# Heart Disease Pridiction Using Machine Learning and Data Analytics Approach

N Pravallika[1], B. Murali[2]
[1]*PG Student, CSE, Quba College of Engineering & Technology*
[2]*Associate professor, CSE, Quba College of Engineering & Technology*

***Abstract*— In this proposed system we implement Heart Disease Prediction using Artificial Neural Network (ANN). The project "Heart Disease Prediction using Artificial Neural Network (ANN)" aims to develop a predictive model for the early detection of heart disease using ANN's. ANN's are powerful machine learning algorithms that can learn patterns and relationships in data, making them an ideal choice for predicting complex medical conditions like heart disease. The proposed system is implemented using Cleveland Heart Disease data set available on UCI machine learning repository / Kaggle.**

***Index Terms*— Disease Prediction, artificial Neural Network, Machine learning, heart diseases.**

## I. INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world [1]. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen [2]. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time [3]. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available [4]. The effective diagnosis The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Rathinam. and proper treatment can save the lives of many people [5]. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually [6]. Most of the people in the United States are suffering from heart disease [7]. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze [8]. Thus, to develop a noninvasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a 107562 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/ VOLUME 8, 2020 J. P. Li et al.: HD Identification Method Using ML Classification in E-Healthcare result, the ratio of death decreases [9] and [10].

The Cleveland heart disease data set was used by various researchers [11] and [12] for the identification problem of HD. The machine learning predictive models need proper data for training and testing. The performance of machine learning model can be increased if balanced data set is use for training and testing of the model. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. In literature various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD.

## II .LITERATURE SURVEY

### 2.1 INTRODUCTION

In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnosis HD. This research study present some

existing machine learning based diagnosis techniques in order to explain the important of the proposed work. Detrano et al. [11] developed HD classification system by using machine learning classification techniques and the performance of the system was 77% in terms of accuracy. Cleveland data set was utilized with the method of global evolutionary and with features selection method. In another study Gudadhe et al. [22] developed a diagnosis system using multi-layer Perceptron and support vector machine (SVM) algorithms for HD classification and achieved accuracy 80.41%. Humar et al. [23] designed HD classification system by utilizing a neural network with the integration of Fuzzy logic. The classification system achieved 87.4% accuracy. Resul et al. [19] developed an ANN ensemble based diagnosis system for HD along with statistical measuring system enterprise miner (5.2) and obtained the accuracy of 89.01%, sensitivity 80.09%, and specificity 95.91%. Akil et al. [24] designed a ML based HD diagnosis system. ANN-DBP algorithm along with FS algorithm and performance was good. Palaniappan et al. [17] proposed an expert medical diagnosis system for HD identification. In development of the system the predictive model of machine learning, such as navies bays (NB), Decision Tree (DT), and Artificial Neural Network were used. The 86.12% accuracy was achieved by NB, ANN accuracy 88.12% and DT classifier achieved 80.4% accuracy. Olaniyi et al. [18] developed a three-phase technique based on the artificial neural network technique for HD prediction in angina and achieved 88.89% accuracy. Samuel et al. [20] developed an integrated medical decision support system based on artificial neural network and Fuzzy AHP for diagnosis of HD. The performance of the proposed method in terms of accuracy was 91.10%. Liu et al. [25] proposed a HD classification system using relief and rough set techniques. The proposed method achieved 92.32% classification accuracy.

## III. SYSTEM ANALYSIS

### 3.1 EXISTING SYSTEM:

- The existing system demonstrates the prediction of heart disease using multiple machine learning classification algorithms such as Naive Bayes, Random Forest, SVM etc., and compares their accuracy scores. The existing system for predicting heart disease uses Naive Bayes, Random Forest, and Support Vector Machines (SVM). These algorithms are commonly used in machine learning for classification tasks.
- The system collects patient data, such as age, gender, blood pressure, and cholesterol levels, and applies these algorithms to predict the likelihood of heart disease.
- The existing system developed a heart disease algorithm based on Ensemble Learning Technique, known as Stacking Classifier, where different classification models will be developed and trained with the training set's help to compare their performance by using their accuracy scores.

DISADVANTAGES OF EXISTING SYSTEM:

- Over-fitting: Ensemble learning techniques can sometimes overfit the training data, resulting in high accuracy on the training data but poor performance on new, unseen data.
- This can lead to incorrect predictions and compromised patient outcomes.
- Limited Interpretability: Ensemble learning techniques can be complex and difficult to interpret, making it challenging to understand the underlying factors that contribute to heart disease prediction.

