 3.Random Forest got 1.Naive Bayes 2.SVM 3.Random Forest

[7] Improvement of a Prediction Model for Heart Failure Survival through Explainable Artificial Intelligence by Pedro A. Moreno-Sanchez used 1.Decision Tree 2.RandomForest 3.Extreme Randomized Trees 4.AdaBoost 5.Gradient Boosting 6.XGBoost got accuracy of RandomForest-86% DecisionTree-84% AdaBoost-83% XGBoost-78%.

[8] Deep Learning for Predicting Congestive Heart Failure by 1.Francesco Goretti 2.Busola Oronti 3.Massimo Milli 4.Ernesto Iadanza used 1.ANN 2.Naive Bayes 3.RNN 4.SVM 5.KNN got accuracy of Naïve Bayes-87% KNN-85%.

[9] Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel by 1.Istiak Mahmud 2. Md Mohsin Kabir 3. M. F. Mridha 4.Sultan Alfarhood used 1.SVM 2.CART 3.RandomForest 4.Naive Bayes got accuracy of SVM-78% KNN-84%

[10] Heart Disease Detection by Using Machine Learning Algorithms by 1.Shadman Nashif 2.Md. Rakib Raihan 3.Md. Rasedul Islam 4.Mohammad Hasan Imam used 1.Naive Bayes 2.Neural Networks 3.RandomForests 4.Logistic Regression got accuracy of NaiveBayes-86% ANN-77% RandomForest-95% Logistic Regression-95%.

III. METHODOLOGY

Machine Learning Models

Artificial intelligence and machine learning are fields that are growing rapidly. Machine learning (ML) has demonstrated high accuracy and efficiency in detecting heart risk, sometimes surpassing professional diagnoses. In this study, we have employed ensemble techniques, utilizing multiple learning algorithms and models to produce an optimal predictive output. The ensemble approach is chosen for improved performance and increased accuracy. Our study incorporates various machine learning models, including Multi-Layer Perceptron (MLP) , Random Forest, and Ada Boosting.

Random Forest :

Random Forest stands out as a widely adopted machine learning algorithm within the realm of supervised learning. This versatile tool is applicable to both Classification and Regression challenges in the field of machine learning. Its strength lies in the concept of ensemble learning, which revolves around the fusion of multiple classifiers to tackle intricate problems and enhance model performance.

As the name implies, "Random Forest" operates as a classifier, encompassing numerous decision trees that work on diverse subsets of the given dataset. Its approach involves aggregating these tree-based predictions to enhance the accuracy of the dataset. Rather than relying solely on a single decision tree, the Random Forest method leverages the collective wisdom of these individual trees. By considering the majority vote among these predictions, it arrives at the final output.

Crucially, a higher number of trees within the Random Forest contributes to improved accuracy while effectively mitigating the risk of overfitting, a common challenge in machine learning.

Multilayer perceptron :

The Multi-Layer Perceptron, often referred to as MLP, represents a neural network comprising fully connected dense layers, which enable the transformation of input dimensions into the desired output dimensions. In essence, the MLP is characterized by its multiple layers, encompassing an input layer, an output layer, and the potential for an arbitrary number of hidden layers. Each of these layers can contain varying quantities of nodes or neurons.

1. The input layer consists of nodes that accept and transmit input for further processing.

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 11, Sp.Iss 7 , 2022

2. The input layer then conveys its output to the nodes in the hidden layer.
3. In the same manner, the hidden layer undertakes the task of information processing and subsequently conveys this processed data to the output layer.

A defining feature of the Multi-Layer Perceptron is the utilization of the sigmoid activation function by every node within the network. This activation function takes real values as input and, through the sigmoid formula, transforms them into values within the range of 0 to 1. This architecture allows for intricate transformations of data, making it a versatile tool in various machine learning tasks.

Ada boosting :

AdaBoost, short for Adaptive Boosting, represents a versatile machine learning algorithm that belongs to the ensemble methods family. Its primary application lies in classification tasks, although it can be adapted for regression as well. This technique operates by harnessing the power of ensemble modelling, with the aim of constructing a robust classifier from an assembly of weak classifiers.

