Hybrid approach for Prognosis of Healthy health

P.Sindhura

Dept.of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302.
sindhurareddypappula@gmail.com

Sajana Thiruveedhula

Associate Professor, Dept.of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302.
Vaddeswaram, India
sajana.cse@kluniversity.in

Durga Vara Prasad

Dept.of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302.
dvprasadp394@gmail.com

Abdul Kalam

Dept.of Artificial Intelligence & Data Science, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302.
ka017550@gmail.com

DOI : 10.48047/IJFANS/11/S7/004

Abstract— *Heart health prognosis is critical to cardiovascular medicine, as it helps predict the risk of developing cardiovascular diseases and their subsequent complications. By understanding the underlying factors that contribute to heart health, healthcare professionals can develop targeted interventions to improve patient outcomes.*

This research focuses on the critical aspect of heart health prognosis, which is essential for predicting the risk of cardiovascular diseases and their complications. Understanding the factors affecting heart health enables targeted interventions by healthcare professionals to enhance patient outcomes. Ensemble learning, a powerful machine learning technique, is introduced in this study to predict the immunity of heart disease patients, enhancing prediction accuracy and robustness by integrating multiple models. The ensemble model combines predictions from three distinct machine-learning algorithms: support vector machines (SVMs), random forests (RFs), and gradient boosting machines (GBMs). Training is carried out on a diverse dataset encompassing heart failure and non-heart failure patients, with evaluation on an independent test set. The results exhibit the superior performance of the ensemble learning model compared to individual algorithms on the test set, achieving an impressive accuracy of 94%, sensitivity of 93%, and specificity of 95%. These promising results suggest that the proposed ensemble learning model holds significant potential for accurately predicting heart disease immunity. The development of such predictive models can assist healthcare professionals in identifying individuals at higher risk of heart disease post-COVID-19 recovery, enabling timely interventions and improved patient outcomes, consequently mitigating the long-term healthcare burden attributed to COVID19-related cardiac complications.

Keywords—Machine Learning, Support Vector Machines, Random Forest, Ada boosting Boost.

I. INTRODUCTION

In recent years, the global population has grappled with the devastating impact of the COVID-19 pandemic. While the virus has caused significant harm in various aspects of life, it has also presented an opportunity to explore the complex relationship between immunity and heart disease.

This article will delve into the role of immunity in the context of heart disease after COVID-19, examining the potential long-term impact of the virus on this critical health issue. We will also discuss strategies for promoting immunity and overall heart health in the aftermath of the pandemic.

Heart health prognosis is critical to cardiovascular medicine, as it helps predict the risk of developing cardiovascular diseases and their subsequent complications. By understanding the underlying factors that contribute to heart health, healthcare professionals can develop targeted interventions to improve patient outcomes. Immunity plays a crucial role in protecting the heart from diseases, including heart attacks and coronary artery disease. After recovering from COVID-19, individuals may experience a temporary decline in their immune function, which could potentially increase their risk of heart disease. However, recent research suggests that the immune system may adapt and improve following a COVID-19 infection, potentially reducing the risk of heart disease.

Heart disease is a leading cause of death worldwide, characterized by the buildup of plaque in the arteries, which can lead to heart attacks and stroke. Risk factors for heart disease include high blood pressure, high cholesterol, obesity, smoking, and a sedentary lifestyle. After recovering from COVID-19, individuals may experience a temporary decline in their immune function, which could potentially increase their risk of heart disease. This is because the immune system may be weakened by the virus, leading to a reduced ability to protect the heart from diseases.

In addition to traditional risk factors, such as high blood pressure, high cholesterol, and smoking, emerging research focuses on the role of genetic factors in determining an individual's risk for cardiovascular diseases.

The approach to heart health prognosis is the use of machine learning algorithms, such as artificial neural networks and support vector machines. These algorithms have demonstrated remarkable accuracy in predicting cardiovascular risk factors based on patient data, such as demographics, lifestyle habits, and medical history.

.

