International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4s, November 2018

Analysis of Prediction Accuracy of Heart Diseases using Supervised Machine Learning Techniques for Developing Clinical Decision Support Systems

Kiruthikaa K V, Vijay Franklin J, Yuvaraj S

Abstract: ***Heart diseases are taking on hands as the vital mortality deciding factor in the current era. Most of the people around the world are experiencing a time-scheduled and stressful work life, which often leads to increase in the percentage of healthy people affected by heart diseases. It is mandatory to solve this raising issue by predicting the occurrence of the disease as earlier as possible with the help of variety of available solutions. Machine learning techniques can be applied to analyze and predict whether a person is likely to have heart disease or not. In this paper, we made a detailed investigation on prediction accuracy rate of heart diseases using different supervised machine learning techniques, which will pave the way for researchers to choose the efficient technique(s) in order to design and develop clinical decision support systems that predicts the occurrence of heart diseases in people efficiently.***

Keywords- Heart Disease, Machine Learning, Prediction Accuracy, Clinical Decision Support Systems

I. INTRODUCTION

Heart disease is common among people who have the triggering factors such as unhealthy diet, obesity, smoking, cholesterol, high blood pressure, high blood glucose, lipids, physical inactivity, risky use of alcohol and hypertension. World Health Organization (WHO) stated that cardiovascular diseases are the primary cause of increasing global death. According to WHO, 17.7 million people around the world die due to cardiovascular diseases, which estimated 31% of all deaths worldwide as of 2015 report. More precisely, 7.4 million people around the world die due to coronary heart diseases and 6.7 million were due to stroke.

A. Types of Heart Diseases

Heart disease is the term referred to the malfunctioning or disorder of the heart. Heart diseases can be classified into various types and some of the major types include:

- Congenital heart disease – Malformation of heart or parts of the heart since birth such as hole in the two chambers of heart, insufficient oxygen around the body and reduced blood flow to the chambers of heart.
- Coronary artery disease – Narrowing of coronary arteries which cause the heart to accept adequate oxygen and nutrients due to plaque build-ups.
- Arrhythmia – Irregular heartbeat due to improper working of electrical impulses in the heart.
- Myocardial infarction – Popularly known as heart attack, caused by a blood clot in one of the coronary arteries or if an artery narrows or spasms unexpectedly due to reduced blood flow.
- Heart failure – Heart does not pump an adequate amount of blood around the body.
- Dilated cardiomyopathy – Dilation of heart chambers due to reduced oxygen supply to heart muscle, which eventually leads to heart muscle weakness.
- Angina pectoris – Part of the heart fails to receive enough oxygen due to physical exertion.

II. SUPERVISED MACHINE LEARNING TECHNIQUES

Machine learning techniques integrate a variety of supervised machine learning techniques and algorithms that helps to develop Decision Support Systems (DSS) in the healthcare domain that involve large datasets and variables. These techniques and algorithms are used to predict the possibility of heart diseases in healthy people with the help of supporting documents such as electronically generated health records and history of heart disease patients obtained from collected datasets. Supervised machine learning techniques include classification and regression algorithms such as Support Vector Machines (SVM), Discriminant Analysis, Naive Bayes, Nearest Neighbor, Linear Regression, Support Vector Regression, Decision Trees, Ensemble Methods and Neural Networks.

III. CLINICAL DECISION SUPPORT SYSTEMS

Generally, Decision Support Systems (DSS) are intelligent and able to do take decisions from the vast amount of information based on certain situations. They are more complex in nature, to perceive and act upon, the decisions are made. Clinical Decision Support System (CDSS) is a special class of computer program application that analyzes large volume of patient's data and presents it to the medical professionals and practitioners to make decisions with best optimum results more efficiently, thus serving to the healthcare industry.

Manuscript published on 30 November 2018.
***Correspondence Author(s)
Kiruthikaa K V, Department of Computer Science Engineering, Bannari Amman Institute of Technology, Sathyamangalam-638401, India.
Dr. Vijay Franklin J, Department of Computer Science Engineering, Bannari Amman Institute of Technology, Sathyamangalam-638401, India.
Yuvaraj S, Department of Computer Science Engineering, Bannari Amman Institute of Technology, Sathyamangalam-638401, India.

IV. PREDICTION ACCURACY METRICS

Prediction of heart disease can be determined using accuracy metrics as:

Accuracy = (TP+TN)/ (TP+TN+FP+FN) (1)

Where,

- TP – Number of true positives (number of instances which are unhealthy and predicted correctly)
- TN – Number of true negatives (number of instances which are healthy and predicted correctly)
- FP – Number of false positives (number of instances which are healthy but predicted wrongly)
- FN – Number of false negatives (number of instances which are unhealthy but predicted correctly)

The obtained result is multiplied by 100 to get the value in percentage.