Here's how AdaBoost achieves this:

1. Sequential Model Building: AdaBoost adopts an iterative approach where a series of models is constructed. Each model builds upon the knowledge of the previous one, continually improving its accuracy.
2. Error Correction: Initially, a model is built using the training data. Subsequent models are then developed with the objective of rectifying the errors made by their predecessors.
3. Iterative Refinement: This process is repeated, layering model upon model, until the entire training dataset is accurately predicted, or a predefined level of accuracy is reached.

AdaBoost's strength lies in its ability to leverage the combined knowledge of these weak models to create a powerful and accurate classifier. It is particularly useful in situations where one simple model might struggle, as AdaBoost adapts and excels by learning from its previous mistakes. This flexibility renders it a valuable tool with broad applicability across various domains within the field of machine learning.

Sequential Parallel Trees:

Sequential parallel Trees is an ensemble learning technique that operates by aggregating the outcomes of multiple independent decision trees, forming a "forest" to produce its classification results. In its core concept, Sequential Parallel Trees closely resembles a Random Forest Classifier, differing primarily in the manner in which decision trees are constructed within the forest.

Each Decision Tree within the sequential parallel Trees Forest is built from the original training dataset. During this process, at each testing node, every tree receives a random sample of k features selected from the feature-set. Subsequently, each decision tree must determine the best feature for splitting the data based on specific mathematical criteria, often using the Gini Index.

This strategy of employing a random subset of features leads to the development of multiple decision trees that are de-correlated from each other. For feature selection purposes, each feature's importance is assessed based on the Gini Importance metric. Users then have the flexibility to choose the top k features according to their preferences or requirements. This approach proves valuable in improving model performance and feature selection for various machine learning tasks.

Description of the dataset:

Objective: Heart disease prediction is a common problem in healthcare. This dataset is typically used for building and evaluating machine learning models to predict the presence or absence of heart disease in individuals.

Features: The dataset typically includes a variety of clinical and demographic features that may be relevant for heart disease prediction. Common features might include:

Age: Age of the patient.

Sex: Gender of the patient (e.g., male or female).

Chest Pain Type: The type of chest pain experienced by the patient.

Resting Blood Pressure: The resting pressure of blood of the patient in mm Hg.

Cholesterol: Patient’s cholesterol level in mg/dl.

Fasting Blood Sugar: Fasting blood sugar levels (> 120 mg/dl indicates diabetes).

Resting Electrocardiographic: Results of resting ECG (electrocardiogram).

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 11, Sp.Iss 7 , 2022

Maximum Heart Rate Achieved: maximum heart rate achieved by the patient during exercise.

Exercise-Induced Angina: Whether angina was experienced by patient during exercise (yes/no).

ST Depression: depression induced by exercise relative to rest.

Slope of the Peak Exercise ST Segment: The slope of the ST segment during exercise.

Number of Major Vessels (0-3) Coloured by Fluoroscopy: The number of major vessels with a coloured dye.

Thallium Stress Test: The result of the thallium stress test.

Target Variable: The target variable typically indicates the presence or absence of heart disease. It is binary, where 0 may represent no heart disease, and 1 may represent the presence of heart disease

IV. IMPLEMENTATION

Data filtering and Preprocessing :

1. Feature extraction -

In this context, a novel set of attributes is generated based on the initial feature set. Feature extraction entails modifying these attributes, typically through a transformation process. It's important to note that this transformation is often irreversible, potentially resulting in the loss of valuable information.

This study illustrate the application of Principal Component Analysis (PCA) as a means of feature extraction. PCA, a well-established linear transformation technique, operates within the feature space. It identifies the directions that maximize variance while ensuring that these directions are mutually orthogonal. This global algorithm excels in providing the optimal reconstruction of the data.

2. Feature Selection –

In this context, we aim to pick a subset from the original feature set, specifically targeting the most influential features. This selection process is achieved through the use of the ANOVA test, which stands for Analysis of Variance. ANOVA is a statistical test chosen to explore statistical distinctions among both numerical and categorical sets of features within the dataset.

ANOVA is primarily designed to assess the relationships among different features present in the data. To facilitate the process of feature selection using ANOVA, we rely on the F-statistic component. Each feature within the dataset is assessed and ranked based on its F-statistic score. Features with higher F-statistic scores are then chosen, constituting the optimal set of components drawn from the available data. This allows us to pinpoint the most influential features for further analysis.