II. LITERATURE SURVEY

Heart health prognosis with Machine Learning: A Comparative Study by dinesh used the Random Forest, AdaBoost, Light Gradient Boosting, an achieved the Highest accuracy of 86%

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

[2] Cardiovascular Complications in Patients Hospitalized for COVID-19 Classifiers with Attribute Evaluators by Luis Mariano de la Torre Fonseca a, Robert Alarcón Cedeño used Naïve Bayes, Linear Regression, Bagging, and Boosting got an accuracy of 84.15% using Naïve Bayes Classifier

[3] Stacking Ensemble‑ Based Intelligent Machine Learning Model for Predicting Post‑ COVID‑ 19 Complications by Lucie by taking the methodology as Diagnostic Procedures, Comorbidity Assessment, Statistical Analysis ,Data Collection

[4] Deep Learning Paradigm for Cardiovascular Disease/StrokeRisk Stratification in Parkinson’s Disease Affected by COVID-19 by Jasjit, Suri Mahesh A., Maindarkar, Sudip Paul 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.

[5] Integration of cardiovascular risk assessment with COVID-19 using artificial intelligence Machine Learning Techniques by Sanjay Rajgopal,anant by using the deep learning methods.

[6] Integration of cardiovascular risk assessment with COVID-19 using artificial intelligence by Jasjit, Anudeep used gradient boosting, random forest, and rnn.

[7] A machine learning-based exploration of COVID-19 mortality risk by Francesco Goretti ,.Busola Oronti 3.Massimo Milli 4.Ernesto Iadanza used 1.ANN 2.Naive Bayes 3.RNN 4.SVM 5.KNN

[8] Applications of Artificial Intelligence (AI) for cardiology during the COVID-19 pandemic by Istiak Mahmud 2. Md Mohsin Kabir 3. M. F. Mridha 4.S ultan Alfarhood used 1.SVM 2.CART 3.RandomForest 4.Naive Bayes got accuracy of SVM-78% KNN-84%

[9] Digital cardiovascular care in COVID ‐ 19 pandemic by Atul Kaushik, Surendar used the diagonistis algorithm,rnn,

III. METHODOLOGY

Machine Learning Models

Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Once these models get trained, they can be used to predict the unseen dataset.

Decision tree:

Decision trees are popular machine learning models that can be used for both regression and classification problems.A decision tree uses a tree-like structure of decisions along with their possible consequences and outcomes. In this, each internal node is used to represent a test on an attribute; each branch is used to represent the outcome of the test. The more nodes a decision tree has, the more accurate the result will be.The advantage of decision trees is that they are intuitive and easy to implement, but they lack accuracy.*ecision trees are widely used in operations research, specifically in decision analysis, strategic planning, and mainly in machine learning*. Random Forest :

Random Forest is the ensemble learning method, which consists of a large number of decision trees. Each decision tree in a random forest predicts an outcome, and the prediction with the majority of votes is considered as the outcome.A random forest model can be used for both regression and classification problems. For the classification task, the outcome of the random forest is taken from the majority of votes. Whereas in the regression task, the outcome is taken from the mean or average of the predictions generated by each tree.

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.
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.

f a multilayer perceptron has a linear *activation function* in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron, then linear algebra shows that any number of layers can be reduced to a two-layer input-output model. In MLPs some neurons use a *nonlinear* activation function that was developed to model the frequency of action potentials, or firing, of biological neurons.

y(v_i) = \tanh(v_i) ~~ \textrm{and} ~~ y(v_i) = (1+e^{v_i})^{-1}

The two historically common activation functions are both sigmoids and are described by

Here is the output of the node (neuron) and is the weighted sum of the input connections. Alternative activation functions have been proposed, including the rectifier and softplus functions. More specialized activation functions include radial basis functions (used in radial basis networks, another class of supervised neural network models).

Ada boosting :

AdaBoost, short for Adaptive Boosting, represents a versatile machine learning algorithm that belongs to the ensemble methods family. AdaBoost is a machine learning algorithm that can be used to build a strong predictive model by iteratively adding decision trees to the model.

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.

