IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis of Heart Disease Detection Techniques Using Machine Learning

[1]Prachi Chanchlani, [2]Dr. Madan Lal Saini
[1]M.Tech Student, [2]Assosiate Professor
[1]*Department of Computer Science, Poornima University, Jaipur, Rajasthan,
[2]Department of Computer Science, Poornima University, Jaipur, Rajasthan
[1]prachic1496@gmail.com, [2] madan.saini@poornima.edu.in

ABSTRACT

Heart Disease is the life threatening disease which is increasing at a very fast pace as many people suffer due to this every year and at present, heart disease is the number one cause of death worldwide. This disease affects heart as well as other body part so to provide an efficient analysis of disease, helps to the medical team in treatment. Early detection of this disease can save many lives by proper treatment. The traditional diagnostic methods like blood tests, electrocardiogram, cardiac computerized tomography scan, cardiac magnetic resonance imaging, etc are time consuming and/or invasive. In this review paper, many heart disease detection techniques were studies which were proposed in past few years. This paper narrates about the methodologies used by the researchers and accuracy claimed by them. For examining the claimed accuracy a common data set was used and a comparison table of various techniques was given in result discussion section. This paper presents the techniques which having acceptable accuracy level.

Keywords - Heart Disease Detection Techniques, Machine Learning and Deep Learning for Heart Disease**INTRODUCTION**

Health is a most crucial challenges faced by the world. For individuals health is the fundamental rights according to World Health Organization. Proper Case of health is required to keep people fit. Due to these concern computer technologies and machine learning algorithms are being used in the recent times to develop various software for assisting the experts in the process of decision-making related to cardiac ailments in the early stage and then predicting the probability of the risk of a significantly diseases fatality which helps in extracting the meaningful patterns and knowledge [1]. The prediction system for heart diseases will help better treatment in small period of time, resulting in saving number of lives. Heart is the most vital part of the human body. At the size of a closed fist, heart is the most hard-working muscle in the body. The heart beats 115,000 times in a day, at an average rate of 80 times per minute. In an average lifespan of 70-year a human heart will beat more than 2.5 billion times. Even when a person is at rest, the heart continuously beats [2].

Effective and efficient automated heart disease prediction systems can be beneficial in healthcare sector for heart disease prediction. Our work attempts to present the detailed study about the different data mining techniques which can be successfully deployed. This will reduce the number of tests done. The comparison of different data-mining techniques for finding out their suitability for the desired job is required for the integration of varied data mining techniques with the existing medical decision support system.

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Among various life threatening diseases, heart disease has garnered a great deal of attention in medical research .The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. The diagnosis of heart disease is usually based on signs, symptoms of the patient. The natures of heart disease are complex and hence, the disease must be handled carefully.

I. METHODOLOGY

This paper exhibits detections of various heart diseases by various data mining techniques and their comparison. It will help our medical data analysts or practitioners to detect the heart diseases accurately and examine it in less cost. Main technique which is used in our work is examining the various techniques' efficiency by taking same dataset on the basis of various attributes. After comparison, on the basis of accuracy a model is proposed. The comparison is done for the source codes which are available on open source depository. In this paper a neural network was also proposed which consists of 3 hidden layers first hidden layer has 22 neurons, the second has 30 and the third layer has 40.

The data that we have used to build the decision support system consists of two different dataset that have the same number and types of attributes. One is the Cleveland Heart Diseases Dataset [3]. This was published in 1988 and is divided into four smaller datasets from different hospitals. The other dataset is the Statlog Heart Dataset [4] which was published in 1993. The Cleveland dataset contains 303 instances with 14 attributes and the statlog dataset contains 266 instances with the exact same attributes (the target) us represented in the Cleveland dataset 0 means absence of any heart diseases and other values represents presence of a heart diseases while in Statlog dataset 1 represents absence and 2 indicates presence of heart diseases. During the pre-processing this discrepancy is resolved. Below the reader can find a table with the description of the 14 attributes and their types

The Described data has been used to train and test our classifiers. The Dataset is also quite balanced, with 310 Instances belonging to the absence of category and 259 belonging to the presence of disease category.

II. EXPERIMENTATION

Following experimental scenarios were considered during experimentation in the research work.

- Accomplish Cleveland Heart disease dataset (303 instances and 14 attributes) and statlog heart dataset [266 instances and 14 attributes] by collecting it.
- Accomplish Classification accuracy of heart disease dataset on the basis of classifier performance

In classifications problem detection of heart diseases is possible by the help of patient's measurements. In order to find the best classifier we have tested several well-known methods: Ada boost, Deep Neural Network [5].