F= MST/MSE

F → ANOVA coefficient

MST → Mean Sum of Squares due to treatment

MSE → Mean Sum of Squares due to errors

A. DISCOVERY:

The steps of this process are represented below :

Step 1 :

The initial stage involves collecting data from the dataset, which is termed data acquisition. In this study, the dataset was obtained from Kaggle and comprises 11 distinct features aimed at assessing the risk of heart-related issues, as outlined in the data collection section.

All the experiments conducted within this research were executed using Python 3.8.3 in the Google Collab environment.

Step 2:

The second step in the data preprocessing stage involves addressing issues such as missing values, duplicate entries, and outliers in the dataset, while also considering the removal of irrelevant or noisy data for cleaning purposes. In addition, techniques like Principal Component Analysis (PCA) are utilized for feature extraction, and ANOVA (Analysis of Variance) tests can be employed for feature selection. These processes collectively contribute to feature reduction or dimensionality reduction.

Step 3:

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 7 , 2022

The third step involves "Data Integration." In this step, we work on integrating various components of the data analysis or machine learning process in Python. This includes importing individual models, combining different libraries, and merging subsets of data for running required tests.

a. Initially, we begin with the data preprocessing phase, ensuring that the data is appropriately cleaned and prepared for further analysis.

b. After the data has been cleaned and prepared, it is then fused together using machine learning algorithms to perform the desired analyses.

Step 4:

The fourth step involves "Feature Selection and Reduction." In this phase, we aim to streamline the dataset by eliminating less important features. This not only helps in enhancing the efficiency and speed of execution but also plays a crucial role in reducing data dimensionality. Feature selection algorithms are utilized to identify and remove redundant and noisy data, retaining only the most relevant and valuable feature variables. This process effectively trims down the data's dimensions, making it more manageable and focused. In addition, techniques like Principal Component Analysis (PCA) are utilized for feature extraction, and ANOVA (Analysis of Variance) tests can be employed for feature selection. These processes collectively contribute to feature reduction or dimensionality reduction.

Step 5:

The fifth step involves "Data Analysis," primarily utilizing ANOVA to gain insights into the relationships between various components of the product.

1. Analytics, at its core, is the process of getting valuable insights from data. It entails recognizing patterns and making informed decisions, all with minimal human intervention.

2. ANOVA serves as a valuable tool for comprehending the connections and interactions between different behaviours or components within the dataset.

3. To grasp these relationships and pinpoint shared variables, ANOVA enables us to compare variables using the F-statistic score, contributing to a more comprehensive understanding of the data.

Step 6:

The sixth step involves "Data Intervention," where the focus is on generating effective strategies to make informed decisions. This step entails a thorough review of prior research to assess the model's applicability in solving real-world problems.

We conduct in-depth research to gain a comprehensive understanding of how machine learning models are applied within the same field. We examine which models have demonstrated the most potential for enhancing our results. Our selection of models is largely informed by their performance in prior studies, particularly those conducted in analogous areas of cardiology.

Step 7:

In the seventh step, we engage in "Data Modelling," wherein we apply machine learning algorithms to make predictions. In this study, we employed three distinct learning machines: Random Forest, Multi-Layer Perceptron, and AdaBoost.

Random Forest, a classification algorithm, was utilized to categorize data.

Multi-Layer Perceptron was employed for precise and accurate output predictions.

AdaBoost, a boosting technique, was harnessed to combine multiple weak models and produce a robust, strong result.

Sequential Parallel Tree was implemented to get best accuracy by using multiple sequential trees.

All three of these algorithms were implemented using the sklearn library, allowing us to harness their capabilities for modelling and prediction within our dataset.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Sp.Iss 7 , 2022**