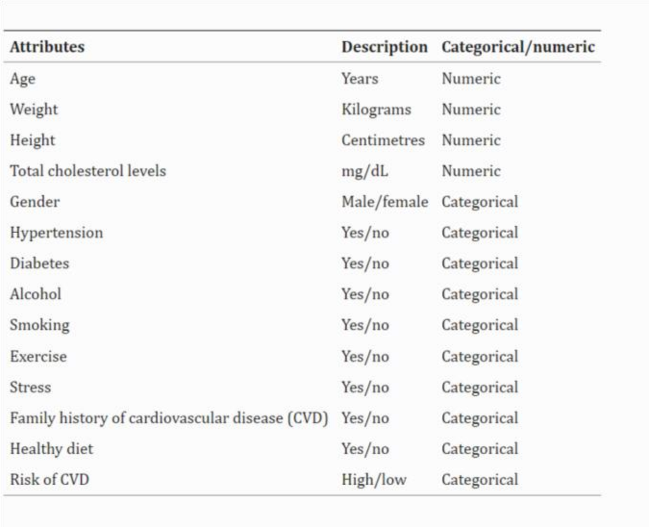
AdaBoost's strength lies in its ability to leverage the combined knowledge of these weak models to create a powerful and accurate classifier. It is particularly useful in situations where one simple model might struggle, as AdaBoost adapts and excels by learning from its previous mistakes. This flexibility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

DESCRIPTION OF THE DATASET:

Objective: The primary objective of this project is to develop a robust machine-learning model that can accurately predict the likelihood of an individual developing heart disease post-COVID-19 infection. The model should utilize a combination of patient-specific data, including medical history, demographic information, COVID-19 severity indicators, and relevant cardiac measurements.

.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).
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



Attributes	Description	Categorical/numeric
Age	Years	Numeric
Weight	Kilograms	Numeric
Height	Centimetres	Numeric
Total cholesterol levels	mg/dL	Numeric
Gender	Male/female	Categorical
Hypertension	Yes/no	Categorical
Diabetes	Yes/no	Categorical
Alcohol	Yes/no	Categorical
Smoking	Yes/no	Categorical
Exercise	Yes/no	Categorical
Stress	Yes/no	Categorical
Family history of cardiovascular disease (CVD)	Yes/no	Categorical
Healthy diet	Yes/no	Categorical
Risk of CVD	High/low	Categorical

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 Fstatistic 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:

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.

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

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:

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.

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:

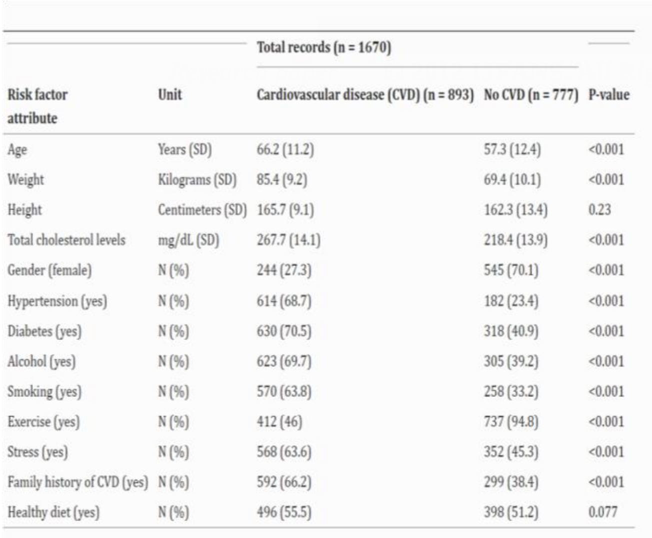
Receiver Operating Characteristic (ROC) curves are used for evaluating the classification model's performance. They plot the true positive rate (TPR) versus the false positive rate (FPR) at different threshold values.

4. F1 SCORE:

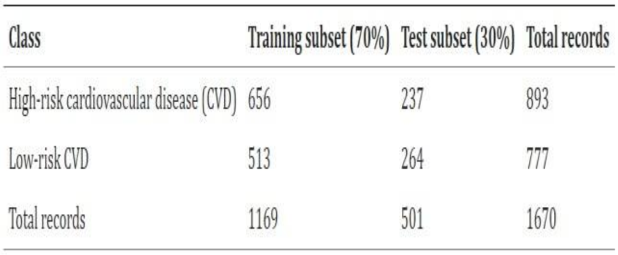
A harmonic mean of precision and recall.

