International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Heart Attack Prediction by using Machine Learning Techniques

Sangya Ware, Shanu k Rakesh, Bharat Choudhary

Abstract*: *Heart disease is most common now a days and it is a very serious problem. Machine learning provides a best way for predicting heart disease. The aim of this paper is to develop simple, light weight approach for detecting heart disease by machine learning techniques. Machine learning can be implemented in heart disease prediction. In this paper different machine learning techniques have been used and it compares the result using various performance metrics. This study aims to perform comparative analysis of heart disease detection using publicly available dataset collected from UCI machine learning repository. There are various datasets available such as Switzerland dataset, Hungarian dataset and Cleveland dataset. Here Cleveland dataset is used which is having 303 records of patients along with 14 attributes are used for this study and testing. These datasets are preprocessed by removing all the noisy and missing data from the dataset. And then the preprocessed dataset are used for analysis. In this study six different machine learning techniques were used for comparison based on various performance metrics. The analysis shows that out of six techniques SVM gives the best result with 89.34%. A GUI is developed for the prediction of heart disease.

***Keywords*: *SVM, Machine Learning, datasets, Heart disease prediction, analysis*.

I. INTRODUCTION

Heart disease has created a lot of serious problems; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. There are various medical instruments available in the market for predicting heart disease but they are very much expensive and they are not efficient enough to be able to calculate the chance of heart diseases. There is a need to find better and efficient approach to diagnose heart diseases at early stage [1]. With advancement of computer science in different research areas including medical sciences, this has been made possible. A machine-learning system is trained rather than the explicitly programmed. Machine learning could be a better choice for achieving high accuracy for detection of heart diseases. This paper is dedicated for wide scope survey in the field of machine learning technique in prediction of heart disease. Later part of this paper will discuss about various machine learning algorithms and their relative comparison based on various performance metrics like F1 Score, specificity, accuracy etc.

Machine learning techniques

"With the help of Machine Learning computers can learn and act like humans, and improve their learning by feeding them the data and information in the form of observations." There are various machine learning techniques available. In this study six different algorithms were used for analysis and comparison.

A. *SVM*

Support vector machines (SVM) are supervised learning models used for classification and analysis. The main objective is to create a hyper plane.

B. *Random forest*

It is a supervised classification algorithm. As a name suggests, algorithm creates the forest with a number of trees and higher the number of trees in the forest higher will be the accuracy. It will handle the missing values. If there are more trees in the forest, random forest classifier won't overfit the model.

C. *KNN*

The k-nearest neighbors (KNN) algorithm is a simple and easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

D. *Decision Tree*

A decision tree is tree-like graph or model of decisions and their possible consequences. In decision tree each internal node represents a "test", each branch represents the outcome of the test, and each leaf node represents a class label.

E. *Naïve Bayes*

This classifiers calculates the probabilities for every factor. It is based on the Bayes Theorem for calculating probabilities and conditional probabilities.

F. *Logistic Regression*

Logistic regression is a predictive analysis technique. In Machine Learning it is used for binary classification problems. It predicts the outcome in a binary variable which has only two possible outcomes.

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.

***** Correspondence Author

Sangya ware*, M. Tech scholar, CSE Department, ChoukseyEngineering College, Lalkhadan Bilaspur (C.G.). Email : sangyaware@gmail.com

Shanu K Rakesh, Assistant Professor, CSE Department, Choukesy Engineering College, Lalkhadan Bilaspur (C.G.) Email: Shanu.kuttan28@gmail.com

Bharat Choudhary, Assistant Professor, CSE Department, Chouksey Engineering College, Lalkhadan Bilaspur(C.G.) Email : hi.bharat2002gmail.com

It is a technique to analyze a data-set which has a dependent variable and one or more independent variables. Dependent variable is also referred as target variable and the independent variables are called the predictors.

II. PROPOSED WORK

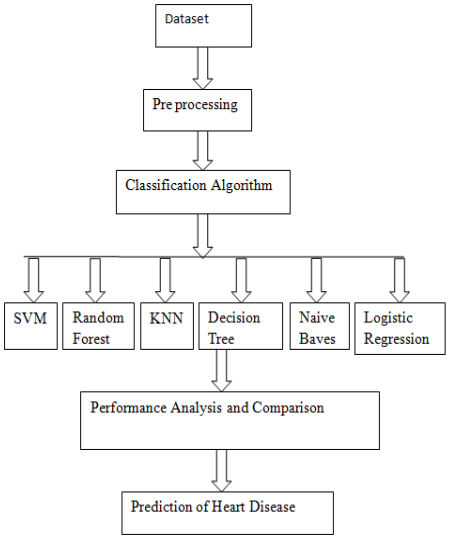

Figure 1: Proposed work

Figure 1 shows the proposed work. Where the first step is dataset, select the dataset for analysis and comparison, second step is preprocessing. Preprocessing is a process of removing all the noisy and missing data. Third step is classification algorithm where different algorithms are used for comparison. In the next step performance is analyzed and algorithms are compared and find out the best algorithm based on accuracy. At last it will predict whether a person has heart disease or not.