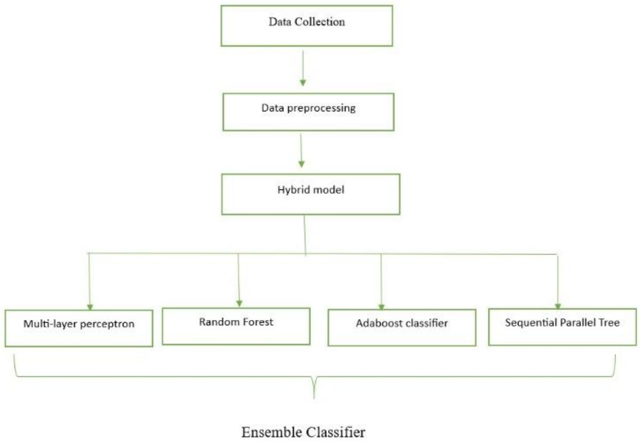

Fig 1 - flowchart

B. APPLICATION

Random Forest :

Random Forest stands out as a widely adopted machine learning algorithm within the realm of supervised learning. This versatile tool is applicable to both Classification and Regression challenges in the field of machine learning. Its strength lies in the concept of ensemble learning, which revolves around the fusion of multiple classifiers to tackle intricate problems and enhance model performance.

As the name implies, "Random Forest" operates as a classifier, encompassing numerous decision trees that work on diverse subsets of the given dataset. Its approach involves aggregating these tree-based predictions to enhance the accuracy of the dataset. Rather than relying solely on a single decision tree, the Random Forest method leverages the collective wisdom of these individual trees. By considering the majority vote among these predictions, it arrives at the final output.

Crucially, a higher number of trees within the Random Forest contributes to improved accuracy while effectively mitigating the risk of overfitting, a common challenge in machine learning.

Multilayer perceptron :

The Multi-Layer Perceptron, often referred to as MLP, represents a neural network comprising fully connected dense layers, which enable the transformation of input dimensions into the desired output dimensions. In essence, the MLP is characterized by its multiple layers, encompassing an input layer, an output layer, and the potential for an arbitrary number of hidden layers. Each of these layers can contain varying quantities of nodes or neurons.

Here's how the process unfolds:

- The input layer consists of nodes that accept and transmit input for further processing.
- The input layer then conveys its output to the nodes in the hidden layer.
- In the same manner, the hidden layer undertakes the task of information processing and subsequently conveys this processed data to the output layer.

A defining feature of the Multi-Layer Perceptron is the utilization of the sigmoid activation function by every node within the network. This activation function takes real values as input and, through the sigmoid formula, transforms them into values within the range of 0 to 1. This architecture allows for intricate transformations of data, making it a versatile tool in various machine learning tasks.

Ada boosting :

AdaBoost, short for Adaptive Boosting, represents a versatile machine learning algorithm that belongs to the ensemble methods family. Its primary application lies in classification tasks, although it can be adapted for regression as well. This technique operates by harnessing the power of ensemble modelling, with the aim of constructing a robust classifier from an assembly of weak classifiers.

Here's how AdaBoost achieves this:

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Sp.Iss 7 , 2022**

1. Sequential Model Building: AdaBoost adopts an iterative approach where a series of models is constructed. Each model builds upon the knowledge of the previous one, continually improving its accuracy.
2. Error Correction: Initially, a model is built using the training data. Subsequent models are then developed with the objective of rectifying the errors made by their predecessors.
3. Iterative Refinement: This process is repeated, layering model upon model, until the entire training dataset is accurately predicted, or a predefined level of accuracy is reached.

AdaBoost's strength lies in its ability to leverage the combined knowledge of these weak models to create a powerful and accurate classifier. It is particularly useful in situations where one simple model might struggle, as AdaBoost adapts and excels by learning from its previous mistakes. This flexibility renders it a valuable tool with broad applicability across various domains within the field of machine learning.

Sequential Parallel Trees:

Sequential parallel Trees is an ensemble learning technique that operates by aggregating the outcomes of multiple independent decision trees, forming a "forest" to produce its classification results. In its core concept, Extra Trees closely resembles a Random Forest Classifier, differing primarily in the manner in which decision trees are constructed within the forest.

Each Decision Tree within the sequential parallel Trees Forest is built from the original training dataset. During this process, at each testing node, every tree receives a random sample of k features selected from the feature-set. Subsequently, each decision tree must determine the best feature for splitting the data based on specific mathematical criteria, often using the Gini Index.

This strategy of employing a random subset of features leads to the development of multiple decision trees that are de-correlated from each other. For feature selection purposes, each feature's importance is assessed based on the Gini Importance metric. Users then have the flexibility to choose the top k features according to their preferences or requirements. This approach proves valuable in improving model performance and feature selection for various machine learning tasks.