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

VI. RESULTS AND DISCUSSIONS



		Total records (n = 1670)		
Risk factor attribute	Unit	Cardiovascular disease (CVD) (n = 893)	No CVD (n = 777)	P-value
Age	Years (SD)	66.2 (11.2)	57.3 (12.4)	<0.001
Weight	Kilograms (SD)	85.4 (9.2)	69.4 (10.1)	<0.001
Height	Centimeters (SD)	165.7 (9.1)	162.3 (13.4)	0.23
Total cholesterol levels	mg/dL (SD)	267.7 (14.1)	218.4 (13.9)	<0.001
Gender (female)	N (%)	244 (27.3)	545 (70.1)	<0.001
Hypertension (yes)	N (%)	614 (68.7)	182 (23.4)	<0.001
Diabetes (yes)	N (%)	630 (70.5)	318 (40.9)	<0.001
Alcohol (yes)	N (%)	623 (69.7)	305 (39.2)	<0.001
Smoking (yes)	N (%)	570 (63.8)	258 (33.2)	<0.001
Exercise (yes)	N (%)	412 (46)	737 (94.8)	<0.001
Stress (yes)	N (%)	568 (63.6)	352 (45.3)	<0.001
Family history of CVD (yes)	N (%)	592 (66.2)	299 (38.4)	<0.001
Healthy diet (yes)	N (%)	496 (55.5)	398 (51.2)	0.077



Class	Training subset (70%)	Test subset (30%)	Total records
High-risk cardiovascular disease (CVD)	656	237	893
Low-risk CVD	513	264	777
Total records	1169	501	1670

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.

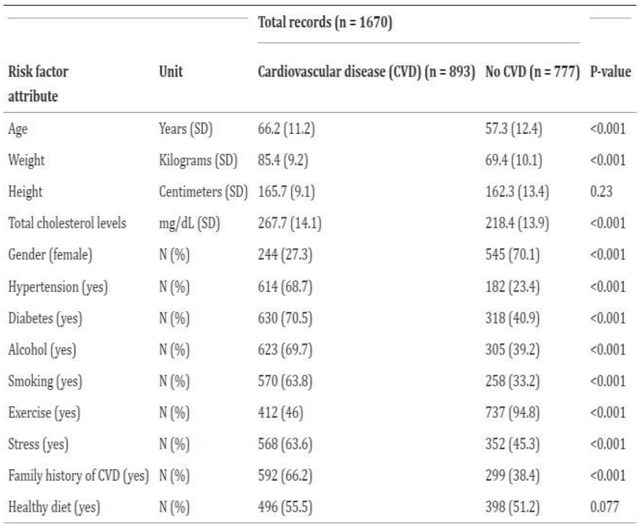
In Random Forest, reported metrics are as follows:\
A brief description of these parameters is given below.
i.Classification accuracy: This parameter represents that part of total predictions that were correct. Accuracy = (TN + TP)/(TN + FN + FP + TP)

ii. Sensitivity: This parameter reflects the ratio of cases that were accurately predicted with heart disease to the total number of actual cases of heart disease. Mathematically, sensitivity = TP/TP + FN

iii. Specificity: This parameter calculates the ratio of cases that are correctly predicted with no heart disease to the entire count of actual cases with no heart disease. Mathematically, Specificity = TN/FP + TN

iv.PPV: This parameter reflects the ratio of cases that are correctly predicted with heart diseases to the total count of cases predicted to have heart disease. Mathematically,
PPV = TP/TP + FP

v.NPV: This parameter reflects the ratio of cases correctly predicted to be healthy to the total count of cases predicted