V. COMPREHENSIVE ANALYSIS OF PREDICTION ACCURACY

T. S. Brisimi et al. (2018)[1] formulated his solution for heart disease prediction using SVM, sparse logistic regression and random forests. In order to provide solution for the identified problem, the authors proposed two novel methods: K–LRT, a likelihood-ratio-based method, and Joint Clustering and Classification (JCC) method to identify the hidden patient clusters. Each of these methods adapts different classifiers to each identified cluster. While solving JCC, Mixed Integer Programming (MIP) problem was raised which resulted in creation of binary indicator variables that are used to assign clusters.

Due to absence of general polynomial-time algorithms, it was not possible to process large data sets and solve MIP problems. As a result, Alternating Clustering and Classification (ACC) approach was proposed to overcome the identified drawbacks. ACC depends on re-clustering the positive samples. The dataset was obtained from Boston Medical Center (BMC) which consists of 3033 patients with heart disease. Heart data pre-processing consists of 212 features for each patient. 60% of the dataset was used for training and 40% was used for testing. During their study, ROC curve was produced based on the patients affected by 10-year risk cardiovascular disease provided in the Framingham Heart Study (FHS) using random forest technique. The features used in the prediction include age of the patient, presence of diabetes, presence of smoking habit, blood pressure, cholesterol measure, level of high-density lipoprotein (HDL), and the value of body mass index (BMI). Framingham risk factor (FRF) was calculated for all the patients and classification is done based on the risk factor. After experimental analysis, in terms of average prediction accuracy, random forests technique (81.62%) was predicted as the best technique followed by the ACC approach (77.06%).

J. Zhang (2017)[2] proposed a system that makes use of Fast Fourier Transformation and ensemble model. Fast Fourier Transformation is applied in order to decompose the time series data and thereby extract the measurable features which are then fed into to the ensemble model. The output of the system is the prediction concerning whether a patient requires clinical test or not. Experimental analysis was conducted on real-time dataset obtained from Tunstall Healthcare which consists of data from 6 patients contributing 7,147 different time series health records. 75% of the dataset was used as training data and the remaining 25% was used for testing. The accuracy of the proposed model was 87.00%, 93.00% and 94.83% for three-feature set, six-feature set and eight-feature set respectively. This result was compared with three techniques: Neural Networks, Least Square-Support Vector Machine (LS-SVM) and Naive Bayes which exhibited less accuracy when measured against the proposed system.

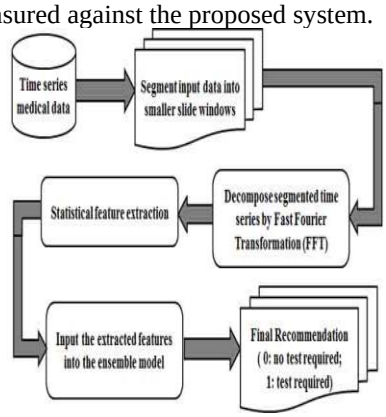

Fig 1. Fast Fourier transformation ensemble model [2]

R. G. Saboji (2017)[3] proposed a scalable solution for heart disease prediction using random forest technique. The technique was implemented on Apache Spark framework for efficient decision making. The dataset used for the experiment was obtained from Cleveland, Switzerland and Hungary databases of UCI machine learning repository which consists of 14 attributes and 600 records related to heart disease. The training dataset and testing dataset consists of data in the proportion of 70:30. The proposed solution using random forest generated the prediction accuracy rates as 88%, 96% and 98% for the number of records 200, 400 and 600 respectively. The result was compared with Naive Bayes technique which produced 44%, 55% and 64% prediction accuracy rates for 200, 400 and 600 records respectively. From the comparison result, it is analyzed that, though random forest technique outperformed naive bayes, the increase in accuracy rate was 8% initially when carrying out the experiment from 200 to 400 records and the rate enormously diminished to 2% increase when modifying the dataset from 400 to 600 records.

S. Pouriyeh et al. (2017)[4] made a detailed investigation and comparison on machine learning techniques that are used for predicting the possibility of heart diseases. The authors extracted the dataset from Cleveland data set of UCI repository that consists of 303 instances and 14 attributes in order to predict the possibility of heart diseases.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The different types of machine learning techniques used for comparison includes Decision Tree (DT), Naive Bayes (NB), Multi Layer Perceptron (MLP), K Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM). The authors used 10-fold cross validation technique for data portioning. When the seven mentioned techniques are applied with 10-fold cross validation, DT attained 77.55% accuracy, NB 83.49%, K-NN 83.16%, MLP 82.83%, RBF 83.82%, SCRL 69.96% and SVM 84.16%. From the results, it was clear that SVM attained the highest prediction accuracy rate.