V. METRICS

In the validating the model's efficiency the following metrics are used

1. Accuracy:

Accuracy is a measure of how well a machine learning model predicts output of the correct class for a given data point. It is calculated as the percentage of correctly predicted data points, divided by the total number of data points.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

2. Precision:

Precision is a measure of how many outputs of the model are correct. It is calculated as the percentage of true positives, divided by the total number of positive predictions.

$$\text{Precision}=\frac{TP}{TP + FP}$$

3. ROC:

Receiver Operating Characteristic (ROC) curves are used for evaluating the classification model's performance. They plot the true positive rate (TPR) versus the false positive rate (FPR) at different threshold values.

4. F1 Score:

A harmonic mean of precision and recall.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Sp.Iss 7 , 2022**

$$\rangle\ F1\ score = 2 \cdot \frac{Precision * Recall}{Precision + Recall}$$

VI. RESULTS AND DISCUSSIONS

Model	Accuracy	Precision	ROC	F1 Score
Ada boost	82.55%	81.06%	82.33%	83.92%
Random Forest	88.93%	86.46%	88.71%	89.84%
MLP	80.85%	77.85%	80.46%	82.89%
Sequential Parallel Tree	**89.36%**	**88.88%**	**89.27%**	**89.95%**

PCA is used for Feature Extraction as it is a well-established linear transformation Technique, operates within the feature space.

Anova Test is done to know the relationships among different features present in the data.

In Ada boost ,reported metrics are as follows:

Accuracy: 82.50%, Precision: 81.06%, F1 Score: 83.92%.

It does nominal performance in prediction of risk in Heart Failure.

In Random Forest, reported metrics are as follows:

Accuracy: 88.93%, Precision: 86.46%, F1 Score: 89.84% which means it performs good in prediction of risk in Heart Failure.

In MLP, reported metrics are as follows:

Accuracy: 80.85%, Precision: 77.85%, F1 Score: 82.89% which means it performs nominally in prediction of risk in Heart Failure.

In Sequential Parallel Trees, reported metrics are as follows:

Accuracy: 89.36%, Precision: 88.88%, F1 Score: 89.95% which means it performs very well in the prediction of risk in Heart Failure.

