## ENSEMBLE MACHINE LEARNING APPROACH TO CARDIOVASCULAR DISEASES RISK PREDICTION

## Introduction

Cardiovascular diseases are conditions that affect the structures or function of your heart, such as:

* + Abnormal heart rhythms, or arrhythmias
  + Aorta disease and Marfan syndrome
  + Congenital heart disease
  + Heart attack
  + Heart failure
  + Heart muscle disease (cardiomyopathy)
  + Stroke etc

Cardiovascular diseases are Non-Communicable Diseases (NCD) like cancers, chronic respiratory diseases, diabetes and Mental Health Disorders. They are the world’s biggest killers and have now been termed “a silent epidemic”. These diseases share common risk factors namely; tobacco use, harmful use of alcohol, unhealthy diet, physical inactivity and air pollution(WHO, 2019). More than 36 million people die annually from NCDs accounting for 63% of all global deaths. Low- and middle-income countries unfortunately, already bear 86% of the burden of these premature deaths giving rise to estimated cumulative economic losses of US$7 trillion over the next 15 years (WHO, 2019).

Cardiovascular diseases (CVDs) are the number 1 cause of death accounting for 17.5 million deaths annually with high blood pressure being the leading risk factor. CVD’s are disorders of the heart plus blood vessels and they include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Four out of five CVD deaths are due to heart attacks and strokes(WHO, 2019).According to Dr. Nnenna Ezeigwe, the National Coordinator for NCD’s in the Federal Ministry of Health (FMoH ), “CVD is a significant public health concern responsible for 11% of over 2 million NCD deaths in Nigeria annually. It is also responsible for a high burden of morbidity and disability. Most people with CVDs are not aware until catastrophes like stroke, heart attack or death occur.”(WHO, 2019)

## Background of the Study

Heart disease (HD) is one of the most common diseases nowadays, due to number of contributing factors, such as high blood pressure, diabetes, cholesterol fluctuation, exhaustion and many others. An early diagnosis of such disease has been sought for many years, and many data analytics tools have been applied to help health care providers to identify some of the early signs of HD.

Data Mining has proved itself to be very effective in forecasting diverse scenarios for numerous fields. With data mining and deep learning paired with each other, many models have been designed to forecast specific scenarios for us to operate on them. Due to the need of making accurate prediction of heart disease machine learning algorithms are used in order to point out prediction based on many factors, the human mind cannot process too much estimation and can therefore provide incorrect feedback several times, leading to vital risk to the patient.

Describe the problem they address

What they did

What method they adopt

Weakness of the method

According to a research carried out by (Miao *et al.* 2016), they developed an advanced ensemble machine  
learning technology, utilizing an adaptive Boosting algorithm, developed for precise heart disease diagnosis and  
aftermath predictions. The developed ensemble learning classification and prediction models were applied to 4  
different data sets for heart disease diagnosis, including patients diagnosed with heart disease from Cleveland Clinic  
Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland  
University Hospital (SUH). The testing results showed that the advanced ensemble learning classification and  
prediction models accomplished model accuracies of 80.14% for CCF, 89.12% for HIC, 77.78% for LBMC, and  
96.72% for SUH. These results exceeded the accuracy of previously published research. According to the ensemble  
method applied for identifying valvular heart disease (Das and Sengur 2010), a higher sensitivity rate (97.3%) is  
obtained using the FCM-CHMM. However, FCM–CHMM created the 92% specificity value; the ensemble method  
produced the highest specificity rate (100%). SVM technique improvements the worst specificity rate (90%). ANN,  
WPNN, and LDA–ANFIS approaches achieve the same 94% specificity rate, and ANN, LDA–ANFIS, and PCA–AIS and fuzzy k-NN produce the same sensitivity rate (95.9%). Therefore, in the end, SVM and WPNN produce a lesser  
sensitivity rate (94.5%).