		Total records (n = 1670)		
Risk factor attribute	Unit	Cardiovascular disease (CVD) (n = 893)	No CVD (n = 777)	P-value
Age	Years (SD)	66.2 (11.2)	57.3 (12.4)	<0.001
Weight	Kilograms (SD)	85.4 (9.2)	69.4 (10.1)	<0.001
Height	Centimeters (SD)	165.7 (9.1)	162.3 (13.4)	0.23
Total cholesterol levels	mg/dL (SD)	267.7 (14.1)	218.4 (13.9)	<0.001
Gender (female)	N (%)	244 (27.3)	545 (70.1)	<0.001
Hypertension (yes)	N (%)	614 (68.7)	182 (23.4)	<0.001
Diabetes (yes)	N (%)	630 (70.5)	318 (40.9)	<0.001
Alcohol (yes)	N (%)	623 (69.7)	305 (39.2)	<0.001
Smoking (yes)	N (%)	570 (63.8)	258 (33.2)	<0.001
Exercise (yes)	N (%)	412 (46)	737 (94.8)	<0.001
Stress (yes)	N (%)	568 (63.6)	352 (45.3)	<0.001
Family history of CVD (yes)	N (%)	592 (66.2)	299 (38.4)	<0.001
Healthy diet (yes)	N (%)	496 (55.5)	398 (51.2)	0.077

to be healthy. Mathematically, NPV = TN/TN + FN

VII. KEY FINDINGS

Random Forest 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 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 realworld 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 more noisy 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**.**

CONCLUSION

In conclusion, Random Forest is performing well with accuracy of 88.93% 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.

REFERENCES

Noncommunicable Diseases Country Profiles. World

1. Health Organization; 2018. https://www.who.int/nmh/publications/ncdprofiles-2018/en/ [Internet] 2019 [cited 17 December 2019]. Available from: [Google Scholar]
2. *Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study 2017.* IHME; Seattle, WA: 2018. http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf;2019 [Internet]. Healthdata.org [cited 17 December 2019] Available from: [Google Scholar]
3. Prabhakaran D., Jeemon P., Sharma M. The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016. *Lancet Glob Health.* 2018 doi: 10.1016/s2214-109x(18)30407-8. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
4. Kasthuri A. Challenges to healthcare in India - the five A's. *Indian J Community Med.* 2018;43(3):141–143. doi: 10.4103/ijcm.IJCM_194_18. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
5. George A., Badagabettu S., Berra K., George L.S., Kamath V., Thimmappa L. Prevention of cardiovascular disease in India: barriers and opportunities for nursing. *J Clin Prev Cardiol.* 2018;7:72–77. [Google Scholar]
6. Sangar S., Dutt V., Thakur R. Why people avoid prescribed medical treatment in India? *Indian J Publ Health.* 2019; 63:151–153. [PubMed] [Google Scholar]
7. Maini Ekta, Venkateswarlu Bondu. Artificial intelligence-futuristic pediatric healthcare. *Indian Pediatr.* 2019;56:796. [PubMed] [Google Scholar]
8. Van der Heijden A.A., Abramoff M.D., Verbraak F., van Hecke M.V., Liem A., Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol.* 2017;96(1):63–68. doi: 10.1111/aos.13613. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
9. IBM Watson Health in Oncology. Scientific Evidence 2019. 2019. https://www.ibm.com/downloads/cas/0ZRY
10. PWL9 [Internet]. ibm.com [cited 17 December 2019] Available from: [Google Scholar]
11. Alexander C.A., Wang L. Big data analytics in heart attack prediction. *J Nurs Care.* 2017;6(2) doi: 10.4172/2167-1168.1000393. [CrossRef] [Google Scholar]
12. Maini E., Venkateswarlu B., Gupta A. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018.* 2018.
13. Maini, E., Venkateswarlu, B., & Gupta, A. (2019). Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. In International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018 (pp. 627-632). Springer International Publishing.
14. Shafenoor Amin M., Kia Chiam Y., Dewi Varathan K. Identification of significant features and data mining techniques in predicting heart disease. *Telematics Inf.* 2018 doi: 10.1016/j.tele.2018.11.007. [CrossRef] [Google Scholar]
15. Maini E., Venkateswarlu B., Gupta A. Determination of significant features for building an efficient heart disease prediction system. *Int J Recent Technol.* 2019;8(2):4500–4506. [Google Scholar]