

Predictive Analysis of Heart Disease Based on Machine Learning Approaches

Shaik Shameeda¹, S. Vasundra²

¹PG Scholar, Department of CSE, Jawaharlal Nehru Technological University College Of Engineering (Autonomous) Anantapuramu

²Professor, Department of CSE, Jawaharlal Nehru Technological University College Of Engineering (Autonomous) Anantapuramu

Abstract

Heart disease, alternatively known as cardiovascular disease, indicates various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and it is needed to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Predicting and diagnosing heart disease is the biggest challenge in the medical industry and relies on factors such as the physical examination, symptoms and signs of the patient. Machine learning algorithms play an essential and precise role in the prediction of heart disease. A hybrid machine learning approach is used to predict stroke via imbalanced and in complete medical data set. The existing system uses a hybrid approach model by combining the characteristics of Random Forest and Linear model approaches collectively termed as HRFLM (Hybrid Random Forest Linear Model). This model makes use of all the features without any restrictions while selecting them and uses artificial neural networks with back propagation concept. Heart disease dataset is collected from UCI machine learning repository with 13 clinical features as input. The Cleveland dataset contains an attribute with the name num to show the diagnosis of the heart disease in patient on different scales from 0 to 4. The proposed system uses other combination of hybrid approach by combining RBF SVM along with Logistic regression. RBF SVM uses kernel function to solve non-linear problems and Logistic regression provides great training efficiency for timely improving the diagnosis of the heart disease.

Keywords— Machine Learning, Prediction, Classification Technique, Random Forest, Decision Tree, Feature Selection, Prediction Model, Cardiovascular Disease (CVD), Radial Basis Function

I. INTRODUCTION

Machine learning is a method of data analysis that automates analytical model using a set of algorithms which are performed automatically with provided user data. As ML is one of the sections of artificial intelligence which provides a series of steps through which user interacts with training and learning of datasets, various patterns of datasets to make automatic decisions with minimal human intervention. Now a days ML is widely used in many applications such as medicine, Statistics, Agriculture, Aviation, Speech Recognition etc., Through various ML Conventional Algorithms all industrial and other sectors data is used to perform needed tasks automatically without maximum user interaction.

Now a days ML is widely for various diseases prediction accurately with provided and trained datasets. This paper provides is a study of Predictive Analysis of Heart Disease Based on Machine

Learning Approaches. As cardiovascular disease is the kind of disease which can cause the emergency if not predicted early. Many people are losing their life's due to false predictions and later stages predications. As heart disease is a defect related coronary decency which can be occurred due to various reasons in the heart like weakened walls, blockages, insufficient blood supply to arteries. To make a better and faster analysis now days Machine learning (ML) a branch of artificial intelligence (AI) is increasingly utilized within the field of cardiovascular medicine for better, faster and accurate analysis.

It is essentially how computers make sense of data and decide or classify a task with or without human supervision. The conceptual framework of ML is based on models that receive input data (e.g., images or text) and through a combination of mathematical optimization and statistical analysis predict outcomes (e.g., favorable, unfavorable, or neutral). Several ML algorithms have been applied to daily activities. As an example, a common ML algorithm designated as SVM can recognize non-linear patterns for use in facial recognition, handwriting interpretation. Too many automated techniques to detect the heart disease are implemented like data mining, machine learning, deep learning, etc. This paper will provide brief introduction about machine learning techniques. In this paper we train datasets using the machine learning repositories. There are some risk factors based on which the heart disease is predicted. Risk factors are: Age, Sex, Blood pressure, Cholesterol level, Family history of coronary illness, Diabetes, Smoking, Alcohol, Being overweight, Heart rate, Chest Pain.

So-called boosting algorithms used for prediction and classification have been applied to the identification and processing of spam email. Another algorithm, denoted random forest (RF), can facilitate decisions by averaging several nodes [5]. While convolutional neural network (CNN) processing, combines several layers and applies to image classification and segmentation. Previously described technical details of each of these algorithms are implemented, but no consensus has been emerged to guide the selection of specific algorithms for clinical application within the field of cardiovascular medicine.

The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naïve Bayes (NB) [1], [2]. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

Although selecting optimal algorithms for research questions and reproducing algorithms in different clinical datasets is feasible, the clinical interpretation [3] and judgement for implementing algorithms are very challenging. A deep understanding of statistical and clinical knowledge in ML practitioners is also a challenge. Machine learning algorithms play an essential and precise role in the prediction of heart disease. HML (Hybrid Machine Learning) is an advancement of the ML workflow that combines different algorithms and processes. Most ML studies reported a discrimination measure such as the area under K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naïve Bayes (NB) All models make use of all the features without any restrictions while selecting them and uses Artificial neural networks with back propagation concept. The said algorithms can diagnosis heart disease in patient on different scales from 0 to 4. Most importantly, an acceptable cutoff for different scales to be used in clinical practice, interpretation of the cutoff, and the appropriate/best algorithms to be applied in cardiovascular datasets remain to be

evaluated.

