Scientific Journal of Impact Factor (SJIF): 5.71
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 8, Issue 04, April-2021

Efficient heart disease classification approach using data mining techniques

Hetasvi R. Ribadiya[1], Prof. Swati Sharma[2]

[1]M.E. Scholar, Computer Engineering, Darshan Institute of Engineering & Technology, Rajkot
[2]Assistant Professor, Computer Engineering, Darshan Institute of Engineering & Technology, Rajkot

ABSTRACT - *In this inquire about, we point to foresee exactness, whether the person is at chance of a heart illness. This forecast will be done by applying machine learning calculations on preparing information that we offer. Once the individual enters the data that is requested, the calculation is connected and the result is produced. Clearly, the exactness is anticipated to diminish when the medical information itself are deficient. We execute the expectation show over real-life healing center information. We have proposed assist by applying SHGB(STOCHASTIC HIGH GRADIENT BOOSTING) ALGORITHM data mining or machine learning calculations over the preparing information to foresee the hazard of maladies, comparing their accuracies so that ready to conclude the foremost exact one. Properties can too be adjusted in an endeavor to make strides the precision assist in terms of accuracy.*

Keywords: *accuracy, performance, Recall, Precision, KNN,* Naïve Bayes, *SHGB*

I. INTRODUCTION

Data mining is the process is to extract information from a data set and transform it into an understandable structure. In the health care industry, data mining plays an important role in predicting diseases. For detecting a disease number of tests should be required from the patient. But using the data mining technique the number of a test should be reduced. This reduced test plays an important role in time and performance.

Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease. In day-to-day life, many factors affect the human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. The diagnosis of heart diseases is very important and is itself the most complicated task in the medical field.

Heart disease (HD) is one of the most common diseases nowadays, due to a number of contributing factors, such as high blood pressure, diabetes, cholesterol fluctuation, exhaustion, and many others.

My Research focuses on try to detect possibility of the heart diseases at early stage.

With the help of data mining techniques, doctors will be able to make future predictions. The data mining techniques and prediction models are responsible for making accurate predictions if a patient is likely to get a heart disease in the future.

If we can identify patients who are vulnerable to a heart disease in the future, then the doctor can take appropriate action to help the patient.

Major challenge is how to extract the information from these data because the amount is very large so some data mining and machine learning techniques can be used. But the main problem of data mining is, it uses different algorithms for detection of heart disease. Some algorithms are less accurate and time consuming.

Also, the expected outcome and scope of this project is that if disease can be predicted than early treatment can be given to the patients which can reduce the risk of life and save life of patients.

II. LITERATURE SURVEY

According [1], they propose to use convolutional neural network algorithm as a disease risk prediction algorithm using structured and perhaps even on unstructured patient data. The accuracy obtained using the developed model ranges between 85 and 88%. With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis. Various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, rather than the previously selected

Characteristics. To solve these problems, the structured and unstructured data can be combined in healthcare to assess the risk of disease. They propose to extend our algorithm to incorporate unstructured data as well. As of now, all attributes and laboratory tests considered have been approved by medical doctors.

According to [2], Heart disease and stroke have had an impact on 28:1% of total deaths in India in 2016 as compared to 15:2% in 1990. With the rising use of learning algorithms. In this edition paper, we have developed a system for predicting heart disease that can predict heart disease by using a modified random forest algorithm. The proposed algorithm is trained

with a dataset consisting of 303 instances which help to predict the occurrence of heart disease with an accuracy of 86:84% and can be implemented in the medical field to improve the overall diagnosis about heart disease. Random Forest is another learning algorithm that also applies to the nonlinear tendency of the data set as well as provides a better outcome compared to the decision tree algorithm. Random Forest is made up of large quantities of trees along with deliberately random inputs. Proper adjustments should be needed to get better results in the random forest so that by changing parameters such as randomness, number of trees, and the maximum depth, the accuracy could be increased. The proposed algorithm equally responds better in real-time and its accuracy can be increased by collecting more data and by implementing other deep learning-based techniques and convolutional neural network. This machine-based prediction based methodology will help to reduce human errors while detecting heart disease.