In order to improve the estimation performance authors applied bagging technique, which showed improvement of accuracy in some of the classifiers such as DT, NB, MLP and more drastically in SCRL from 69.96% to 80.52% while NB and RBF showed decreased accuracy and interestingly SVM maintained the same accuracy rate. To better the performance results of weak classifiers, the authors applied boosting techniques, and achieved improvement in DT and SCRL techniques while others remained approximately same. As a final experiment, the authors applied stacking technique on the combination of stated classifiers, which resulted that SVM + MLP produced the best accuracy rate of 84.15%.

Tulay Karayilan and Ozkan Kilic (2017)[5] developed a heart disease prediction system which uses Artificial Neural Network's Backpropagation Algorithm. The authors executed their experiment on Cleveland database which contains 303 instances and 14 attributes related to heart disease prediction. Initially the experiment was conducted with hidden layer size varying from 3 to 12 neurons that resulted in the highest prediction accuracy rate as 86.67%. To perk up the performance percentage based on dimensionality reduction, Principal Component Analysis (PCA) method was executed which reduced the number of neurons at the input layer from 13 to 8 neurons. The generated results showed that the proposed technique outperformed the other techniques with 95.56% prediction accuracy rate.

K. Pahwa and R. Kumar (2017)[6] compared two approaches namely random approach and the proposed hybrid approach using Naive Bayes and Random Forest techniques to predict the occurrence of heart disease. For analysis purpose, the authors have used Cleveland Heart Disease dataset that consists of 303 records. Out of 13 attributes available, the proposed framework had chosen the best suitable features to predict heart disease by applying SVM-Recursive Feature Elimination (SVM-RFE) algorithm and gain ratio. In random approach, 10 features were selected for accuracy prediction using both the Naive Bayes and Random Forest techniques. The results showed that prediction accuracy was 78.22% for Naive Bayes and 76.90% for Random Forest. In hybrid approach, prediction accuracy is determined for different number of features in the order of 12, 10, 8 and 6. As a result, Naive Bayes attained the highest prediction accuracy of 84.16% for 10 best features, whereas Random Forest technique attained only 82.84% for the same 10 feature set. Interestingly, the accuracy prediction was highest in Random Forest technique with 84.16% for 12 best features compared to that of Naive Bayes with 83.83%.

Abhishek Rairikar et al. (2017)[7] employed three techniques namely k -nearest neighbors algorithm (KNN), Decision Trees (DTs) and Naive Bayes to analyze the clinical dataset that consists of 909 heart disease patients records with 13 medical attributes. The authors proposed that KNN was very accurate and faster compared to the other two techniques but they do not specify the numerical values of accuracy.

Shahed Anzarus Sabab et al.(2016)[8] applied three supervised learning techniques: C4.5 DT, SVM and Naive Bayes on medium sized dataset obtained from Department of Computing of Goldsmiths University of London, which consists of 303 cardiovascular patient records with 14 attributes. 66% data of the dataset were contributed for training and the remaining 34% were contributed for testing. Initially their prediction accuracy was 83.17% for Naive Bayes, 82.84% for SVM and 73.93% for Decision Tree, which was relatively poor compared to previous works in the similar domain. So the authors applied ranker algorithm for feature selection to remove the extraneous attributes. The implementation results showed that the prediction accuracy was improved by 3.63% with 8 best attributes for Naive Bayes, 4.95% for SVM with 9 best attributes and 5.94% for Decision Tree with 5 best attributes. Thus the largest accuracy was given by SVM (87.79%), followed by Naive Bayes (86.60%) and Decision Tree (79.87%) respectively.

M. A. Jabbar and S. Samreen (2016)[9] make use of Hidden Naive Bayes (HNB) model for prediction of heart disease. For experimental analysis, the authors utilized heart Statlog dataset from the UCI repository which consists of 270 instances of records with 14 attributes. Out of 270 instances, 273 were chosen as training data and the remaining 27 were chosen as testing data. The 10 fold cross validation technique was used for evaluating HNB technique. The authors applied pre-processing filter discretization and Inter-Quartile Range (IQR) filter to divide the data set into quartiles. The proposed HNB + IQR model achieved 100% accuracy in predicting heart disease compared to other related works which used Naive Bayes. The accuracy was improved by 14.82% in the proposed model wherein the Naive Bayes approach along with CFS Feature selection attained only 85.18%, as proposed by T. John Peter and K. Somasundaram (2012)[10].