Specialists previously proposed the methodology to conduct ML research in medicine. Systematic review and meta-analysis, the foundation of modern evidence-based medicine, have to be performed in order to evaluate the existing ML algorithm in cardiovascular disease prediction. Here, we performed the first systematic review and meta-analysis of ML research over a million patients in cardiovascular diseases [6][7][8]. Our proposed system uses other combination of hybrid approach by combining RBF SVM along with Logistic regression. RBF SVM uses kernel function to solve non-linear problems and Logistic regression provides great training efficiency for timely improving the diagnosis of the heart disease.

II LITERATURE SURVEY

The number of works has been done related to disease prediction systems using different machine learning algorithms in medical Centers. Senthil Kumar Mohan et al,[10] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. In this strategy the objective is finding the critical condition by applying Machine Learning concepts, aiming about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. This concept produced an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) [9] they likewise educated about Diverse data mining approaches and expectation techniques, such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Sonam Nikhar et al [11] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point-by-point portrayal of Naive Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system.

Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), [3] Prediction of Heart Disease Using Machine Learning, In this paper the proposed system uses the neural network algorithm and multi-layer perceptron (MLP) to train and test the dataset. This algorithm will be having multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through the hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b , which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement.

Abhay Kishore et al,[4] developed Heart Attack Prediction Using Deep Learning. This paper proposes a heart attack prediction system by using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that implemented based on Deep Learning approach in Artificial Neural Network. The paper talks in detail about the significant modules of the framework alongside the related hypothesis. The proposed model uses deep learning and data mining concepts to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another way of heart attack prediction platform.

Lakshmana Rao et al,[14] Machine Learning Techniques for Heart Disease Prediction in which the

contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc.). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the severity of heart disease among people. The idea of CHD identification is difficult, in addition the disease must be dealt with warily. Not doing early identification, may impact the heart or may cause sudden death. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning a procedure that causes the framework to gain from past information tests, models without being expressly customized. Machine learning makes rationale dependent on chronicled information.

Mr. Santhana Krishnan.J and Dr. Geetha.S, [15] Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The idea about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in detail in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through this techniques heart disease is predicted in this System. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques.

Avinash Golande et al,[16] proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease. Usually utilized methodology utilized are decision tree, k- closest and Naive Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self- arranging guide and SVM (Bolster Vector Machine). The following area obviously gives subtleties of systems that were utilized in the examination.

V.V. Ramalingam et Al,[17] proposed heart disease prediction using machine learning techniques in which Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance.

Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers and systems have been applied to different clinical datasets to robotize the investigation of huge and complex information. Numerous scientists, as of late, have been utilizing a few Machine Learning algorithms and techniques. They have been applied to various medical datasets to automate the analysis of large data.

Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and in the diagnosis of heart related diseases. This paper provides a survey of various models based on various algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K- Nearest Neighbor (KNN), Naive Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers. strategies to enable the wellbeing to mind industry and the experts in the analysis of heart related sicknesses.

This paper provides a review of different models dependent on such calculations and methods and

analyze their performance. Models in light of directed learning calculations, for example, Support Vector Machines (SVM), K- Nearest Neighbor (KNN), Navy Bayes, Decision Trees (DT), Random Forest (RF) and group models are discovered extremely well known among the scientists.

III. BACKGROUND METHODS

Machine learning is a hot topic in research and industry, with new methodologies developed all the time. The speed and complexity of the field makes keeping up with new techniques difficult even for experts and potentially overwhelming for faster analysis.

Logistic regression

Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. It is one of the supervised learning and is used to estimate the target object value's possibility. It is a tool to calculate the statistical values and make results on binary output. In the linear method, which is calculated by the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts $P(Y=1)$ as a function of X . Here, y is the linear model's output trained with logistic regression produce value between zero and one.

Naïve Bayes

In the Naïve Bayes network, all features are independent. Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. When there is a change in one feature, it does not affect another. This is suitable for large datasets. The assumption from Conditional independence is that an attribute value is independent of the values, which are from other attribute values in a class. Bayes' Theorem is based on probability theory.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as: Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Support Vector Machine (SVM)

SVM is used both for regression and classification tasks. The SVM model represents the data in the space described so that the examples in various categories are divided by a distance as large as possible. That divides sensitive information with the maximum separable space between them and is calculated so that many of the points belong to one group fall on the plane's one side.

Radial Basis Function (RBF)

An Artificial Neural Network that uses non linear Radia basis function as activation functions and gives linear output using combination of radial basis functions of the inputs and neuron parameters. RBF is mainly used in SVM classification, which maps input space in new dimensional space [12]. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function

used in various kernelized learning algorithms. It is the default kernel used within the sklearn's SVM classification algorithm. A kernel is a function that takes the original non-linear problem and transforms it into a linear one within the higher dimensional space.

KNN

K-Nearest Neighbor is an anti-parametric method, which is used for regression and classification. It is essentially a grouping method, consider the distance between a point and the coordinates (x, y) and its neighbors. The distance between the Euclidean its neighbors are determined from the point and eventually located in the region nearest to its neighboring points. The KNN algorithm assumes that the similar things exists in the nearest proximity.