According to [3], Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. Currently, the recent development in medical supportive technologies based on data mining, machine learning plays an important role in predicting cardiovascular diseases. In this paper, we propose a new hybrid approach to predict cardiovascular disease using different machine learning techniques such as Logistic Regression (LR), Adaptive Boosting (AdaBoostM1), Multi-Objective Evolutionary Fuzzy Classifier (MOEFC), Fuzzy Unordered Rule Induction (FURIA), Genetic Fuzzy System-LogitBoost (GFS-LB) and Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML). For this purpose, the accuracy and results of each classifier have been compared, with the best classifier chosen for a more accurate cardiovascular prediction. The dataset used in this article is taken from the UCI Repository of Machine Learning Databases. The dataset is a collection of medical analytical reports with a total of 303 records with 14 medical features. The performance evaluation of these algorithms is done based on Accuracy, Sensitivity, Specificity and Error rate using WEKA and KEEL tools.

According to [4], Heart or We can say cardiovascular disease (CVD) is Couse of several illnesses, disability, and death. Working on heart disease patients' databases can be compared to real-life applications. Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. It also provides healthcare professionals with an extra source of knowledge for making decisions. With the motivation from literature that most of the researchers have considered neural network technique for building appropriates prediction approaches hence a hybrid technique has introduced in this paper by using the group of two most popular classification techniques of data mining, Naive Bayes and Neural Network. Naive Bayes algorithm employs a simplified version of Bayes formula to decide which class a feature belongs to. To improve the prediction accuracy of heart disease the intended approach integrating the prediction mechanism of two popular classification techniques namely Naïve Bayes and neural network scheme into a single form. The designed technique works in a layered format at which each layer has predicted entity separately with the use of optimized feature set that has been selected by the integrated attributes selection process. For the evaluation purpose, a number of datasets have been collected from the online available data pool, UCI data repository. Each and every evaluation result has demonstrated the effectiveness and efficiency of the proposed data prediction practice of this investigation.

According to [5], this shows that the main objective of data mining is to collect and choose the relevant data from the previously stored data. There are enormous types of data sets being used in data mining. Heart diseases are one of the major causes of death nowadays. Smoking, consumption of alcohol in large quantities, cholesterol, and pulse rate are the reason for heart diseases. The heart is the operating system of human body, if it will not function properly then it will directly affect the functioning of the other body parts. Smoking is one of the major causes of heart diseases; almost 40% of the population is dying because of this. The heart disease prediction is the problem which is solved with the machine learning techniques. The various machine learning techniques are reviewed for heart disease prediction. In the future, the novel approach of machine learning will be proposed which gives high accuracy for heart disease prediction.

According to [6], Heart disease is accountable for deaths in all age groups and is common among males and females. A good solution to this problem is to be able to predict what a patient's health status will be like in the future so the doctors can start treatment much sooner which will yield better results. It's a lot better than acting at the last minute where the patient is already at risk and hence the prediction of heart disease is widely researched area. A lot of research and technological advancement has been recorded in similar fields. This paper aims to report about taking advantage of the various data mining techniques and develop prediction models for heart disease survivability. In this paper, different classifiers and impact of data processing techniques are studied and experiments are conducted to find the best classifier for predicting the patient of heart disease and understanding the significance of how data processing can be used to increase accuracy. From our observations we understand that Logistic Regression and Naïve Bayes have a high accuracy when run on a high dimensional dataset and algorithms such as Decision Tree and Random Forest give better results on small dimensional data set. Random Forest gives better accuracy than Decision Tree Classifier as the algorithm is an optimized learning algorithm.

According to [7], In this work, the heart disease is effectively predicted by a Machine Learning (ML) model, which is trained with the UCI datasets. Univar ate and Multivariate analysis of the dataset is performed using statistical methods and checked for data - imbalanced, skewness/kurtosis in the distribution of data, and the correlation between the features. Performance is evaluated using metrics such as confusion matrix, Accuracy score, precision-recall curve (PRC), and Receiver operating curve

(ROC). An effective Machine Learning Model is developed, where the Dataset transformation using the Power transform technique essential, after cleaning the data and removing the outlier from the data set is incorporated to provide a good score. Dimension reduction technique - Kernel PCA used to reduce the features as low as 5 features from 13 features to improve model effectiveness not only accuracy, but also computational performance.