It's important to note that since we are trying to detect a disease that can cause the death of the patients if he or she does not receive the treatment, we need take into account other performance measures apart from accuracy. The confusion matrix is an effective tool for visualizing the performance of the classifiers and it also allows us to impute precision and recall measures. In the context of heart diseases detection we consider two different classification errors. From the point of views of the classifiers, both types of errors should be minimized in the same way but, from the point of views of the patients a false negative is much more critical than false positive, because it puts the life of the patient at risk.

Table 1: Dataset's Attribute Description

Attribute	Value
Age	Real
Sex	Binary
Cp	Nominal
Chol	Real
Fbs	Binary
Thalach	Binary
Trestbps	Real
Restecg	Nominal
Exang	Binary
Old Peak	Real
Slope	Nominal
Ca	Real
Thal	Nominal
Num	Nominal

Precision is a kind of measurement that allows us to measure the ratio of type 1 and 2 errors separately [6]. In our case we will consider the precision of the type two error or negative predictive value (NPV) because it represents the proportion of healthy patients that are classified as means that we are minimizing number of false negatives, which represents that classified that tested in the first places the logistic regression classifier and we got the worst results with and accuracy value of 86.84% and a NPV of 0.85, which indicates that the system won't be able to detect 15% of the patients with heart diseases [7]. Then we

tried the nonlinear ensemble classifier ada boost, which turned out to be slightly better than logistic regression with accuracy of 89.51and an NPV of 0.89. After that we tried but problem with this configuration is that we have two hyper parameters; C parameter that we used to balance the classes and the gamma that us a parameter of the radial basis function. When we have hyper parameters and try to gain some knowledge about the relation between performance and the parameter value. But this approach is not rigorous and can be useless. In consequence we searched for hyper parameters optimization [8]. To have an idea about the effectiveness of opportunity we have compared the performance of a support vector machine after comparing the performance of the different classifiers we can conclude that the best method to diagnose heart disease based on the available features is the fully connected feed-forward deep neural network, therefore it is the model implemented in the final Decision Support System. However, we must add that even though the neural network shows charley the best before performance, its performance has some variability with respect to the initialization of the weights, while the other techniques are much more consistent with their performance [9]. The results of the different classifiers are all gathered in the table 2. The obtained results prove the effectiveness of the optimization of hype parameters, but we must say that given the small amount of data and the very low number of hyper parameters we have, we believe the same performance could be achieved by manually several combinations of Para filters [10]. The performance of the optimized support vector machine classifier was even better than Ada Boost with an accuracy of 92.30 and of 0.93 finally, for the last classifier we tested a connected deep neural network. The final neural networks consist of hidden layers, the input layer has 22 nodes and the output layer has two nodes. The hidden layer has 22 neurons, the second hidden layer has 30 nodes and the third hidden layer has 40 nodes. The network has a learning rate of 0.01, is trained for 1000 epochs and the batch size is 100. With this structure and hyper parameters value and positive predictive value are computed from the confusion matrix [10].

Table 2 Comparison of Machine Learning Techniques

Dataset	Technique	Accuracy
Statlog	Ensemble of 2 ANN with 1 hidden layer	81.92%
	SVM with RBF Kernel	83.70%
	Naïve Bayes	84.50%
	K-NN	82.90%
	Our Logistic Regression	82.35%
	AdaBoost	88.23%
	SVM	82.35%
	ANN	77.90%
Cleveland	C4.5[9]	77.56%
	Naïve Bayes[9]	83.50%
	SVM[9]	84.12%
	Logistic Regression	86.66%
	AdaBoost	85.33%
	SVM	86.66%
	ANN	80%

The Cleveland and Statlog datasets are old datasets (Late 1988 and 1993 respectively) and have been used in several research works. Therefore we provide a comparison between the performances. The techniques we have tested in this system the results of the older research works. It is important to note that the previous works always use only one of the two datasets so we have retested our models with the datasets separately.

This can be explained by the fact that we now have about half of the original number of samples. Other interesting things to renovations are that the AdaBoost classifier gives very with the Statlog dataset compared to the other techniques, and that our ANN has the worst accuracy. The reason for the bad of the ANN is that we are using the same model architecture that we use with the two datasets which is optimal for that particular setting, and with such a significant reduction of sample.

III. PROPOSED MODEL

More concretely feed forward multilayer perception, CNN architecture of neural network. Specifically, we train the neural network with a stochastic optimization method: Adaptive Moment Estimation. The model consists of 3 hidden layers with a

ReLu activation function and an output layer with a softmax activation function. The first hidden layer has 22 neurons, the second has 30 and the third layer has 40.