IV. EXISTING ANALYSIS

Prediction using traditional methods and models like Hybrid Random Forest with a linear model(HRFLM) approach is used by combining the characteristics of Random Forest (RF) and Linear Method (LM) which involves various risk factors and it consists of various measures of algorithms such as datasets, programs and much more to add on. In existing techniques

HRFLM makes use of 13 clinical attribute features as the input and identify whether the patient has heart disease or not. The Cleveland dataset consists of 303 records with 14 attributes. The feature selection and modeling keep on repeating for various combinations of attributes. A very high and low risk patients will be classified on the basis of various tests which are done in the selected group, but proposed models are only based on clinical situations based which uses supervised learning methods for predictions.

In existing system algorithms like KNN, Navy Bayes, RF [13] and others which predicts the disease of various patients suffering from various symptoms. But the given performance measures of the proposed algorithms statistics are using various other tools but they are predicting up to 90% accuracy rate only. The information of patient statistics, results, disease history in recorded in EHR, which enables to identify the potential data centric solution, which reduces the cost of medical case studies. Existing system can predict the disease but not the sub type of the disease and it fails to predict the condition of the people, the predictions of disease have been indefinite and nonspecific.

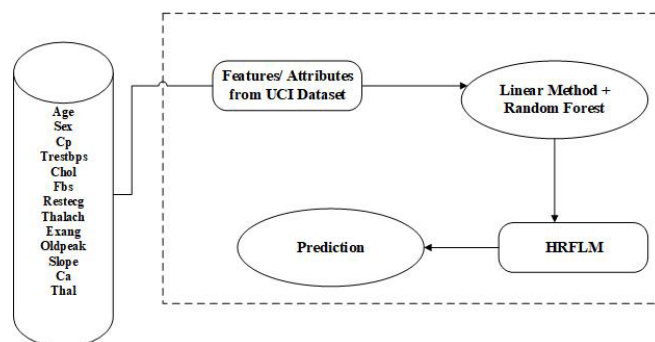


Fig 1: Existing Architecture showing HRFLM Prediction Process

Drawbacks:

- The effects of heart events are hard to forecast.
- Information systems would not contribute towards

- Statistical methods for medical information are too Heavy
- This model makes use of all the features without any restrictions while selecting them.
- Prediction accuracy is less.

V. PROPOSED WORK

In our existing system which is called hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The proposing system will make use of other combination of hybrid approach and will improve the accuracy of the prediction for timely improving the diagnosing of the heart disease and Ensemble learning technique is used.

RBF SVM and Logistic Regression algorithms are used and the predictions are stacked. The existing architecture contains the input layer followed by a combination of RF and LM methods with 14 attributes along with activation function, in the subsequent combination two techniques was performed with extended attributes with previous parameters, also applied the in all the layers for prediction probability calculations, added an output layer.

In our proposed architecture is shown in Figure 2. The cleaned data is split into Logistic regression and RBF SVM and predictions are done as p_1, p_2 and the meta classifier for enhanced accurate final prediction using various training data sets will be built. The same dataset is tested with different machine learning classifiers such as Logistic Regression (LR), NB, KNN, and SVM with different kernels, such as linear and RBF and simple neural networks.

In this paper, we proposed our combination technique to predict whether the patient have heart disease or not. The prediction accuracy is increased to diagnose the patient disease on time to reduce mortality rate.

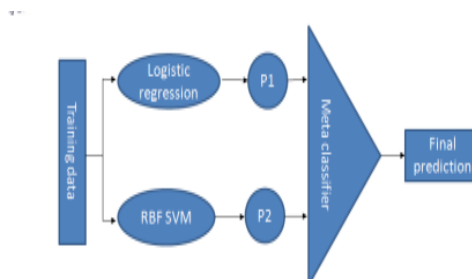


Fig 2: Proposed Architecture showing Stacking Classifiers

The proposed model accuracy is compared with the existing ML models which were depicted in Table 1. Naïve Bayes achieved 80.62% training accuracy and 77.04% testing accuracy. KNN achieved 79.76% training accuracy and 68.86% testing accuracy. SVM (Linear) achieved 90.61% training accuracy and 86.83% testing accuracy. Neural Network achieved 88.95% training accuracy and 86.97% testing accuracy. The proposed network achieved 92.8% training accuracy and 90% testing accuracy. The graphical representation of existing and proposed method accuracies is shown in given table 1.