Researches have been done on the area of heart disease prediction using different tools like weka, sklearn and Matlab with different machine learning techniques like logistic regression, Support Vector Classifier, Random Forest, Decision trees and with deep learning models eg Convolutional Neural Network (CNN). Many tests can be performed on potential patients to take the extra precautions measures to reduce the effect of having such a disease [1], and reliable methods to predict early stages of HD, such as the methods proposed in this paper. The proposed in this paper is an ensemble model comprising of *Bayesan Network, Random forest Classifier, Naïve Bayes and Logistic regression,* which have proven very efficient in predicting heart disease. The ensemble method will combine all the predicting power of the models to make a single prediction

## Problem Statement

In terms of time, accuracy, and cost, medical dictation has always remained a high maintenance field. Human beings are susceptible to mistakes and can make errors. Cases of cardiovascular diseases are rising at an exponential pace, and that is very troubling and early detection is crucial in preventing the progression of the disease. Patients usually visits a hospital when the disease reaches an advanced stage. This leads to increase in the cost of treatment and also weakens the chances of recovery and often leads to loss of life. Cardiac diseases can be prevented if these are diagnosed timely. The importance of developing a constraint-free, reliable and timely prediction system has long been recognized by the health sector. Machine learning based prediction system can prove to be an effective tool in detecting the heart diseases in an early stage.

## Objective of the Study

The main aim of the paper is to develop an accurate and fast hybrid method for prediction of heart disease. To achieve these, the work considered the following specific objectives:

* Study the best performing models in heart disease prediction.
* Combine the strength of the models (an ensemble model) to overcome the weakness of an individual model
* To compare results obtained from the ensemble model and that of the individual models.

## Literature Review

Research has been done in this field and people have produced methods to predict cardiovascular disease using supervised machine learning algorithms. Several research papers have been written on this topic. A short survey has been in which analyzes performance of various models based on machine learning algorithms and techniques is shown below:

In one of the works by (Prasad, 2020) , Bayesan network was used to predict heart disease. Bayesian networks apply Bayes’ Theorem (also known as Bayes’ rule or Bayes’ law). In Bayes’ theorem, a prior (unconditional) probability represents the likelihood that an input parameter will be in a particular state; the conditional probability calculates the likelihood of the state of a parameter given the states of input parameters affecting it; and the posterior probability is the likelihood that parameter will be in a particular state, given the input parameters, the conditional probabilities, and the rules governing how the probabilities combine. The network is solved when nodes have been updated using Bayes’ Rule:

P(A|B) =

Where P(A) is the prior distribution of parameter A; P(A|B) is the posterior distribution, the probability of A given new data B; and P(B|A) the likelihood function, the probability of B given existing data.

Although Bayesian Network are “white box” models i.e their reasoning process can be understood by humans and their answers always come with probabilities attached. Bayesian networks (BN) have been used to build medical diagnostic systems as doesn't produce a range outcome but rather a probability for each potential predicted event. Baysian network have some drawbacks ranging from being Computationally expensive. Eg: Approximate structure learning is too NP-Complete tending to perform poorly on high dimensional data, quality of the results of the network depends on the quality of the prior beliefs or model i.e A variable is only a part of a Bayesian network if you believe that the system depends on it.

In another paper by (Aniruddha Dutta, 2020) , Convolutional Neural Network (CNN) was used to make heart disease prediction. Convolutional Network is very useful in the field of image recognition. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. Convolutional Neural Network (CNN) is known to provide better outcomes than other machine learning algorithms if tuned better and feeded a good amount of data, It also poses some drawbacks ranging from being significantly slower due to an operation such as maxpool thereby making training process take a lot of time if the computer doesn’t consist of a good GPU, to the fact that it requires a large Dataset to process and train the neural network