According to [8], heart disease has become more common these days. The life of people is at a risk. Variation in Blood pressure, sugar, pulse rate etc. can lead to cardiovascular diseases that include narrowed or blocked blood vessels. It may causes Heart failure, Aneurysm, Peripheral artery disease, Heart attack, Stroke and even sudden cardiac arrest. Many forms of heart disease can be detected or diagnosed with different medical tests by considering family medical history and other factors. But, the prediction of heart diseases without doing any medical tests is quite difficult. The aim of this project is to diagnose different heart diseases and to make all possible precautions to prevent at early stage itself with affordable rate. We follow 'Data mining' technique in which attributes are fed in to SVM, Random forest, KNN, and ANN classification Algorithms for the prediction of heart diseases. The preliminary readings and studies obtained from this technique is used to know the possibility of detecting heart diseases at early stage and can be completely cured by proper diagnosis. To diagnose the disease at early stage at affordable cost is the important aim of this paper total 20 attributes of nearly 2200 and above patients were collected. This collected data were then sorted and arranged systematically in Excel format. Using this data, it can be subjected to different data mining algorithms. From the medical profiles twenty attributes are extracted such as age, sex, blood pressure and blood sugar etc. to predict the likelihood of patient getting heart diseases. These attributes are fed in to SVM, Random forest, KNN, and ANN classification Algorithms in which ANN gave the best result with the highest accuracy. Valid performance is achieved using ANN algorithm in diagnosing heart diseases and can be further improved by increasing the number of attributes.

III. DATASET DESRIPTION

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "Target" field refers to the presence of heart disease in the patient. (0=not presence and 1=presence)

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

age	sex	Cp	trestbps	…	ca	thal	target
63	1	3	145	…	0	1	1
37	1	2	130	…	0	2	1
41	0	1	130	…	0	2	1
56	1	1	120	…	0	2	1
. .	. .	. .		. .	. .	. .	. .
45	1	3	110	…	0	3	0
68	1	0	144	…	2	3	0
57	1	0	130	…	1	3	0
57	0	1	130	…	1	2	0

[303 rows × 14 columns**]**

***Table 1. Heart Disease dataset containing age*, sex, cp, trestbps, chol, fbs, restecg, thalach, Exang, oldpeak, slope, ca, thal, target**

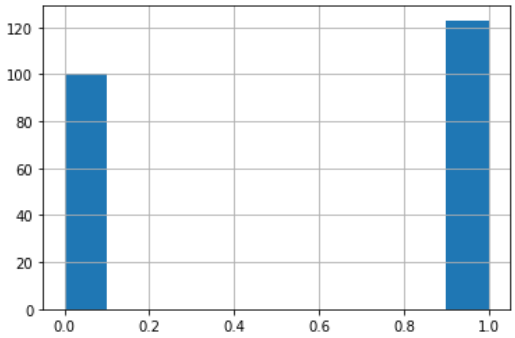

Figure 1. Class Distributions for training dataset (0: not present, 1: present) of original dataset

IV. VARIOUS CLASSIFICATION MODELS AND OVERSAMPLING TECHNIQUE

4.1 Naïve Bayes Classification Algorithm

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. Naïve Bayes Classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.
Bayes theorem on Mathematical Representation:

$$P\ (A\backslash B) = \frac{P(B\backslash A) * P(A)}{P(B)}$$

Here,
P (A) => independent probability of A (prior probability)
P (B) => independent probability of B
P (B\A) => conditional probability of B given A (likelihood)
P (A\B) => conditional probability of A given B (posterior probability).
Naïve Bayes is a simple and powerful algorithm for predictive modeling. This model is the most effective and efficient classification algorithm which can handle massive, complicated, non-linear, dependent data. Naïve comprises two part namely naïve & Bayes where naïve classifier assumes that the presence of the particular feature in a class is unrelated to the presence of any other feature.