Table 1. Comparison Methods

Model		Trainin g Accura cy	Testing Accuracy
Naive Bayes		80.62	77.04
KNN		79.76	68.86
SVM	Linear	90.61	85.29
Neural Network		88.95	86.97
Proposed Network		92.8	90

Benefits:

1. Decrease the amount of work required before receiving results.
2. There is still no proof about cheating.
3. It harness the capabilities of classification models

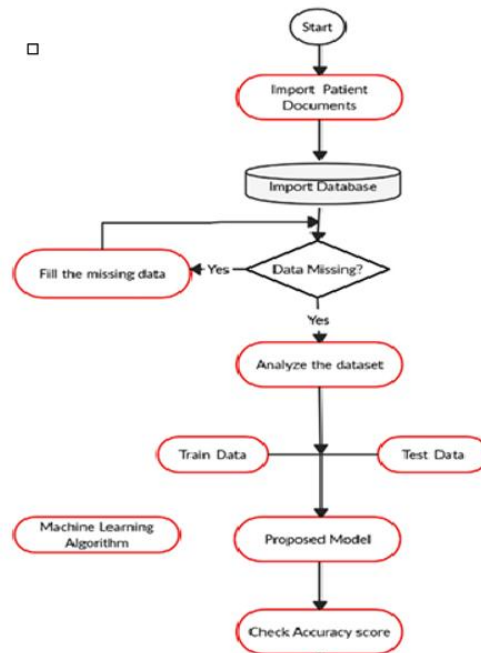


Fig 3: Flow Showing Proposed Architecture showing Stacking Classifiers

Stacking Classifier

The main purpose of this application is to look at the feature selection methods, data preparation, including sample preparation used within the training models. The application emerges as an important in the accuracy during training, verification, and practical validating. As little more than a result, this application had been carried out with the aim of learning little about the models and executing the techniques. The identification of cardiac disease is a severe difficulty. Generally, devices that would identify heart disease, because they're either too expensive or unproductive on calculating the cholesterol levels in human. This same risk of death that chronic implications of ventricular conditions can all be diminished even when they are found early. Unfortunately, this is

better to watch such individuals on a large scale. Because that need significant competence, time, and technique to successfully treat situations plus discuss with something like a physician for 24hrs, it really isn't possible. We sometimes use different classifiers so examine the data finding unstructured data considering they have had so much data in modern day society.

Using medicine data, the hidden patterns might be mistreated for care and emergency. An aim of this study seeks to see if a complete current parameter, also including personality, weight, excessive sweating, overnight blood pressure, and so on, predict that if they are common in patients of cardiac heart attacks. Another collection with both the doctor as well as parameters is retrieved from those in the Extracted features. One should estimate whereas if person could undergo a heart attack using the same knowledge. Regardless as well as not, they include a biostability. To provide it, researchers categories individual business owners on 14 medical indicators if user really at risk for heart disease. Different techniques are now used to train these medical personality traits: artificial neural network (Linear Method and Random Forest). Classifier is really the most accurate of these strategies, with either an accuracy of 90 basis points. Ultimately, we characterize persons of developing a disease If you've a biostability or not, this process is extremely low cost.

This suggested method of illness predictions utilizing ml algorithms is that we've had utilized a variety of methods, algorithms, as well as other tools to implement a situation which its predictions a health diagnosis based on their signs, which we compare to the software's datasets that has been before published. Through correlating those databases to the particular diagnosis, we can approximate the client's disorder %. The dataset and symptoms go to the prediction model of the system where the data is preprocessed for the future references and then the feature selection is done by the user, where they will enter the various symptoms. Then the data goes in the recommendation model, there it shows the risk analysis that is involved in the system and it also provides the probability estimation of the system such that it shows the various probability like how the system behaves when there are n number of predictions are done and it also does the recommendations for the patients from their final result and also from their symptoms like it can show what to use and what not to use from the given datasets and the final results. Here we have combined the overall structure and unstructured form of data for the overall risk analysis that is required for doing the prediction of the disease.

Using the structured analysis, we can identify the chronic types of disease in a particular region and particular community. In unstructured analysis we select the features automatically with the help of algorithms and techniques. This system takes symptoms from the user and predicts the disease accordingly based on the symptoms that it takes and also from the previous datasets, it also helps in continuous evaluation of viral diseases, heart rate, blood pressure, sugar level and much more which is in the system and along with other external symptoms it predicts the appropriate and accurate disease.

VI. METHODOLOGY

We'll start by exploring that collection with Panda's, numpy, matplotlib and seaborn packages of python for evaluating information, Training and experimentation on datasets. The heart Disease Prediction model will be trained on the dataset of diseases to do the prediction accurately and produce on our heart dataset with 14 classifiers. In this project different algorithms were used Logistic Regression, RFB SVM, Stacking classifiers.

We will represent that data samples utilizing bar, or bar plots again using our proposed techniques.

We'll choose certain characteristics from either the database besides research during filtering. Separating the dataset into two for testing and training and Utilizing machine learning methods to find as well as compare the performance, thereafter determining Accuracy, Recall, as well as Point total results. This information gets maintained in order to detect each user inputs. That visitor would determine their consequence by providing mistreated via an Interface built with the Python System.