III. METHODOLOGY

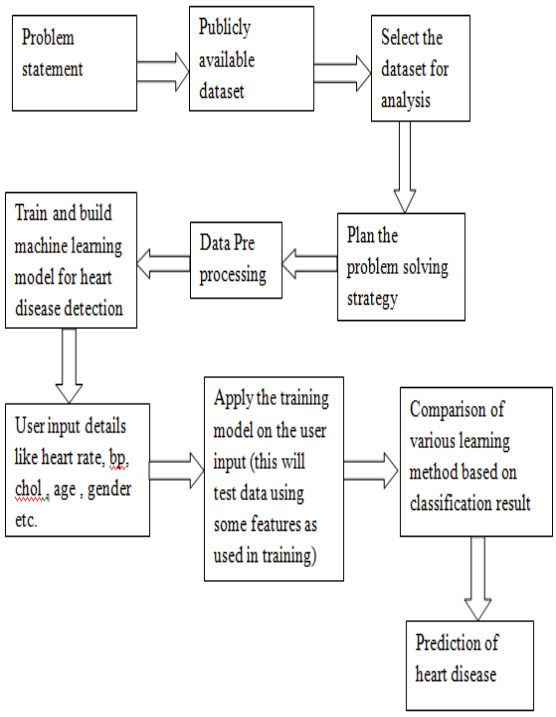

Figure 2: Design and approach

Figure 2 shows the project plan where first step is to identify the problem statement, where the problem is identified weather the person has heart disease or not. The next step is publicly available dataset. There are many datasets available in UCI machine learning repository such as Hungarian dataset, Switzerland dataset and Cleveland dataset. The next step is to select the dataset. So in this thesis used Cleveland dataset for analysis and comparison. The next step is plan the problem solving strategy, for this using six classifiers, they are is SVM i.e. support vector machine and Random forest, Decision Tree, Navie Bayes, KNN, logistic Regression. The next step is Data preprocessing. Data preprocessing is a process of removing noisy and missing data from the data set. The next step is building a training model for prediction in this step build the model by publicly available dataset such as sex, age etc. give training by using different machine learning algorithms and the results were obtained by using different performance metrics. The next step is Apply the model on user data and predict heart disease. The next step is comparison of various machine learning algorithm based on classification results. In the last step it will compare all the classifiers and find out whether a person has heart disease or not.

A. *Problem Statement*

Study the dataset (Cleveland dataset) and to predict whether a person has heart disease or not. If a person has a heart disease it is represented by and if a person has no heart disease it is represented by 0.

B. *Select Dataset*

A data set (or dataset) is a collection of data which is usually presented in tabular form. There are many datasets available at the UCI machine learning repository [2]. Some of the datasets are Hungarian dataset, Switzerland dataset and Clevelanddatasethttps://archive.ics.uci.edu/ml/datasets/Heart +Disease [3]. Only 14 attributes are used for this study. The attributes are as follows–

1. Age
2. Sex
3. CP
4. Trestbps
5. Chol
6. Fbs
7. Restecg
8. Thalach
9. Exang
10. Old peak
11. Slope
12. Ca
13. Thal
14. Heart_disease_label

C. *Problem Solving Strategy*

Machine Learning techniques for good decision making in the field of health care addressed are namely support vector machine, decision trees, Artificial Neural Networks and Naive Bayes. Here 6 different algorithms were used for comparison such as svm, random forest, decision tree, knn, naïve bayes, logistic regression.

D. *Data Preprocessing*

It is a process of removing all the noisy and missing data from the data set.

E. *Train and build machine learning model for heart disease detection*

In this step the dataset is divided into two parts: training dataset and testing dataset. Training dataset contains 60% and testing dataset contains 40% which are selected randomly.

F. *Input Details*

User input details such as Age, sex, cp, Trestbps, chol, Fbs, Exang, Thalach, old peak, slope, ca, thal, restecg, class are the 13 attributes and 1 label.