A. MAIN INDENTIFIED DECISIONS

Heart disease can be detected adequately when it is diagnosed in an early stage. Since obtaining a correct diagnosis is difficult due to complicated components of heart ailments. It is helpful to support the medical staff to give a good analysis on time. The diagnosis procedure is considered as a decision- making process, where the medical personnel must decide based on information available through clinical data in conjunction with the medical staff expertise. In this work, we propose an intelligent decision system. It helps doctors in the treatment.

Heart defects are basically diagnosed in two logics:

- Physical Evaluation.
- Clinical Evaluation.

During the physical evaluation stage, the physician collects the signs and symptoms of the patient. Also he records the various measurements. The patient is suspected to have a heart disease based on all the signs above. If patient is suspected of having a disease in the initial stage then only the doctors order the second stage tests and the treatment. But detecting a disease in the initial stage itself is not an easy task, because some of the patients may have the symptoms and signs and a few does not have any of them. So in such cases of expectation the process of diagnosis totally depends only on the basis of experience of the existing cases in which there are very high chances of incorrect decisions. If a wrong decision is taken and the disease is not taken and the disease is not detected in the early stage then doctors will never suggest for the second test. If incorrect analysis in the second stage tests then it will be a time consuming process itself which will raise the costs too for the treatment.

B. MODEL ARCHITECTURE

There are mainly eight steps of algorithms for prediction of heart disease.

1 Input disease datasets (heart disease dataset).
2 Select the attributes of the disease datasets.
3 Pre-processing of all the attributes can be done by applying filter.
4 Visualize all the attributes of disease dataset.
5 For creating a classifier select data mining classification algorithms.
6 Training and testing of all the parameters for classification is done.
7 Prediction and Classification of a disease are done.
8 Evaluate results and analyse the performance of classifiers.

C. SOFTWARE TOOLS

The Intelligent decision support system developed in python and uses several relevant libraries that add functionalities. We have chosen the programing language Python because it is a high level language that makes the development of software applications very simple and an important library used in this system is scikit-learn because it is used to create, train, and test Logistic Regression, AdaBoost and SVM classifiers. For the creation, training and testing of the deep ANN we use Tensor Flow because it is a very powerful library for developing complex Machine Learning applications.

For the pre-processing of data we use panda's library because it makes all the pre-processing much easier once the data is loaded in a data frame. The library provides a function to load a csv file into a data frame, it allows to easily applying operations to a whole column, which is useful when normalizing and standardizing the data, and finally it provides a function to generate dummy a categorical feature. The library Optunity has been used to find the optimal hyperparameters of the SVM with RBF model. Finally, we use ipywidgets and ipython libraries to create and support the GUI of the system. Through this GUI the user can enter the different measurements from the patient into the system using sliders and dropdowns, run the decision system and see the resulting decision.

V. RESULTS AND DISCUSSION

Two datasets namely Cleveland and Statlog were used to do the comparison analysis and for making decision support system. In both datasets there are total 15 attributes out of these; 8 were taken which are highly relevant in predicting heart diseases. Table 2 shows the accuracy of some machine learning algorithms which has its code on open source repository. We conclude that CNN is one of the best networks for data mining classification as it gives maximum accuracy in detection of risk in disease. MLP gives minimal accuracy with correctly classified instances results. If we compare other classification algorithm on heart cancer dataset so CNN gives maximum accuracy with correctly classified instances result on heart dataset. There were 288 instances in total. So we can say that CNN is better than other algorithm for heart disease detection and diagnosis. Dimension reduction helps to speed up algorithm and increases the accuracy. Dimensions were reduced from 15 to 8 which are highly relevant in predicting heart diseases. The result of this study helps the cardiologists to make more consistent diagnosis of

the heart conditions. Table 3 shows the accuracy for proposed Decision Support System in which the highest accuracy was 95.1% for Deep NN.

Table 3:- Performance Matrix of Proposed Model for Decision Support System

Model ⇨ Performance	Logistic Regression	Ada Boost	SVM	Deep NN
Accuracy	85.31%	89.51 %	92.31 %	95.10 %
Negative Prediction Value	0.84	0.89	0.93	0.94
Positive Prediction Value	0.87	0.91	0.91	0.96

From the above tables we learned that, algorithms performance differs on various situations. It shows performance in every scenario whether it is used or not. Each algorithm has its own intrinsic capacity which makes it better from other algorithms which totally depends on the situations.

Performance of algorithms decreases after boosting the data while algorithms were performing better when features were selected. It shows the necessity of feature selected data before applying boosting. As we used different data mining techniques for classification and prediction the disease but we can only deal with the predefined disease dataset. It means it doesn't found actual impact of parameters. So we should be take real life based parameters of patient's health to see the actual impact of parameters and it improves the classifiers accuracy.