Deployment and analysis on real life scenario the trained and tested prediction model will be deployed

Table 2: All Information's Used for Prediction of Heart Diseases

#	Column	Non-Null Count	Dtype
0	age	303 non-null	int64
1	sex	303 non-null	int64
2	cp	303 non-null	int64
3	trestbps	303 non-null	int64
4	chol	303 non-null	int64
5	fbs	303 non-null	int64
6	restecg	303 non-null	int64
7	thalach	303 non-null	int64
8	exang	303 non-null	int64
9	oldpeak	303 non-null	float64
10	slope	303 non-null	int64
11	ca	303 non-null	int64
12	thal	303 non-null	int64
13	target	303 non-null	int64

Table 3: Data Descriptions of used Column Sets

S.NO	Attribute	Description
1	Age	Age in years
2	Sex	Sex (1 = male; 0 = female)
3	Cp	Chest pain type (Categorized into 4 values)
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl; (1 = true; 0 = false)
7	Restecg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	The predicted attribute(Target)	Diagnosis of heart disease represented in 5 values (0 represents absence, 1 to 4 represents presence in different degree)

Data description

Preprocessing

At the principal level stage, the dataset is first cleaned and processed using preprocessing techniques using panda's package. The counterplot of sex and target attributes group is shown in Table 12. After that, using the data visualization procedure, the data frame attributes are shown in Figure 1.

Separate majority and minority classes

from sklearn.utils import resample

Separate majority and minority classes

df_majority = df[df['target']== 1]

df_minority = df[df['target']== 0]

df_minority_upsampled = resample(df_minority, replace=True,n_samples=500,random_state=123)

df_majority_downsampled = resample(df_majority, replace=True,n_samples=500,random_state=123)

Combine minority class with downsampled majority class

df_upsampled = pd.concat([df_minority_upsampled, df_majority_downsampled])

Display new class counts

df_upsampled['target'].value_counts()

Above algorithm separates majority and minority classes with sample of majority and sample minority classes using combination technique of minority and majority classes.

```
In [18]: # split data into features (X) and labels (y)
X = df.drop('target', axis=1)
y = df['target']
print(X.sample(1))
print(y.sample(1))

age      sex      cp      trestbps      chol      fbs      restecg      thalach      exang      oldpeak      \
168      0      1      0      158      234      0      0      147      0      1.4
200      0      0      0      120      280      0      2      130      1      2.0
160      0      1      0      120      230      0      0      120      1      2.0
68      0      1      0      120      230      0      1      170      0      0.0
200      0      0      0      120      255      0      1      105      1      0.0

slope      ca      thal
168      1      1      3
200      1      1      3
160      2      0      2
68      2      0      2
200      0      0      3
155      1
210      0
154      0
200      0
251      0
Name: target, dtype: int64
```

Figure 4: Showing Split data into features(x) and labels (y)

Feature Selection

The feature selection and modeling keep on repeating for various combinations of attributes. A very high and low risk patients will be classified on the basis of various tests which are done in the selected group, but proposed models are only based on clinical situations based which uses supervised learning methods for predictions.

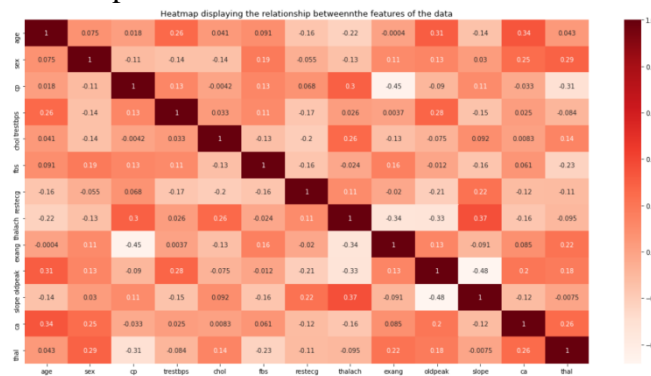



Figure 5: Heatmap displaying the relationship between the features of the data

Figure 5 showing Heatmap displaying the relationship between the features of the data Create correlation matrix through upper triangle and features correlation through drop features technique.

Create correlation matrix

```
import numpy as np
corr_matrix = X.corr().abs()
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))
to_drop = [column for column in upper.columns if any(upper[column] > 0.35)]
X.drop(to_drop, axis=1, inplace=True)
```

Table 4: Data Descriptions with loaded data values



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	oldpeak	ca	thal
281	52	1	0	128	204	1	1	156	1.0	0	0
168	63	1	0	130	254	0	0	147	1.4	1	3
235	51	1	0	140	299	0	1	173	1.6	0	3
249	69	1	2	140	254	0	0	146	2.0	3	3
258	62	0	0	150	244	0	1	154	1.4	0	2