In a work by (Mursal Furqan, 2020), algorithms like Logistic Regression Classifier, K-Nearest Neighbors Classifier, and Random Forest Classifier were individually used to predict heart disease with each resulted in its percentage accuracy. The accuracy gotten from all the models ranges from 72 -84 percent with logistic regression having the highest accuracy of 84%. Using method of running different algorithms is useful to know the performance of each model on a dataset for the major decision of which model prediction to be expected more accurate. Logistic regression which was the best performing model is one of the supervised Machine Learning algorithms used for classification i.e. to predict discrete valued outcome. Although logistic regression being **one of the simplest machine learning algorithms** and is easy to implement yet provides great training efficiency in some cases. It also possesses some drawbacks including model overfitting the training set on high dimensional datasets,it can best perform well when only important and relevant features are used to build the model otherwise the probabilistic predictions made by the model may be incorrect and the model's predictive value may degrade.

The table below shows comparative analysis with previous studies

## Literature Review

The table below shows some of the related literature and their weaknesses

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S/N** | **Title** | **Year** | **Author** | **Method Used** | **Evaluation Metrics** | **Weakness** |
|  | A fast algorithm for heart disease prediction using Bayesian network model | 2020 | Mistura Muibideen and Rajesh Prasad | Bayesian network (BN) | Accuracy, Precision, Recall, F1 Score | Bayesian Network requires a comparatively large amount of effort to be able to learn patterns in data. Bayesian learning is extremely **computationally expensive** |
|  | A Novel Approach to the Diagnosis of Heart Disease using Machine Learning and Deep Neural Networks | 2020 | Sahithi Ankireddy | K-Nearest  Neighbors (KNN), Support Vector Machine (SVM), Random  Forest Classifier (RF), Naive-Bayes (NB) and Deep Neural  Network (DNN) | ROC-AUC | It requires very large amount of data in order to perform better than other techniques. It is also extremely expensive to train due to complex data models. |
|  | Coronary Artery Disease Diagnosis; Ranking the Significant Features Using Random Trees Model | 2020 | Javad Hassannataj Joloudari , et al. | Random Trees, Decision Trees, support vector machine (SVM) | Accuracy, ROC curve, Gini, Gain and Confidence | The model used (Decision Tree) is easily leads to overfitting of the data that are particularly noisy |
|  | An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction | 2020 | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton | Convolutional Neural Network (CNN) | AUC | The training process takes a lot of time on non-GPU systems. It also requires a large Dataset to process and train the **neural network** |
|  | Improving the performance of heart disease prediction system using ensembling techniques | 2021 | Ekta Maini, and Bondu Venkateswarlu | Ensembling techniques (Naïve Bayes, SVM, Logistic Regression and and Multilayer Perceptron) | Accuracy | The merged ensemble method can be extremely **computationally expensive** due to Neural network used |
|  | Heart Disease Prediction using Machine Learning Algorithms | 2020 | Mursal Furqan, Hiba Rajput,Sanam Narejo | Logistic Regression Classifier, K-Nearest Neighbours  Classifier, and Random Forest Classifier | Accuracy | The models used can easily overfits data that are particularly noisy |
|  | Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means | 2021 | Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria | K-Means with PCA and  Hybrid Genetic Algorithm | Accuracy, Clustering Error, Recall, Precision,F1 Score | K-Means has a tendency of been trapped at the local optimum and GAs has a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem |
|  | Heart Disease Prediction Using Hybrid Machine Learning Algorithms | 2020 | S.Raguvaran  R.Anandhi  A.Anbarasi  T.Megala | Logistic  Regression, Random Forests Classifier Algorithm, Neural network, KNN (K-Nearest Neighbour) |  |  |
|  | Heart disease prediction using machine learning techniques | 2021 | Apurv Garg, Bhartendu Sharma and Rijwan Khan | K-Nearest Neighbor (K-NN) and Random  Forest | Accuracy |  |
|  | Prediction System for Heart Disease Based on Ensemble  Classifiers | 2020 | Joshua Emakhu, Sujeet Shrestha and  **Suzan Arslanturk** |  |  |  |