4.2 KNN

KNN is a non-parametric method Used for classification. And also used regression. It is a type of instance-based learning. Where the function is only approximated locally. All computation is deferred vary till classification. This algorithm is among the simplest of all machine learning algorithms. Useful technique can be used to assign weight to the contributions of the neighbours. So that the nearer neighbours contribute more to the average than the more distant ones.

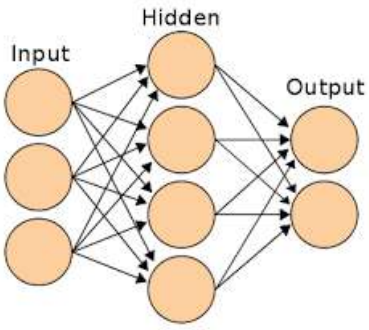

Figure 2. KNN

4.3 SHGB(STOCHASTIC HIGH GRADIENT BOOSTING)

A big insight into bagging ensembles and random forest was allowing trees to be greedily created from subsamples of the training dataset. This same benefit can be used to reduce the correlation between the trees in the sequence in gradient boosting models.

This variation of boosting is called stochastic gradient boosting. At each iteration a subsample of the training data is drawn at random (without replacement) from the full training dataset. The randomly selected subsample is then used, instead of the full sample, to fit the base learner.
A few variants of stochastic boosting that can be used:
Subsample rows before creating each tree.
Subsample columns before creating each tree
Subsample columns before considering each split. Using column sub-sampling prevents over-fitting even more so than the traditional row sub-sampling.
In boosting, the individual models are not built on completely random subsets of data and features but sequentially by putting more weight on instances with wrong predictions and high errors. The general idea behind this is that instances, which are hard to predict correctly ("difficult" cases) will be focused on during learning, so that the model learns from past mistakes. When we train each ensemble on a subset of the training set, we also call this Stochastic Gradient Boosting, which can help improve generalizability of our model.

Synthetic Minority Oversampling Technique
SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbour's for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

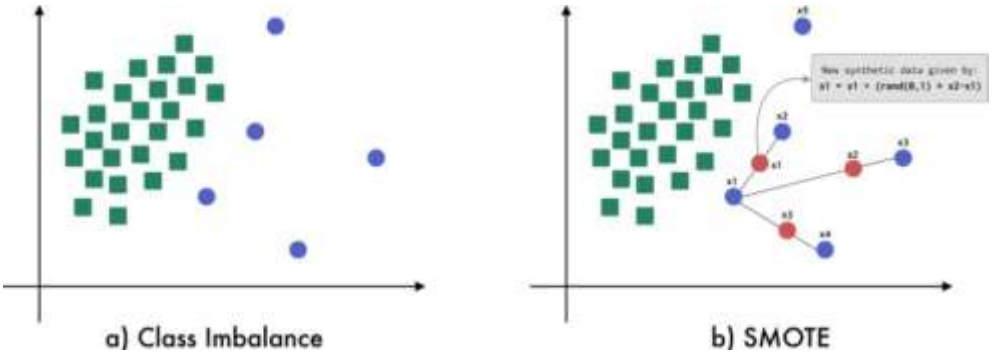

Figure 3. SMOTE

Figure 4 demonstrate the applied SMOTE technique after target = 0 and target =1 are of equal ratio.