Applying Machine Learning Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
import math
c = [10000, 1000, 100, 10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001]
train_auc = []
cv_auc = []
for i in c:
    clf = LogisticRegression(C=i)
    clf.fit(X_train, y_train)
    prob_cv = clf.predict(X_test)
    cv_auc.append(accuracy_score(y_test, prob_cv))
    prob_train = clf.predict(X_train)
    train_auc.append(accuracy_score(y_train, prob_train))
optimal_c = c[cv_auc.index(max(cv_auc))]
c = [math.log(x) for x in c]
#plotauc vs alpha
x = plt.subplot( )
x.plot(c, train_auc, label='AUC train')
x.plot(c, cv_auc, label='AUC CV')
plt.title('Accuracy vs hyperparameter')
plt.xlabel('c')
plt.ylabel('Accuracy')
x.legend()
plt.show()
print('optimal c for which auc is maximum : ', optimal_c)
```

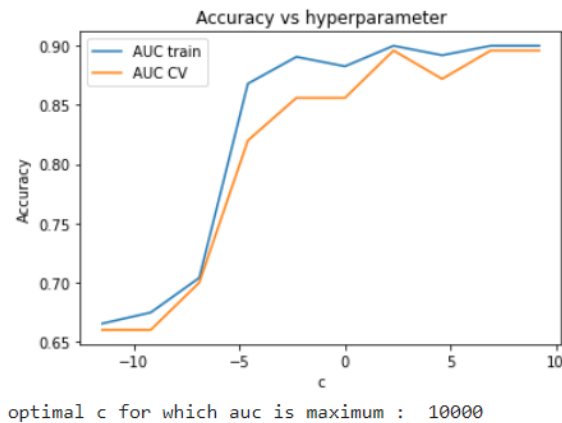


Figure 6: Accuracy vs Hyper parameter using Logistic regression

Testing AUC on Test data

```
log = LogisticRegression(C=optimal_c)
log.fit(X_train,y_train)
pred_test = log.predict(X_test)
#fpr1, tpr1, thresholds1 = metrics.roc_curve(y_test, pred_test)
pred_train = log.predict(X_train)
#fpr2,tpr2,thresholds2 = metrics.roc_curve(le_y_train,pred_train)
test = accuracy_score(y_test,pred_test)
train = accuracy_score(y_train,pred_train)
print("Accuracy on Test data is " +str(accuracy_score(y_test,pred_test)))
print("Accuracy on Train data is " +str(accuracy_score(y_train,pred_train)))
print("-----")
# Code for drawing seaborn heatmaps
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(y_test, pred_test.round()), index=class_names,
columns=class_names )
fig = plt.figure( )
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")
```

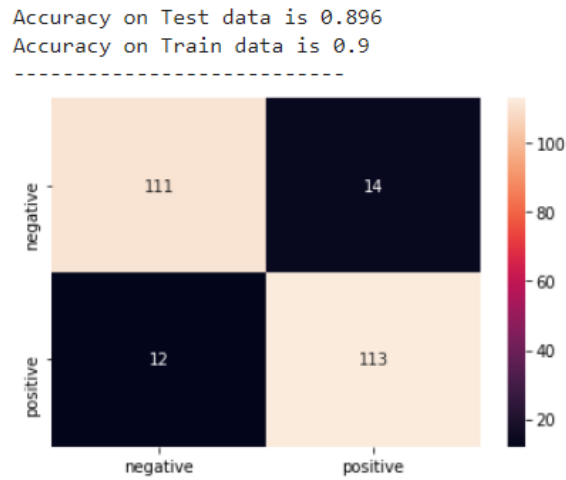


Figure 7: Chart showing Accuracy values on Test and Train data using Logistic regression

Data Frame Showing positive and Negative

```
original = ["Positive" if x==1 else "Negative" for x in y_test[:20]]
predicted = log.predict(X_test[:20])
pred = []
for i in predicted:
    if i == 1:
        k = "Positive"
    else:
        k = "Negative"
    pred.append(k)
# Creating a data frame
df = pd.DataFrame(list(zip(original, pred)),
                   columns=['original_Classlabel', 'predicted_classlebel'])
df
```

	original_Classlabel	predicted_classlebel
0	Negative	Negative
1	Negative	Positive
2	Negative	Negative
3	Negative	Negative
4	Negative	Negative
5	Negative	Negative
6	Positive	Positive
7	Negative	Negative
8	Positive	Positive
9	Negative	Negative
10	Negative	Positive
11	Positive	Positive

Table 5: Data Frame Showing positive and Negative

Applying RBF SVM

```
from sklearn.calibration import CalibratedClassifierCV
from sklearn.svm import SVC
C = [10000,1000,100,10,1,0.1,0.01,0.001,0.0001]
```

```

train_auc = []
cv_auc = []
for i in C:
    model = SVC(C=i)
    clf = CalibratedClassifierCV(model, cv=3)
    clf.fit(X_train,y_train)
    prob_cv = clf.predict(X_test)
    cv_auc.append(accuracy_score(y_test,prob_cv))
    prob_train = clf.predict(X_train)
    train_auc.append(accuracy_score(y_train,prob_train))
    optimal_C= C[cv_auc.index(max(cv_auc))]
C=[math.log(x) for x in C]
#plotauc vs alpha
x = plt.subplot( )
x.plot(C, train_auc, label='Accuracy train')
x.plot(C, cv_auc, label='Accuracy CV')
plt.title('Accuracy vs hyperparameter')
plt.xlabel('C')
plt.ylabel('Accuracy')
x.legend()
plt.show()
print('optimal C for which auc is maximum : ',optimal_C)

```

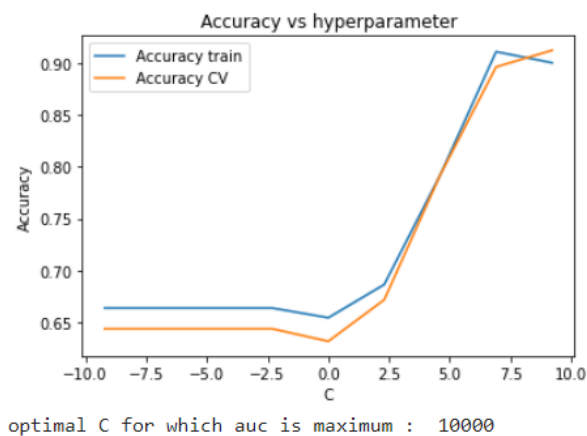


Figure 8: Accuracy vs Hyper parameter using RBF SVM

RBF SVM FOR ACCURACY