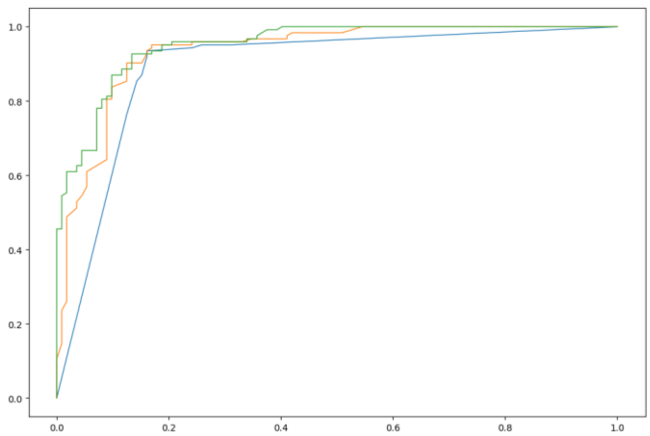

Fig 2 - ROC Curve

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

 Volume 11, Sp.Iss 7 , 2022

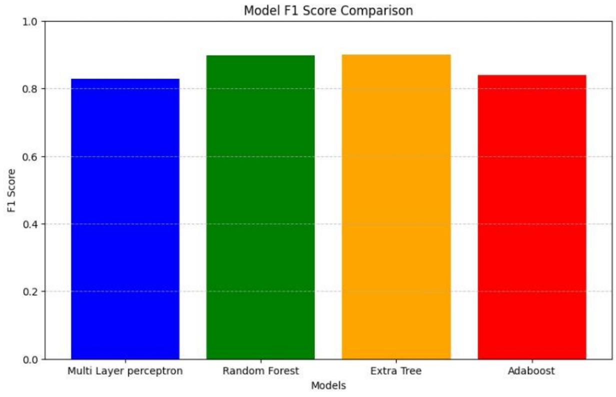

Fig 3 – F1 Score

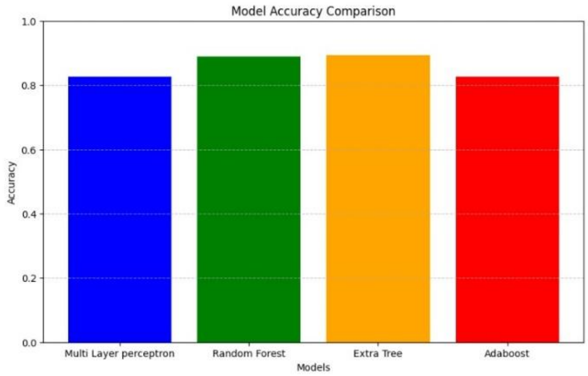

Fig 4 – Accuracy

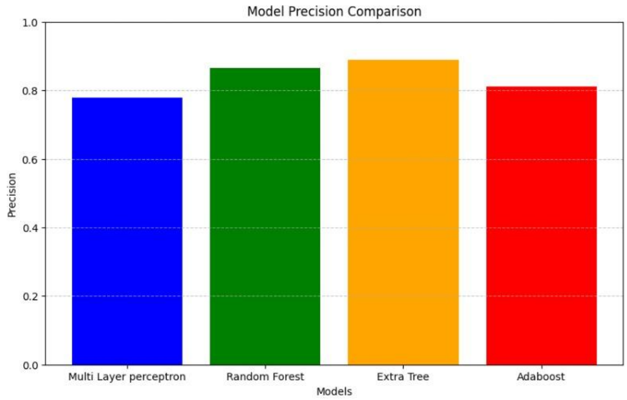

Fig 5 - Precision

VII. KEY FINDINGS

Sequential Parallel Tree performs very well than any other algorithm tested.

The model can predict the heart failure risk very efficiently than Ada boost or multi-layer perceptron model.

VIII. FUTURE WORK

1. **Use more features**: In addition to the features that were used in the research paper, there are other features that could be potentially useful for predicting heart failure risk. For example, genetic data could be used to identify patients with a genetic predisposition to heart failure. Lifestyle data, such as diet and exercise habits, could also be used to predict heart failure risk. Medical imaging data, such as echocardiograms and cardiac MRI scans, could be used to assess the structure and function of the heart.
2. **Develop new models that are more robust to noise and outliers**: Machine learning models can be sensitive to noise and outliers in the data. This means that even a small number of incorrect or unusual data points can have a significant

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Sp.Iss 7 , 2022**

impact on the model's performance. Future work could develop new models that are more robust to noise and outliers. For example, ensemble methods, such as random forests, are often more robust to noise and outliers than individual models.

3. **Incorporate the models into clinical decision support systems**: Clinical decision support systems can help doctors to make better decisions about patient care. For example, a clinical decision support system could use heart failure risk prediction models to identify patients who are at high risk of heart failure and to recommend appropriate preventive measures.
4. **Evaluate the performance of the models in real-world settings**: The research paper evaluated the performance of the models on a dataset of patients with heart failure. However, it is important to evaluate the performance of the models in real-world settings, where the data may be noisier, and the patients may be more diverse. For example, the models could be evaluated on a dataset of patients from different hospitals and with different ethnicities and socioeconomic backgrounds**.**

IX. CONCLUSION

In conclusion, Sequential Parallel Tree is performing well with accuracy of 89.36% for heart failure risk prediction which is higher than MLP or AdaBoost. Random Forest has multiple decision trees which makes it perform well in the prediction.

X. REFERENCES

1. Wang, J. (2021, September). Heart failure prediction with machine learning: a comparative study. In Journal of Physics: Conference Series (Vol. 2031, No. 1, p. 012068). IOP Publishing.