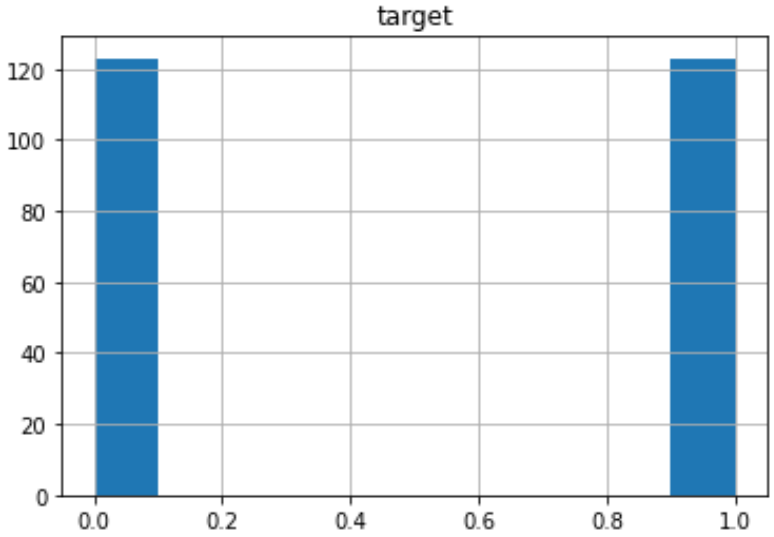

Figure 4. Equally Distributed Classes

V. PROPOSED APPROACH

5.1 Proposed Flowchart

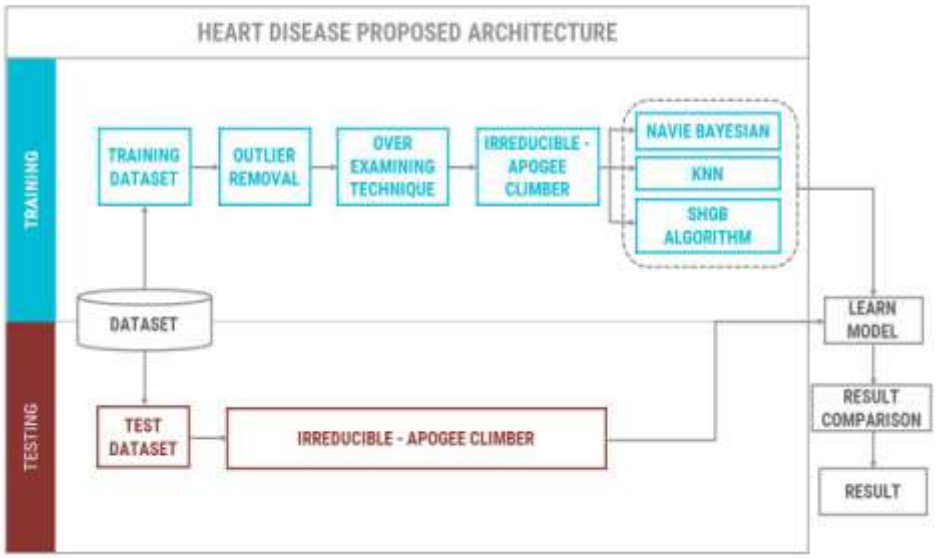

Figure 5. Flowchart for Proposed method

5.2 Proposed Algorithm

BEGIN
Step 1: Take input from a database
Step 2: Data-preprocessing from Dataset.
Step 3: Divide Training and testing data from Dataset.
Step 4: Outlier removal.
Step 5: Use SMOT over examining technique on Dataset
Step 6: Apply Irreducible- Apogee Climber on Dataset.
Step 7: Train Model using NB, KNN, SHGB algorithm
Step 8: Learn Model.
Step 9: Result comparison
Step 10: Result.
End

5.3 Description of Proposed Approach

- It starts with the data collection; here in this step, the collected input data is in the form of csv files.
- A process to gather context to the input data. Understanding the data for preprocessing and cleaning of datasets
- Dataset then divided into training dataset and test dataset among them 80% of the data will be used for training the model while rest 20% will be used for testing the model, which will be highly skewed or imbalanced.
- Now utilize class imbalance solver technique SMOTE on dataset, which is used to balance class distribution by randomly increasing minority class examples by replicating them.
- Then apply Irreducible- Apogee Climber on dataset. It is normalize the data for every feature the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1 and every other value gets transformed into decimal between 0 and 1.
- After the data segregation, the data are fed into machine learning algorithm like Naïve Bayes, KNN and Stochastic Gradient Boosting Classifier. This step is mainly done using training data to teach the machine to increase its predictive accuracy.
- Once the data have learnt enough, our learned model will be ready for testing.
- The learned model is tested using test data to check its predictive accuracy. If the predictive accuracy is up to the desired level, then the model is deployed.