The system we have developed provides highly accurate results (95.1%). But it still remains to be seen that whether the system will work for larger quantities of data, since we the data set we have is very small (569 rows). Another very useful feature would be to enhance the Graphic User Interface (GUI) already presented to the doctors. The GUI will not only display the result of the prediction, but also some insights of the patient. Identify the features that are more important for the prediction such as high cholesterol, etc can be considered in algorithms.

VI. CONCLUSION

The first objective of this paper is to compare various heart diseases detection techniques and second objective is to design a prediction model for detecting heart diseases using machine learning techniques. Two datasets namely Cleveland and Statlog were used to do the comparison analysis and for making decision support system. The result of comparison analysis is shown in Table2 which shows that AdaBoost is giving 88.23% accuracy which is highest compared to others for Statlog dataset. Logistic Regression and SVM are giving 86.66% accuracy for Cleveland dataset. The performance/accuracy of above models was evaluated by using standard metrics on the basis of accuracy, precision, recall and F- measure. The proposed decision support system is giving accuracy 95.10%, 92.31% and 89.51% respectively for Deep NN, SVM, and AdaBoost.

REFERENCES

[1] Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology Conference, *03 September 2015.*

[2] Md. Razu Ahmed, S M Hasan Mahmud, et al., "A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms", IEEE 4th International Conference on Computer and Communications, 2018.

[3] Cleveland Heart Diseases Dataset https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

[4] Statlog Heart Diseases Dataset http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/

[5] Sanchayita Dhar, Pritha Datta, Krishna Roy, Ankur Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Disease", 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[6] Martin Gjoreski, Anton Gradise , et al, "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers", 13th International Conference on Intelligent Environments, 2017.

[7] Noor Basha, Gopal Krishna C, Ashok Kumar P S, Venatesh P, "Early Detection of Heart Syndrome Using Machine Learning Technique", 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, 2019.

[8] Muhammad Fathurachman, Umi Kalsum, Chandra Prasetyo Utomo, Noviyanti Safitri, "Heart Disease Diagnosis using Extreme Learning Based Neural Networks", International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014.

[9] Amin Ul Haq, Muhammad Hunain Memon, et al., "Heart Disease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm for Features Selection", 5th International Conference for Convergence in Technology (I2CT) Pune, India, Mar 29-31, 2019.

[10] Jianping Li, Jalaluddin Khan, et al., "Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection Heart Disease Detection System", IEEE 4th International Conference on Computer and Communications, 2019.

[11] A. Sengur, "An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases." Expert Systems with Applications, 35(1): 214-222, 2008.

[12] M.E. Karar, S.H. El-Khafif and El-Brawany, "Automated Diagnosis of Heart Sounds Using Rule-Based Classification Tree", Journal of medical systems, 41(4): 60, 2017.

[13] M. Durairaj and N. Ramasamy, "A Comparison of the Perceptive Approaches for Preprocessing the DataSet for Predicting Fertility Success Rate," International Journal of Control theory and Applications, vol. vol. 9, pp. 255-260, 2016.

[14] S. Ghwanmeh, et al., "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," Journal of Intelligent Learning Systems and Applications, vol. 5, pp. 176-183, August 2013.

[15] Amin Ul Haq et al. "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol: 2018, pages 21, 2, December 2018.

[16] Chen, A.H., et al. "HDPS: Heart disease prediction system", in Computing in Cardiology, IEEE, 2011.

[17] Bashir, S., U. Qamar, and M.Y. Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis", in International Conference on Information Society (i-Society), IEEE, 2014.

[18] Noh, K., et al., "Associative classification approach for diagnosing cardiovascular disease", in intelligent computing in signal processing and pattern recognition, Springer. p. 721-727, 2006.

[19] Rajeswari, K., V. Vaithiyanathan, and T.R. Neelakantan, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks", Procedia Engineering, 41: p. 1818-1823, 2012.

[20] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach", International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5, doi: 10.1109/WITS.2019.8723839.

[21] P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-7, doi: 10.1109/ INOCON 50539.2020.9298354.

[22] A. U. HAQ et al., "Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection System," 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2019, pp. 130-138, doi: 10.1109/ ICCWAMTIP 47768.2019.9067519.

[23] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.

[24] M. Fathurachman, U. Kalsum, N. Safitri and C. P. Utomo, "Heart disease diagnosis using extreme learning based neural networks," 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), Bandung, Indonesia, 2014, pp. 23-27, doi: 10.1109/ICAICTA.2014.7005909.