3.2.1.ADVANTAGES:

- ANN's can learn from large datasets and can handle noisy and complex data, making them an ideal choice for medical diagnosis.
- In the proposed system, developing an ANN model that can learn from the preprocessed data and accurately predict the likelihood of heart disease in patients

### 3.3.SYSTEM REQUIREMENTS

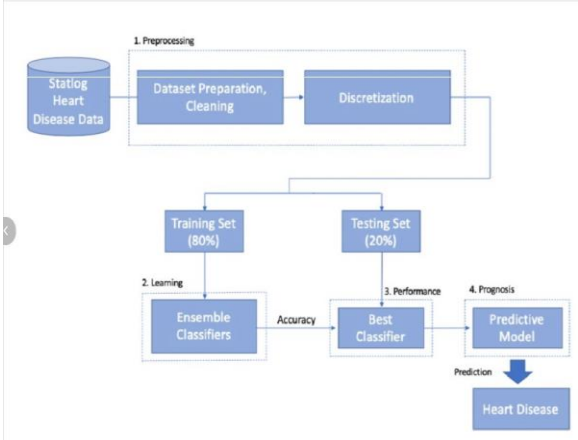
3.3.1.HARDWARE REQUIREMENTS(minimum):

- System : Pentium IV 2.4 GHz
- Hard Disk : 1 TB
- Ram : 2 GB.

3.3.2. SOFTWARE REQUIREMENTS:

| | | |
|---|---|---|
| Operating system | : | Windows XP/7. |
| Coding Language | : | Python |
| Server | : | tomcat |
| Database | : | MYSQL |

## IV. SYSTEM ARCHITECTURE

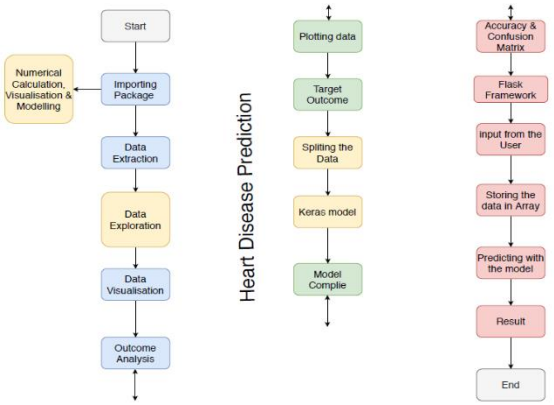



## V. SYSTEM DESIGN:DATA FLOW DIAGRAM

### 5.1 DATA FLOW DIAGRAM:

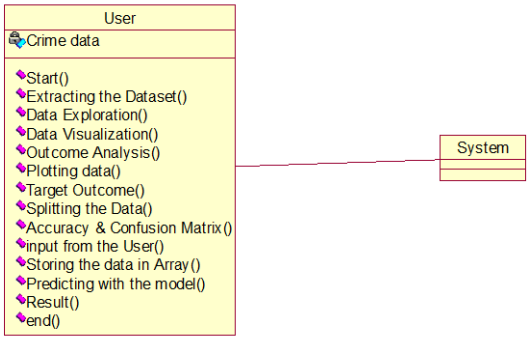
A graphical tool used to describe and analyze the moment of data through a system manual or automated including the process, stores of data, and delays in the system. Data Flow Diagrams are the central tool and the basis from which other components are developed. The transformation of data from input to output, through processes, may be described logically and independently of the physical components associated with the system.

DFD's are the model of the proposed system. They clearly should show the requirements on which the new system should be built. Later during design activity this is taken as the basis for drawing the system's structure charts. The Basic Notation used to create a DFD's are as follows:




### 5.2 CLASS DIAGRAM:

Class diagrams area unit the foremost common diagrams employed in UML. Category diagram consists of categories, interfaces, associations and collaboration. Category diagrams primarily represent the thing directed read of a system that is static in nature. Active category is employed in a very category diagram to represent the concurrency of the system.




## VI. SOFTWARE ENVIRONMENT

Python Technology: What is Python: -

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following

Machine Learning

- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like OpenCV, Pillow)
- Web scraping (like Scrapy, Beautiful Soup, Selenium)
- Test frameworks
- Multimedia

Advantages of Python: -

Let us see how Python dominates over other languages.

1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2. Extensible

As we have seen earlier, Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add scripting capabilities to our code in the other language.