G. *Comparison of various machine learning algorithms*

It this step the comparison is done between the classifiers. Different classifiers such as svm, random forest, knn, naive bayes, decision tree, logistic regression are compared based on the accuracy, precision, recall and f1 score.

H. *Prediction of heart disease*

A GUI is developed in python by using tkinter to generate a simple dialog box which takes input for all the values necessary for evaluation. After the input is taken from the user, prompt appears which decides whether a person has a presence of heart disease or not.

IV. RESULTS

```
Percentage of Persons Having No Heart Disease: 54.13%
Percentage of Persons Having Heart Disease: 45.87%
Accuracy of SVM Algorithm: 89.34%
Precision:  0.94
Recall is:  0.8245614035087719
f1 score is: 0.8785046728971964
Accuracy using Random Forest : 86.89%
Precision:  0.9361702127659575
Recall is:  0.7719298245614035
f1 score is: 0.8461538461538461
Accuracy using knn : 80.33%
Precision:  0.9230769230769231
Recall is:  0.631578947368421
f1 score is : 0.7499999999999999
Accuracy using Decision Tree: 74.59%
Precision:  0.7166666666666667
Recall is:  0.7543859649122807
f1 score is : 0.735042735042735
Accuracy using Naive Bayes: 85.25%
Precision:  0.8679245283018868
Recall is:  0.8070175438596491
f1 score is: 0.8363636363636363
Accuracy using lr: 85.25%
Precision:  0.8679245283018868
Recall is:  0.8070175438596491
f1 score is : 0.8363636363636363
>>>
```

Figure 3: Result

Table 1: Accuracy of an algorithm

Algorithm	Accuracy %
SVM	89.34
Random Forest	86.07
KNN	78.69
Decision Tree	74.59
Naïve Bayes	85.25
Logistic regression	81.97

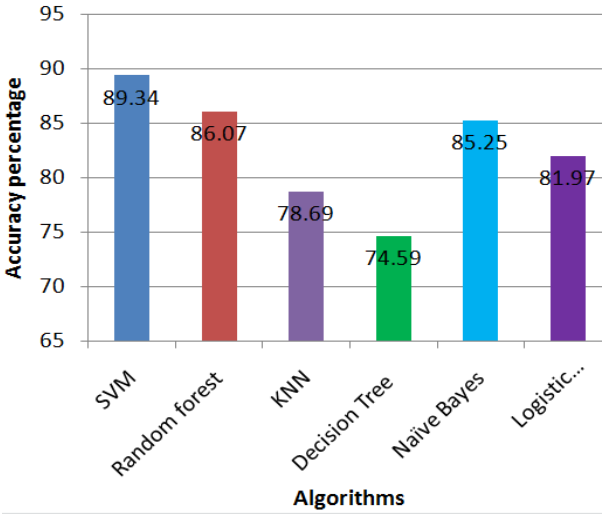

Figure 4: Accuracy of an algorithm

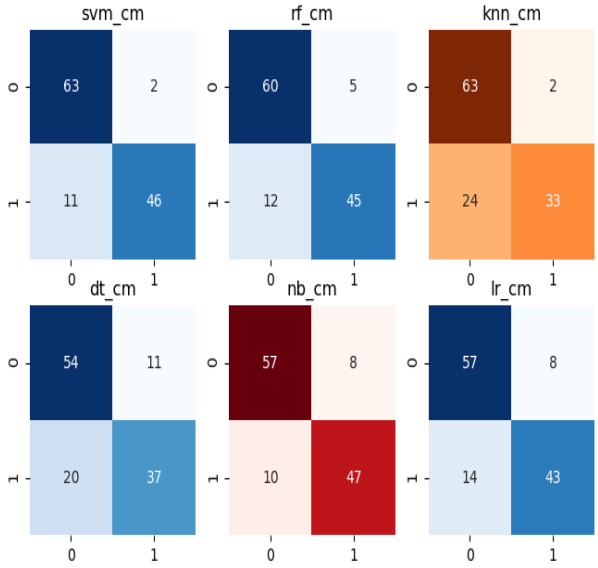

Figure 5: confusion matrix

Table 2: Comparative Analysis of Machine learning Algorithm

Algorithm	Precision	Recall	F1 Score
SVM	95.83	80.70	87.61
Random Forest	90	78.94	84.11
KNN	94.28	57.89	71.73
Decision Tree	77.08	64.91	70.04
Naïve Bayes	85.45	82.45	83.92
Logistic regression	84.31	75.43	79.62