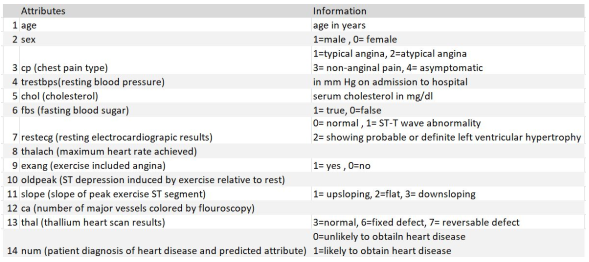
## Methodology

The primary research method for this study is literature review and conceptual modeling. This study will first review various research done under cardiovascular disease prediction and their characteristics. Based on this understanding, a classification method will be developed resolve the limitations of those research.

In order to accurately, easily and effectively make prediction for cardiovascular disease, the research plan will proceed in two phases. This first phase of the proposed method is developing an effective model to predicting heart disease and the second phase is deploying the model for user’s consumption. The first phase is made up of:

1. The dataset to be used
2. Data Retrieval
3. Data preprocessing
4. Handling Missing Values
5. Feature Selection
6. Target Class Transformation and
7. the model to be used for the study.
8. **Dataset**

This research is proposed to use the heart disease dataset from the UCI machine learning repository called Cleveland Heart Disease Data set. Cleveland Heart Disease Dataset is a publicly available supervised dataset provided by the Cleveland Clinic Foundation was used for the ML model. This data set contains 75 total attributes of patient medical information for 303 patients. 14 attributes out of the 75 were chosen. The figure below shows the chosen attributes, and its information. These attributes have been selected by other researchers and healthcare professionals because they are known to be the best determining factors of heart disease.



1. **Data Retrieval**

Data retrieval is usually the first step. Data can be gotten from various sources. It can be as easy as someone handing over a file on a drive for you to analyze them directly. Or you need to download it or issue a database query to collect the data. Our dataset is to be downloaded from the UCI machine repository.

1. **Data Preprocessing**

Data preprocessing is also known as cleaning data. It is one of the most important steps to achieve the best from the dataset. This is a process whereby data inconsistencies such as missing values, out of range values, unformatted data, and noise are removed from the data. The process is usually time-consuming because it involves a lot of experimentation trying out various data analysis tools. Our preprocessing would involve data retrieval, handling missing values, target class transformation and data discretization

1. **Handling Missing Values**

Missing data values is a common problem faced by analysts. This occurs due to different reasons such as incomplete extraction, corrupt data, failure to load the information, etc. This is a great challenge that must be fixed because good models are generated when you make the right decisions on how to fix it. These are 5 ways of handling missing data:

* 1. Deleting Rows ii.
  2. Replacing with mean/median/mode iii.
  3. Assigning a unique category iv.
  4. Predicting the missing values v.
  5. Using algorithms which supports missing values

We would adopt the method that proves best.

1. **Feature Selection**

In order to select the best features for our model. Recursive feature elimination is proposed. RFE is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

1. **Target Class Transformation**

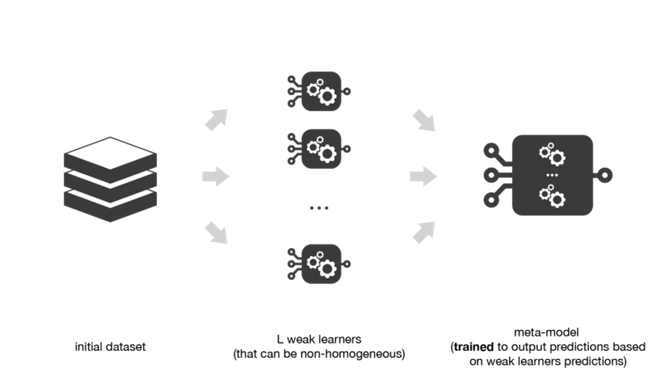
As stated in the data set description, the target class contains values (0, 1, 2, 3, 4). Where 0 means healthy (no heart disease) and (1, 2, 3, 4) means the presence of sickness of varying degrees. Interest is in the absence or presence of heart disease, so the need to limit the class to (0, 1). Level (1, 2, 3, 4) was converted to 1.