```

#Testing AUC on Test data
model =SVC(C = optimal_C)
clf = CalibratedClassifierCV(model, cv=3)
clf.fit(X_train,y_train)
import pickle
filename = 'heart_rbf.pkl'
pickle.dump(clf, open(filename, 'wb'))

```

```

pred_test = clf.predict(X_test)
#fpr1, tpr1, thresholds1 = metrics.roc_curve(y_test, pred_test)
pred_train = clf.predict(X_train)
#fpr2,tpr2,thresholds2 = metrics.roc_curve(le_y_train,pred_train)
test = accuracy_score(y_test,pred_test)
train = accuracy_score(y_train,pred_train)
print("Accuracy on Test data is " +str(accuracy_score(y_test,pred_test)))
print("Accuracy on Train data is " +str(accuracy_score(y_train,pred_train)))
print("-----")
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(y_test, pred_test.round()), index=class_names,
columns=class_names )
fig = plt.figure( )
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

```

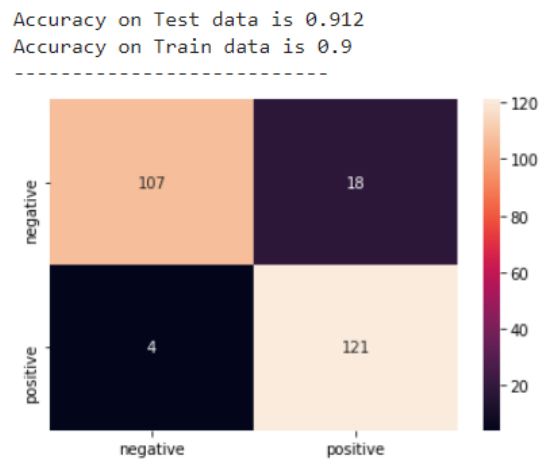


Figure 9: Chart showing Accuracy values on Test and Train data using RBF SVM

Data Frame Showing positive and Negative

```

original = ["Positive" if x==1 else "Negative" for x in y_test[:20]]
predicted = clf.predict(X_test[:20])
pred = []
for i in predicted:
    if i == 1:
        k = "Positive"
    pred.append(k)
    else:
        k = "Negative"
    pred.append(k)
# Creating a data frame
df = pd.DataFrame(list(zip(original, pred,)),
columns=['original_Classlabel', 'predicted_classlebel']) df

```


Table 6: Data Frame Showing positive and Negative

	original_Classlabel	predicted_classlabel
0	Negative	Negative
1	Negative	Positive
2	Negative	Negative
3	Negative	Negative
4	Negative	Negative
5	Negative	Negative
6	Positive	Positive
7	Negative	Negative
8	Positive	Positive
9	Negative	Negative
10	Negative	Positive
11	Positive	Positive
12	Negative	Positive
13	Negative	Positive

Stacking Classifier

```

from mlxtend.classifier import StackingClassifier
log = LogisticRegression(C=optimal_c) # initialising KNeighbors Classifier
model = SVC(C = optimal_C)
rbf = CalibratedClassifierCV(model, cv=3)
clf_stack = StackingClassifier(classifiers=[log, rbf], meta_classifier = rbf, use_probabilities = True,
use_features_in_secondary = True)
#training our model for max_depth=50,min_samples_split=500
clf_stack.fit(X_train,y_train)
print("Accuracy on Test data is " +str(accuracy_score(y_test,pred_test)))
# Code for drawing seaborn heatmaps
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(y_test, pred_test.round()), index=class_names,
columns=class_names )
fig = plt.figure( )
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

```

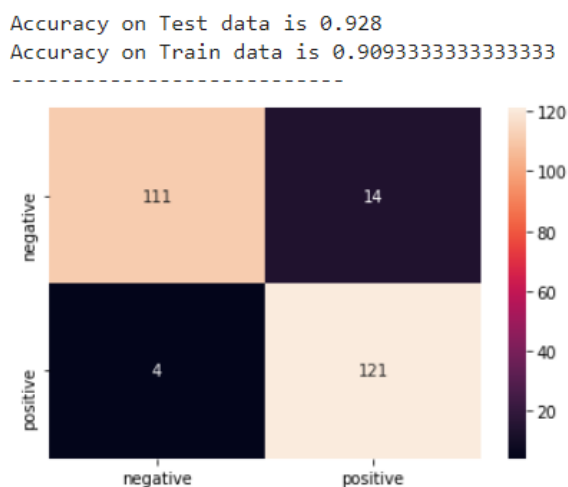


Figure 10: Chart showing Accuracy values on Test and Train data using Stacking classifier