VI. PERFORMANCE ANALYSIS

In this section, we have discussed the following parameters of the performance analysis of the proposed algorithm.
– TP (true positive): is a test result that observes the state when the state is present.
– TN (true negative): is a test result that does not observe the state when the state is absent.
– FP (false positive): is a test result that observes the state when the state is absent.
– FN (false negative): is a test result that does not observe the state when the state is present.
The above-mentioned parameters TP, TN, FP, and FN are used to calculate the AccuracyScore, PrecisionScore, F1Score, and RecallScore of the proposed algorithm.
Results obtained for performance analysis are shown in **Fig-6, Fig-7, Fig-8 and Fig-9**.

$$AccuracyScore = \frac{TP+TN}{TP+TN+FP+FN}$$

$$PrecisionScore = \frac{TP}{TP+FP}$$

$$RecallScore = \frac{TP}{TP+FN}$$

$$F1Score = \frac{2*PrecisionScore\ *RecallScore}{PrecisionScore\ +RecallScore}$$

Confusion Matrix of Heart Disease Dataset

	Predicted Disease	**Predicted Non-Disease**
Actual Disease	TP	FP
Actual Non-Disease	FN	TN

VII. RESULT & DISCUSSION

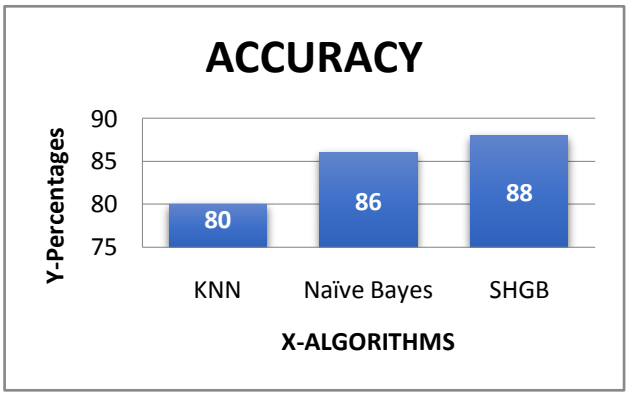

Figure 6 Comparison of Algorithms Based on Accuracy

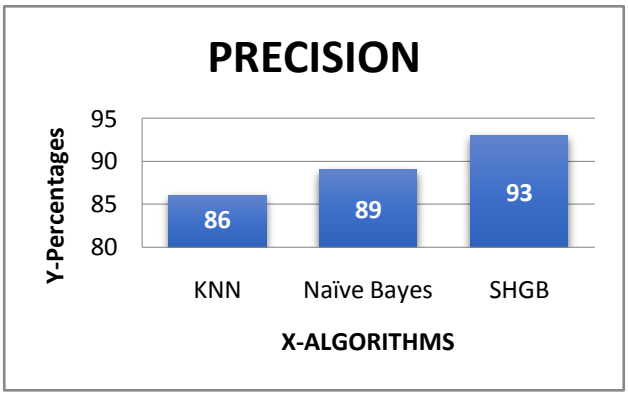

Figure 7 Comparison of Algorithms Based on Precision

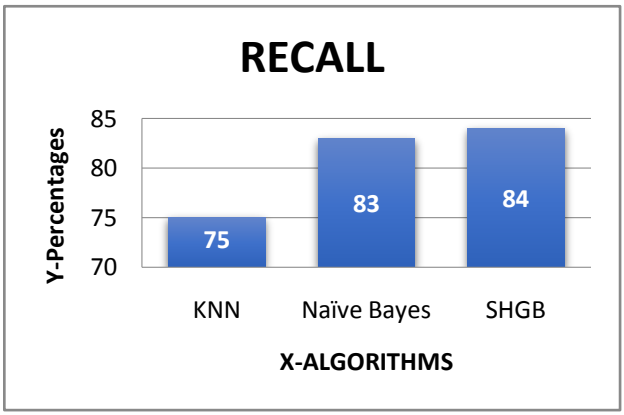

Figure 8 Comparison of Algorithms Based on Recall

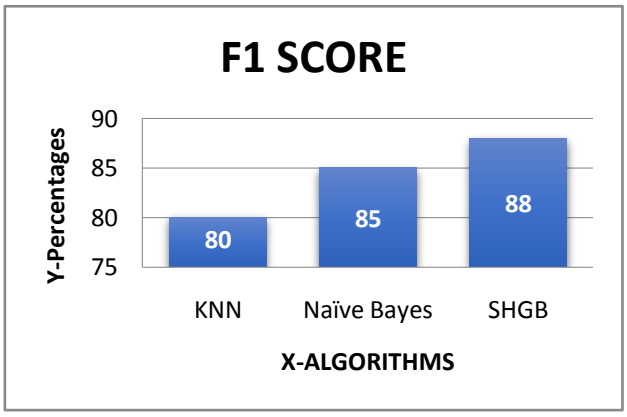

Figure 9 Comparison of Algorithms Based on F1 Score

The aim of this project is to know whether the patient has heart disease or not. The records in the datasets are divided into training set and test sets. After preprocessing the data, data mining classification technique namely naïve Bayes, KNN, SHGB were applied. This section shows the results of those classification model done using Python Programming.

Performance Results

Algorithms	Accuracy	Precision	Recall	F-Measure
SHGB	**88.52**	**93**	**84.37**	**88**
Naïve Bayes	86.87	89	83	85.89
KNN	80.32	86.20	75.75	80

Performance of these three algorithms such as SHGB, Naïve Bayes and KNN are compare based on their accuracy, Precision, Recall and F1-Score. The above figure 6 indicates that among of these three algorithms, SHGB have achieved highest accuracy of 88.52% as compare to Naïve Bayes and KNN algorithms and figure 7, figure 8 and figure 9 also indicates that compare to other two algorithms SHGB achieves the better result in all parameters such as precision, recall and f1-Score. Table 2 illustrates the performance results of these three algorithms.