1. **Proposed Model**

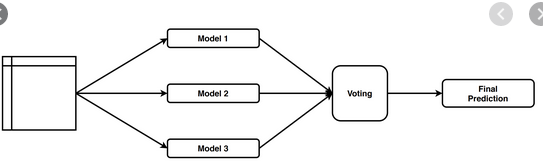
An ensemble model made from different algorithms is proposed. Ensemble models are an ensemble learning method that combines different algorithms together. In this sense, it is a meta-algorithm rather than an algorithm itself. Ensemble learning methods are valuable because they can improve the performance of a predictive model. The value of an ensemble classifier is that, in joining together the predictions of multiple classifiers, it can correct for errors made by any individual classifier, leading to better accuracy overall.

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease** **variance** (bagging), **bias** (boosting), or **improve predictions** (stacking/super learning).

This work proposed the use of Stacking and voting ensemble. The idea of stacking is to learn from several different weak learners and combine them by training a meta-model to output predictions based on the multiple predictions returned by these weak models

Stacking consists in training a meta-model to produce outputs based on the outputs returned by some lower layer weak learners.

Voting ensemble involves summing the votes for class labels from other models and predicting the class with the most votes or summing the predicted probabilities for class labels and predicting the class label with the largest sum probability.



***Proposed Models to use***

The proposed models comprise of Bayesian Network, Random Forest Classifier, Multilayer perceptron (MLP) and Naïve Bayes. This is because the models proved to have high performance on predicting heart disease from previous studies.

Although ensemble machine learning techniques are difficult to interpret and have longer training and prediction time than single models, since you need the predictions of many models, they proved beyond doubt to provide better accuracy

1. **Performance Metrics**

Performance metrics are used to evaluate how different algorithms perform based on various criteria such as accuracy, precision, recall etc. They are discussed below.

***Confusion Matrix***

The confusion matrix shows the performance of the algorithm. It depicts how the classifier is confused while predicting. True positive value signifies that the positive value is correctly predicted, false positive means the positive value is falsely classified, false negative means the negative value is falsely predicted while the true negative means the negative value is correctly classified.

***Accuracy***

Accuracy is the ratio of the number of correctly classified instances to all the cases. It is the sum of TP and TN divided by the total number of instances.

***Precision***

Precision is the proportion of true positive instances that are classified as positive. It reflects the closeness of predicted values is to one another.

***Recall***

Recall is the proportion of positive instances that correctly classified as positive. Recall is known as sensitivity.

***F1 Score***

F1 score combines both precision and recall and finds a balance between both.

## References

1. India State-Level Disease Burden Initiative Collaborators,” Nations within a nation: variations in epidemiological transition across the states of India”, 1990–2016 in the Global Burden of Disease Study. Lancet 2017; 390:2437–60
2. <https://www.afro.who.int/news/who-and-nigerian-government-move-curb-cardiovascular-diseases>
3. Maini E, Venkateswarlu B, Gupta A. Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. In: Hemant J, Fernando X, Lafate P, Baig Z(eds). International Conference on Intelligent Data Communication Technologies and Internet of Things ICICI 2018, Lecture Notes on Data Engineering and Communications Technologies. 2019 (26):627-32
4. Aniruddha Dutta, T. B. (2020). An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction .
5. Ankireddy, S. (2020). A Novel Approach to the Diagnosis of Heart Disease using Machine Learning and Deep Neural Networks.
6. Mursal Furqan, H. R. (2020). Heart Disease Prediction using Machine Learning Algorithms . *2nd International Conference on Computational Sciences and Technologies , 17-19 Dec 2020 (INCCST 20).* MUET Jamshoro.
7. Prasad, M. M. (2020). A FAST ALGORITHM FOR HEART DISEASE PREDICTION USING BAYESIAN NETWORK MODEL.