```
original = ["Positive" if x==1 else "Negative" for x in y_test[:20]]
```

```

predicted = clf_stack.predict(X_test[:20])
pred = []

for i in predicted:
    if i == 1:
        k = "Positive"
    pred.append(k)
    else:
        k = "Negative"
    pred.append(k)
# Creating a data frame
df = pd.DataFrame(list(zip(original, pred,)),
                    columns=['original_Classlabel', 'predicted_classlebel'])
df

```

Table 7: Data Frame Showing positive and Negative

	original_Classlabel	predicted_classlebel
0	Negative	Negative
1	Negative	Positive
2	Negative	Negative
3	Negative	Negative
4	Negative	Negative
5	Negative	Negative
6	Positive	Positive
7	Negative	Negative
8	Positive	Positive
9	Negative	Negative
10	Negative	Positive
11	Positive	Positive
12	Negative	Positive
13	Negative	Positive

Table 8: Performance Table Showing Train and Test Accuracy Values

	model	Train-Accuracy	Test-Accuracy
0	Logistic regression	0.900000	0.896
1	RBF SVM	0.900000	0.912
2	Stacking	0.909333	0.928

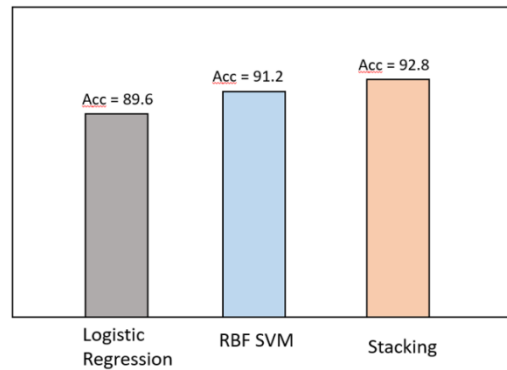


Figure 11: Chart 1 Results Analysis showing the Accuracy Values

VI. CONCLUSION

As heart disease prediction is complex, it should be predicted and diagnose on time to reduce the mortality rate. And the predictions should be more accurate so that the health care industry can start the treatment on time. Many of the state-of-the-art algorithms are not suitable to predict CVD disease correctly. Even doctors are also unable to predict the disease accurately. So, the proposed system supports stacking classifier the prediction accuracy is increased compared to existing system. In this script, proposed system uses other combination of hybrid approach by combining RBF SVM along with Logistic regression. RBF SVM uses kernel function to solve non-linear problems and Logistic regression provides great training efficiency for timely improving the diagnosis of the heart disease. And also, the paper gave a comparison between proposed work and state-of-the-art algorithms. In future we will propose methodology for early prediction of heart disease with high accuracy and minimum cost and complexity.

REFERENCE

- [1] M. Durairaj and V. Revathi, "Prediction of heart disease using back propagation MLP algorithm," *Int. J. Sci. Technol. Res.*, vol. 4, no. 8, pp. 235239, 2015.
- [2] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," in *Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE)*, Feb. 2015, pp. 520525.
- [3] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Mar. 2018, pp. 12751278.
- [4] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [5] Mangesh Limbitote , Dnyaneshwari Mahajan , Kedar Damkondwar , Pushkar Patil, 2020, A Survey on Prediction Techniques of Heart Disease using Machine Learning, *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)* Volume 09, Issue 06 (June 2020),
- [6] A. S. Abdullah and R. R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in *Proc. Int. Conf. Recent Trends Comput. Methods*,

Commun. Controls, Apr. 2012, pp. 22–25.

[7] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, “Using PSO algorithm for producing best rules in diagnosis of heart disease,” in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306–311.

[8] N. Al-milli, “Back propagation neural network for prediction of heart disease,” J. Theor. Appl. Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[9] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, “Analysis of neural networks based heart disease prediction system,” in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[10] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019

[11] S.P.Bingulac, On the Compatibility of Adaptive Controllers, Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994. (Conference proceedings)

[12] SonamNikhar, A.M. Karandikar Prediction of Heart Disease Using Machine Learning Algorithms International Journal of Advanced Engineering, Management and Science (IAEMS) Infogain Publication,[Vol-2, Issue-6, June- 2016].I.S. Jacobs and C.P. Bean, Fine particles, thin films and exchange anisotropy, in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[13] AditiGavhane, GouthamiKokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), Prediction of Heart Disease Using Machine Learning, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1- 5386-0965-1

[14] Abhay Kishore¹, Ajay Kumar², Karan Singp, Maninder Punia⁴, Yogita Hambir⁵, Heart Attack Prediction Using Deep Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04 | Apr-2018.

[15] A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar, Machine Learning Techniques For Heart Disease Prediction, International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019.

[17] Mr.SanthanaKrishnan.J, Dr.Geetha.S, Prediction of Heart Disease Using Machine Learning Algorithms,2019 1st International Conference on Innovations in Information and Communication Technology(ICIICT),doi:10.1109/ICIICT1.2019.8741465.