## VII. SYSTEM IMPLEMENTATION

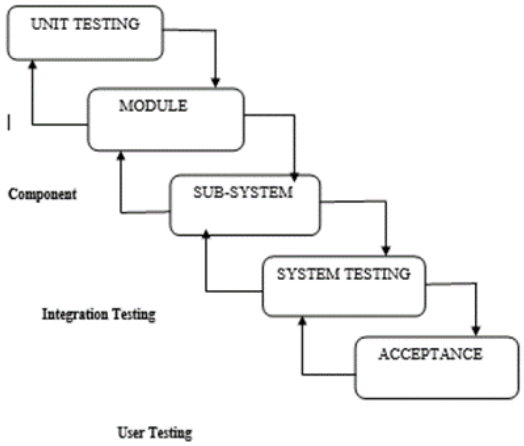
Sample code:

```
from tkinter import messagebox
from tkinter import *
from tkinter import simple dialog
import tkinter
from tkinter import file dialog
from tkinter.file dialog import askopenfilename
from sklearn.linear_model import Logistic
Regression
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import confusion_matrix
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import make_scorer,
accuracy_score,precision_score,recall_score
from sklearn.model_selection import GridSearchCV
from sklearn.neural_network import MLPClassifier
from sklearn.svm import LinearSVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import
RandomForestClassifier, StackingClassifier
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import Random Forest
Classifier, Voting Classifier
```

## VIII SYSTEM TESTING

SYSTEM TESTING

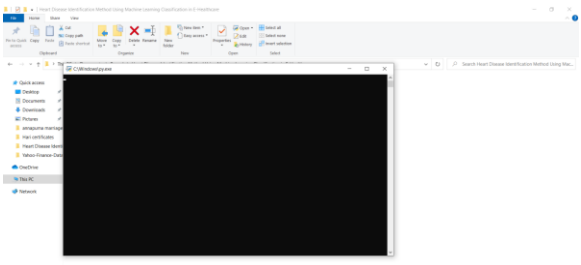
The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies and/or a finished product it is the process of exercising software with the intent of ensuring.

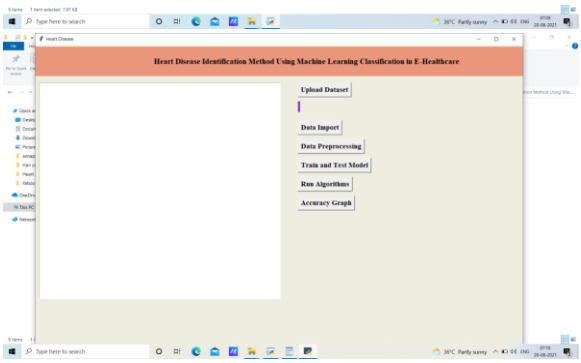


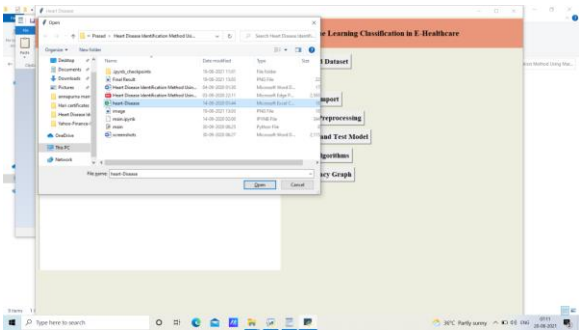

Testing Methodologies:

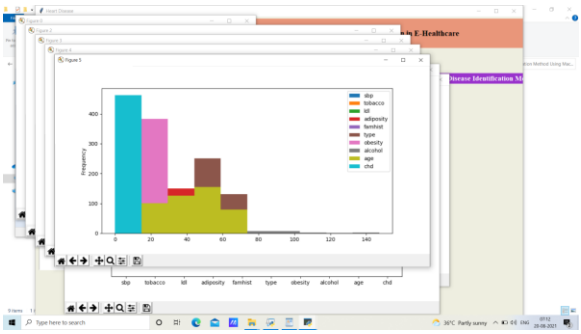
Testing is the process of finding differences between the expected behavior specified by system models and the observed behavior implemented system. From modeling point of view, testing is the attempt of falsification of the system with respect to the system models. The goal of testing is to design tests that exercise defects in the system and to reveal problems.

## IX SCREENSHOTS







## X.CONCLUSION

In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include LR, KNN, ANN, SVM, NB, and DT are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyper parameters selection. The system is tested on Cleveland heart disease data set. Furthermore, performance evaluation metrics are used to check the performance of the identification system. According to Table 15 the specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. Therefore for ANN with relief is the best predictive system for detection of healthy people. The sensitivity of classifier NB on selected features set by LASSO FS algorithm also gives the best result as compared to the sensitivity values of Relief FS algorithm with classifier SVM (linear). The classifier Logistic Regression MCC is 91% on selected features selected by FCMIM FS algorithm.

## REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ‘‘Epidemiology and risk profile of heart failure,’’ Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[2] M. Durairaj and N. Ramasamy, ‘‘A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,’’ Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ‘‘Decision making in advanced heart failure: A scientific statement from the American heart association,’’ Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.