J. Singh et al. (2016)[11] developed a framework for prediction of heart diseases using a hybrid technique that comprises of classifiers such as Naive Bayes, J48 Decision Tree, ZeroR, OneR and k-Nearest Neighbour (IBk). To run their experiment, the authors used 313 instances of records with 14 attributes from Cleveland Heart Disease database from the UCI Machine Learning Repository.

Two associative algorithms: Aprior and FP-Growth is employed to select the best 10 rules from each method. Experimental analysis shows that the prediction of accuracy is 99.19% for IBk, which is the highest rate among the other classifiers, while using Aprior association on heart diseases dataset. Comparatively, the accuracy prediction was 97.55% for Naive Bayes, when used with FP-Growth algorithm on heart diseases dataset. From the results, it is observed that the IBk (k Nearest Neighbor) with Aprior associative algorithms produced better results among all other classifiers used in the proposed model.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

S. Radhimeenakshi (2016)[12] proposed a model to predict the early diagnosis of heart diseases using SVM and ANN techniques. The source dataset was obtained from Cleveland Heart Database and Statlog Database under UCI machine learning dataset repository which consists of totally 14 attributes. The performance evaluation of SVM has shown 84.7% in terms of accuracy using Statlog database whereas it is 81.8% for ANN. The authors do not mentioned the prediction accuracy rate of both SVM and ANN models using Cleveland database.

M.A.Jabbar et al.(2016)[13] developed an intelligent heart disease prediction system using random forest and evolutionary approaches. The authors make use of chi-square and genetic algorithm as feature selection measures on heart disease datasets that include collected T.S dataset and also heart stalog dataset. A total of 11 attributes were selected for heart disease TS dataset and 14 attributes were chosen for heart stalog dataset. 10-fold cross validation technique also applied on all the instances of the data set.

The proposed system attained the highest prediction accuracy rate as 83.70% which outperformed other learning techniques such as PART C4.5 Decision Tree with 75.73%, Naive Bayes 78.56%, Decision Table 82.43%, and Neural Networks 82.77% when executed the experiment using heart Statlog dataset. The proposed system produced 100% prediction accuracy on heart disease T.S dataset compared to Decision Trees which produced only 98.66% for the same dataset.

Zriqat et al.(2016)[14] made a comparative study on five types of classification algorithms: Naive Bayes, DT, Discriminant Analysis, Random Forest, and SVM to predict the occurrence of heart diseases. The authors carried out the experiment on two different datasets obtained from the Cleveland Clinic Foundation that consists of 303 records and Statlog dataset that consists of 270 records. At the beginning of the study, both datasets consists of 76 raw attributes which were then pre-processed and resulted in 14 attributes that can significantly predict heart diseases.

After experimental analysis, the authors showed that decision trees produced the highest prediction accuracy of 99% followed by random forest with 93.4% using Cleveland dataset. Similarly, when Statlog dataset was used, again decision trees showed up the best accuracy 98.15% followed by random forest with 91.48%.

S. Bashir et al. (2014)[15] proposed a decision support framework for intelligent heart disease diagnosis using the majority vote based novel classifier ensemble based on heterogeneous classifiers namely Naive Bayes, DT based on Gini Index (DT-Gini) and SVM. The authors used standard heart disease dataset from UCI repository that consists of 303 instances, in which 297 were complete with 14 attributes related to heart disease prediction. The experimental analysis showed that the proposed ensemble framework attained the best accuracy of 81.82% compared to the accuracies achieved by individual classifiers NB with 78.79%, DT 72.73% and SVM 75.76%.

M. Gudadhe et al. (2010)[16] developed a medical DSS using SVM and ANN for diagnosing the heart disease. To conduct the experiment using SVM technique, the authors used Cleveland Heart Database from UCI repository that consists of 13 attributes and 303 patient instances. Out of 303 instances, 6 tuples had missing values which were ignored during data cleaning process and only 297 instances were taken into consideration. 200 instances were applied for training and the remaining 97 were applied for testing The SVM produced accuracy rate as 80.41%.

In order to execute the experiment using ANN- Multi Layer Perceptron (MLP) technique, the same Cleveland Heart Database was used and Backpropagation was used as the learning algorithm for MLP. In this experiment, 220 instances were selected as training data and 83 were selected as testing data. Since the total number of attributes was 13, the input layer was designed with 13 nodes. The experiment resulted with 97.5% accuracy in predicting the heart disease, which outperformed SVM technique.