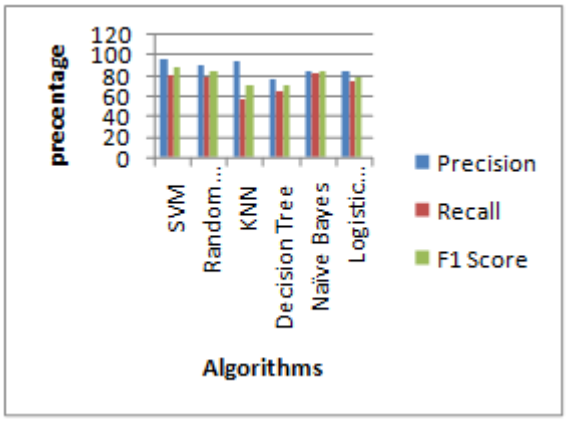

Figure 6: Graph for Precision, Recall and F1 score

V. CONCLUSIONS

The goal is to compare different machine learning algorithms by using performance metrics. Every algorithm performed better in some situation and worse in another. SVM work best in this study. Different devices can be manufactured which will monitor the heart related activities and diagnose the disease. These devices will be helpful where heart disease experts are not available. When tested through various situations, the algorithms performed differently which helped to understand the algorithm's working mechanism. This can be first learning step in heart disease diagnosis with machine learning and it can be extended further for future research. The dataset needed tremendous efforts for cleaning and had a lot of noisy and missing data. This would certainly improve the data quality and hence would improve the classification accuracy. It has proposed Supervised Learning Algorithm for finding the risk of heart disease of a patient using the profiles collected from the patients. This can detect heart related problems by using the model trained from a publicly available dataset. It is believe that only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease. In this paper six different algorithms are compared, which are used to predict heart disease. Form the comparison study; it is observed that the Support Vector Machine model turned out to be best classifier for Heart disease prediction.

FUTURE SCOPE

In future, plan is to extend the research by identifying and including more features. It also plans to use more classification methods like deep learning etc. In future, the plan is to explore and merge more datasets so that will get more effective dataset which includes large diversity of population. Also plan to implement various other classification methods to effectively support prediction. In future, the plan is to explore and merge more datasets to get more effective dataset which includes large diversity of population. In future the plan is to use the feature selection to get more relevant features and effective results.

REFERENCES

1. Noura Ajam, "Heart Disease Diagnoses using Artificial Neural Network", The International Insitute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11.
2. Dr.S.Seema Shedole, Kumari Deepika, "Predctive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.
3. http://archive.ics.uci.edu/ml/datasets/Heart+Disease
4. https://pdfs.semanticscholar.org/a9d8/4eac6a6b8721690bf976e99d2ffa986174fc.pdf
5. https://www.slideshare.net/pierluca.lanzi/machine-learning-and-data-mining-12-classification-rules
6. http://www.csjournals.com/IJCSC/PDF7-1/18.%20Tejpal.pdfhttp://archive.ics.uci.edu/ml/datasets/Heart+Disease
7. https://www.kaggle.com/cdabakoglu/heart-disease-classifications-machine-learning
8. https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c

AUTHORS PROFILE

Miss Sangya ware pursed Bachelor of Engineering from chouksey Engineering, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India in year 2016. Currently pursuing Master of Technology form Chouksey Engineering Bilaspur, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India. And has Published 4 research papers in reputed national and international journals.

Mrs. Shanu K Rakesh pursed Bachelor of Engineering from Institute of Technology, Gurughasidas University, Bilaspur, India in 2005and Master of Technology from RITEE, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India in year 2012. She is currently pursuing Ph.D. from Mewar University, India and currently working as Assistant Professor in Department of Computer Science and Engineering, Chouksey Engineering College, CSVTU, Bhilai, India since 2008. She has published 25 research papers in reputed national and international journals including her main research work focuses on Swarm Intelligence and Swarm Robotics. She has 12 years of teaching experience.

Mr. Bharat Choudhary pursed Bachelor of Engineering from Institute of Technology, Guru Ghasidas University, Bilaspur, India in 2008 and Master of Technology from RGPV University Bhopal in Computer Science and Engineering with specialization in Software Engineering, in year 2014. he is currently working as Assistant Professor in Department of Computer Science and Engineering, Chouksey Engineering College, CSVTU, Bhilai, India since 2010. He has published 15 research papers in reputed national and international journals including his main research work focuses on Software Testing. He has 10 years of teaching experience.