VIII. CONCLUSION AND FUTURE WORK

In this paper, three supervised data mining algorithms was applied on the dataset to predict the possibilities of having heart disease of a patient, were analysed with classification model namely Naïve Bayes Classifier, KNN and SHGB classification. These all algorithms are applied to the same dataset in order to analyse the best algorithm in terms of accuracy. The SHGB has predicted the heart disease patient with an accuracy level of 88 % and Naïve Bayes classifier has predicted heart disease patient with an accuracy level of 86 % and KNN classifier has predicted heart disease patient with an accuracy level of 80 %. Thus, I conclude this project by saying SHGB Classification algorithm is best and better for handling medical dataset. In the future, the designed system with the used machine learning classification algorithm can be used to predict or diagnose other diseases. The work can be extended or improved for the automation of heart disease analysis including some other machine learning algorithms.

IX. REFERENCES

[1] VirenViraj Shankar, Varun Kumar, Umesh Devagade, Vinay Karanth, K. Rohitaksha Jiawei Han and Micheline Kamber, "Heart Disease Prediction Using CNN Algorithm", Springer 2020.
[2] Sarthak Vinayaka, P. K. Gupta "Heart Disease Prediction System Using Classification Algorithms", Springer 2020.
[3] FatmaZahra Abdeldjouad, Menaouer Brahami, Nada Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques",18th International Conference, ICOST 2020.
[4] Manoj Raman , Vijay Kumar Sharma, Saroj Hiranwal, Amit Kumar Bairwa "Efficient Method for Prediction Accuracy of Heart Diseases Using Machine Learning" Springer Nature Singapore Pte Ltd. 2021.
[5] Surendra Singh and Vishal Shrivastava, "The Analysis of Machine Learning Techniques for Heart Disease Prediction", Springer Nature Singapore Pte Ltd. 2021.
[6] Ching-seh (Mike), Mustafa Badshah, Vishwa Bhagwat: "Heart Disease Prediction Using Data Mining Techniques" DSIT'19, July, 2019.
[7] Sateesh Ambesange, Vijayalaxmi A, Sridevi S, Dr. Venkateswaran, Dr. Yashoda B S :" Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques", IEEE 2020.
[8] Mamatha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", IEEE 2019.
[9] Mr.Santhana Krishnan.J , Dr.Geetha.S, "Prediction of Heart Disease Using Machine Learning Algorithms",IEEE