VI. SUMMARY AND FUTURE WORK

Table 1.1 Analysis of Prediction Accuracy Rates

S. No.	Authors	Proposed Best Technique	Prediction Accuracy	No.of Attributes
1.	T. S. Brisimi et al. (2018)[1]	Random Forest	81.62%	212
2.	J. Zhang et al. (2017)[2]	Ensemble Model	94.83%	8
3.	R. G. Saboji (2017)[3]	Random Forest	98%	14
4.	S. Pouriyeh et al. (2017)[4]	SVM + MLP	84.15%	14
5.	T. Karayilan and O. Kilic (2017)[5]	Neural Network	95.56%	14
	K. Pahwa and R. Kumar (2017)[6]	NB+Random Forest	84.16%	10
6.	Abhishek Rairikar et al. (2017)[7]	KNN	NA*	13
7.	S. A.Sabab et al.(2016)[8]	SVM	87.79%,	14
8.	M.A.Jabbarand S.Samreen (2016)[9]	HNB	100%	14
9.	J. Singh et al. (2016)[11]	IBk (k Nearest Neighbor)	99.19%	14
10	S. Radhimeenakshi (2016)[12]	SVM	84.7%	14
11	M.A..Jabbar et al.(2016)[13]	Random Forest	100%	14
12	Zriqat et al.(2016)[14]	DT	99%	14
13	S. Bashir et al. (2014)[15]	Ensemble Model	81.82%	14
14	M. Gudadhe et al. (2010)[16]	ANN	97.5%	14

NA-Not Available

In the recent years, Random Forest algorithm was widely used individually and in combination with any other learning algorithms for developing clinical decision support systems, based on our survey and analysis report. As a future enhancement, we would like develop an efficient clinical decision support system using two-tier architecture that comprises of random forest and support vector machine as learning techniques.

VII. CONCLUSION

Various supervised machine learning techniques are discussed in this paper. Most of the dataset were incomplete with missing values and erroneous data, which should be corrected before using them for evaluation of accuracy rate. This is achieved by pre-processing technique which eliminates all the incomplete and erroneous data.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Also, we have to select features that strongly contribute to the prediction of heart diseases, and hence feature selection using classifiers have high impact on the prediction of heart diseases. From the analysis results of prediction accuracy rates using different supervised machine learning techniques, it is suggested that the feature selection algorithm has highest impact on determining the prediction accuracy rates besides the selection of learning techniques alone. Also, we can understand that the prediction accuracy using same technique varies according to number of feature sets and number of patient instances. So it is clear that, in order to develop a highly efficient CDSS it must achieve highest accuracy in predicting heart diseases. Careful analysis of supervised learning technique, feature selection algorithm and dataset contribute to the problem definition.

REFERENCES

1. T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," in Proceedings of the IEEE, vol. 106, no. 4, pp. 690-707, April 2018.
2. J. Zhang et al., "Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment," in IEEE Access, vol. 5, pp. 10674-10685, 2017.
3. R. G. Saboji, "A scalable solution for heart disease prediction using classification mining technique," International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 1780-1785.
4. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207.
5. T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network," International Conference on Computer Science and Engineering (UBMK), 2017, pp. 719-723.
6. K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), 2017, pp. 500-504.
7. A. Rairikar, V. Kulkarni, V. Sabale, H. Kale and A. Lamgunde, "Heart disease prediction using data mining techniques," International Conference on Intelligent Computing and Control (I2C2), 2017, pp. 1-8.
8. S. A. Sabab, M. A. R. Munshi, A. I. Pritom and Shihabuzzaman, "Cardiovascular disease prognosis using effective classification and feature selection technique," International Conference on Medical Engineering, Health Informatics and Technology (MediTec), 2016, pp. 1-6.
9. M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden Naive Bayes classifier," International Conference on Circuits, Controls, Communications and Computing (I4C), 2016, pp. 1-5.
10. T. John Peter and K. Somasundaram, "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction", International Journal of Scientific and Research Publications, Volume 2, Issue 10, 2012.
11. J. Singh, A. Kamra and H. Singh, "Prediction of heart diseases using associative classification," 5th International Conference on Wireless Networks and Embedded Systems (WECON), Rajpura, 2016, pp. 1-7.
12. S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3107-3111.
13. M.A.Jabbar, B.L.Deekshatulu and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, 2016, Volume 4, pp.175-184.
14. Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh (2016)" A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, 2016.
15. S. Bashir, U. Qamar and M. Younus Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," International Conference on Information Society (i-Society 2014), 2014, pp. 259-264.
16. M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and Artificial Neural Network," International Conference on Computer and Communication Technology (ICCCT), 2010, pp. 741-745.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)