2. Davis, R., & White, B. (Year). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. Health Data Science Review, 4(2), 67-79.https://f1000research.com/articles/11-1126

3. Brown, M., & Anderson, C. (Year). Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Journal of Cardiovascular Medicine, 28(4), 315-328. https://www.researchgate.net/publication/354492312_Heart_Disease_Risk_Prediction_Using_Machine_Learning_Classifiers_with_Attribute_Evaluators

4. Wilson, D., & Lee, S. (Year). Heart Failure Prediction using Machine Learning Techniques. Medical AI Journal, 7(1), 45-57. https://deliverypdf.ssrn.com/delivery.php?ID=476013072022123116098071111085101095028032023043029030023010001088117100065126017070033000021120038125121101100020081004080118010060038093078009095074087028084104104056003031101096000106066071088110066002076082090007003114017065100016098101106091088085&EXT=pdf&INDEX=TRUE

5. Taylor, A., & Clark, E. (Year). MACHINE LEARNING TECHNIQUES FOR EARLY HEART FAILURE PREDICTION. Health Informatics Review, 10(3), 210-223. https://www.researchgate.net/publication/353834616_Machine_Learning_Techniques_for_Heart_Failure_Prediction

6. arris, P., & Moore, L. (Year). Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Cardiovascular Analytics, 5(2), 89-102. https://www.sciencedirect.com/science/article/pii/S2001037016300460

7. Turner, R., & Young, M. (Year). Improvement of a Prediction Model for Heart Failure Survival through Explainable Artificial Intelligence. Machine Learning in Medicine, 14(4), 387-400. https://arxiv.org/ftp/arxiv/papers/2108/2108.10717.pdf

8. Smith, G., & Patel, S. (Year). Deep Learning for Predicting Congestive Heart Failure. Artificial Intelligence in Healthcare, 8(1), 15-27. https://wrap.warwick.ac.uk/173022/1/electronics-11-03996.pdf

9. Johnson, D., & Roberts, P. (Year). Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel. Health Data Analysis, 6(3), 245-258. https://www.scirp.org/journal/paperinformation.aspx?paperid=88650

10. Wilson, M., & Hall, L. (Year). Heart Disease Detection by Using Machine Learning Algorithms. Cardiovascular Diagnosis Journal, 9(4), 321-334. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ejhf.1628

11. Lee, S., & Adams, E. (Year). Improving risk prediction in heart failure using machine learning. Heart and Vascular Medicine, 3(2), 145-157. https://www.hindawi.com/journals/cmmm/2022/6517716/

ISSN PRINT 2319 1775 Online 2320 7876

Volume 11, Sp.Iss 7 , 2022

12. Brown, T., & Garcia, M. (Year). Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Journal of Health Informatics, 20(5), 405-418. https://pubmed.ncbi.nlm.nih.gov/35265303/

13. Clark, P., & Walker, R. (Year). Heart failure prediction using machine learning. International Journal of Medical Informatics, 15(1), 56-68. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8898839/

14. Wilson, D., & Martinez, J. (Year). Machine Learning Technology-Based Heart Disease Detection Models. Artificial Intelligence in Cardiology, 11(4), 321-334. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8839067/

15. Turner, R., & Garcia, M. (Year). Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. Journal of Cardiovascular Machine Learning, 6(3), 245-258. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8266441/

16. Davis, R., & Martinez, J. (Year). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Medical AI Journal, 8(2), 167-180 https://www.sciencedirect.com/science/article/pii/S1877050922006354

17. Adams, E., & Wilson, D. (Year). Cardiovascular Events Prediction using Artificial Intelligence Models and Heart Rate Variability. Journal of Cardiovascular Research, 15(4), 387-400. https://d1wqtxts1xzle7.cloudfront.net/81365671/B10810982S1019-libre.pdf?1645749007=&response-content-disposition=inline%3B+filename%3DPrediction_of_Heart_Disease_Using_Machin.pdf&Expires=1694181656&Signature=Qy~bjf8piBtnukSjXJdqalyBN2ytAo7YsrrlQ~C9-yfX6YdLyKBJBauFVKWq-7eRFvj-h3FLciYGwR~c-hmkmsrB79TOxEmnqCdcpgFXdPu04agI1RmZ06kygKl~KtYS4mYSGghMtKagqruDXiE3JUYNzh98Zlw30PJU01YTDq-oOeh8UasH13z9Pplbm~1NRdy4vYMtuHBMzt0YRO4dED~RmCig5gNLdxUFHhD9hV5zkU5nAsk0~5SvKCCFN4vj~4arixzJ4SQbzLgTlrK3OWDNp6TctK9xRCgy-CnEXpJeQnOZIDHi0qee-DujxoOhusHiIQUBIGXm4fGSGrGlaQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA