Intelligent Diagnosis of Cardiac Disease Prediction using Machine Learning

Ravindhar NV, Anand, Hariharan Shanmugasundaram, Ragavendran, Godfrey Winster

Abstract—Cardiac disease have become worldwide common public health issue, mainly due to lack of awareness of health, poor lifestyle and poor consumption. Practitioners may have different concerns when it comes to disease diagnosis, which result in different decisions and actions. On the other hand, even in the specific case of a typical disease the amount of information available is so massive that it can be difficult to make accurate and reliable decisions. With adequate patient and non-patient medical constraints, it is possible to accurately predict how likely it is that a person with heart disease and to obtain potential information from these systems. A mechanized framework for therapeutic analysis would also dramatically increase medical considerations and reduce costs. We developed a framework in this exploration that can understand the principles of predicting the risk profile of patients with the clinical data parameters. In this article, four machine learning algorithms and one neural network algorithm were used to compare performance measurements to cardiac diseases identification. We evaluated the algorithms with respect to accuracy, precision, recall and F1 settings to achieve the ability to predict cardiac attacks. The results show our method achieved 98 percent accuracy by neural network algorithm to predict cardiac diseases.

Index Terms—Human Cardiac, Disease, Machine Learning, Prediction, Heart disease prediction System, CVD, Neural Network.

I. INTRODUCTION

Healthcare systems have dramatically increased human life expectancy by advancing the use of medications, healthcare services through the medical history management of patients. However, the health support systems are currently facing significant challenges with general healthcare such as minimal medical information, avoidable errors, a threat to information, inconsistencies in diagnosis and delayed transmission of medical data. On the other hand, even in a typical disease the amount of information available is so massive that it may be challenging to make consistent and accurate decisions. There exists many decision support systems for disease diagnosis process like Clinical decision support System (CDSS) and similar other systems like the Electronic Health Recorder (EHR). The subject of investigation aims for designing Disease Prediction Scheme (DPS) and several research communities around the world have considerable interest in this proposed research area. Computerized Medical information systems are always of great interest to researchers. It is possible to predict how likely an individual will suffer from illness with the medical parameters of a substantial number of patients and non-patients.

For instance, of an patient was unable to recollect all the symptoms of the disease and medical diagnosis that exists earlier for the patient, which is a vital resource for clinician helps for efficient decision making process. The management of the health records of patients for more precise diagnosis of the disease at an early stage was used in support of the doctor in the decision-making process. This has also raised an important question as: “What means would we be able to make a viable decision from the Information into helpful data, which can help healthcare service practitioners?” The main objective of this study mainly relies on this. Data mining techniques in this scenario are suitable for the conversion of accumulated raw data to useful, influential data to predict the system. Many methods have been used recently in the mining of data to predict disease in patients [6]. DPSS with data mining approach can be more promising for prediction of accuracy level and to reduce the diagnosis time and cost of the disease [4].

This problem is motivated by the assessment by the World Health Organization (WHO). An estimated 23.6 million persons will be diagnosed as having heart disease by 2030 according to the World Health Organization. The expectation of coronary heart disease should be completed to minimize this risk. Coronary disease analysis is typically carried out with regard to signs, manifestation, and the patient's physical examination. This huge mass of raw information is the rule resource that can be used and inspected before the extraction of important information, which directs or reinforces decision making for costs by remedial society. In order to detect, analyze, and predict various diseases, the ML techniques are more effective. One important goal is the efficiency of disease identification and a reduction of mortality caused by heart disease and prevention.

The organization of the paper goes in this order. The basic ideas related research study is initially presented in section I. Related research study is presented in section II. Section III briefs out the algorithm on the proposed idea. In section IV the analysis on the experiments and the evaluation metrics in...
Heart Disease Classification with forward heart network
Algorithms using Clever Land Sets with common factors involving age, sex, chest pain, blood pressure, steroid cholesterol, blood sugar, electrocardiographic etc have lead to 98% accuracy [5]. The risk was estimated with the use of 1,174 participations, including 456 women and 718. The overall algorithm with several factors dealt in the star of this paragraph repeatedly to find some interesting conclusions. The use of Fuzzy set hypotheses, case-based reasoning and K-NN based on the hybrid reasoning contributes to improved forecast results [6]. Cardiovascular risk prediction model development and validation using NIPPPON DATA have designed a tool for the prediction of cardiovascular risk using data from 2742 Japanese residents [7-8].

The diagnostic design models developed by research communities have been converted into point-based risk score sheets in which the data are divided in group based on age groups of 40 – 84 years. The midpoints for these categories were noted with an reference values with several categories. The marginal category reference values were determined and observations were recorded in this perspective. A reference group (0 points) was established for each variable as the lower risk class— e.g. the 40-49 age class, women with systolic blood pressure < 120 mmHg, and so on [15], Palaniappan and Awang [12] have examined the comparison of different methods of data mining for the diagnosis of patients with heart disease. The comparison included naïve bays, decision tree and the network of neural networks. The results showed that in the diagnosis of heart disease patients the naïve bays can achieve the highest precision [12]. Clustering is one of the most popular clustering techniques, however the first selection of the center is an important problem which strongly affects its results. The study examined different methods of initial centric selection for K-means clustering technique is used for diagnosis of problems with range values, outlier, random attribute values and random row values. Back propagation algorithm wide range of improvements were proposed with weights for training in feed-forward neural network. Such scheme uses gradient descent technique for supervised learning in multilayer perceptron to increase the efficiency with various strategies [14].
The neural network has 3 layers – an input layer, a hidden layer and an output layer. It uses forward propagation to compute, the activation (output) value of the $k^{th}$ output unit and $0$ represents the weights. The input layer comprises $13$ inputs and two for the back propagation perceptron output layer. The Hidden layers are defined by the various aspects such as the Frontline Solvers and Rule of Thumb Methods.

$$N_{max} = \frac{N_z}{\alpha * (N_o + N_f)}$$

$N_f$ - Number of input neurons.

$N_o$ - Number of output neurons

$N_z$ - Number of samples in training data set.

$\alpha$ - An arbitrary scaling factor usually $2$-$10$.

The above equation is the proposed approach by Frontline Solvers [16]. It indicated that the Hidden layers is defined by the ratio of the Total number of data in the training dataset $N_z$ divided by the sum of the inputs $N_f$ and outputs $N_o$, multiplied by an arbitrary scaling factor $\alpha$. There is also diversity approach such as rule-of-thumb [17-19] methods for determining the correct number of neurons to use in the hidden layers, such as the following:

- The number of hidden neurons should be between the size of the input layer and the size of the output layer.
- The number of hidden neurons should be $2/3$ the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than twice the size of the input layer.

The neurons are all similar in MLP. Each has multiple input links and output links. The values obtained from the preceding layer are summarized with some weights if the individual neuron and the bias term. The amount is processed by activation function. In order words, given the outputs of $O_i$ of the layer $n$, the output $Y_j$ of the layer $n+1$ are computed as:

$$u_i = \sum_j w_{ij} n^{j+1} * x_j + w_{i,bias}^{n+1}$$

$$y_{j} = \sigma(u_i)$$

Different Activation functions may be used. Machine Learning Implements three standard functions such as Identity function, Sigmoid and Gaussian function.

Sigmoid Function : $\sigma = \frac{1}{1-e^{-x}}$

In Machine Learning, all the neurons have the same activation functions, with the same free parameters, that are specified by user and are not altered by the training algorithms. Training algorithm calculates weights. The algorithm uses a training set of multiple input vectors and adapts the weights to the desired response from the network to the provided input vectors iteratively. The larger the size of the network, the more flexibility the network is. Arbitrarily small could be the mistake with the training set. But the learned network is also able to “learn” the noise in the training set, so that the error in the test set usually increases after the network size is limited. Moreover, larger networks are much more trained than smaller ones, so that only essential

**IV. ALGORITHM FOR DISEASE DETECTION**

Machine learning techniques have the ability to train and test the data using the raw data. For various application scenarios and types of data, a variety of machine learning algorithms have been developed and improved. We select five algorithms to detect disease or not in this paper. The algorithms selected are classified. The selected algorithms are widely used both in many fields. We would like to investigate how well these algorithms perform with respect to our datasets.

- Logistic Regression is a classification algorithm used as predictive analysis which determines the presence of Cardiac Diseases.
- Naïve Bayes technique is used as a classifier algorithm which classifies the diseases based on the parameters in medical diagnosis.
- Fuzzy Knn algorithm is mainly used as a classifier in supervised learning.
- K Means clustering algorithm is unsupervised learning used to solve clustering problems.
- Neural Network Back propagation used to solve complex problems.

Neural Network Back propagation perceptron is used for training and testing purpose. Every iteration, 50% of the data is used for training and testing purpose from the provided dataset in each iteration. The follow-up procedure after dataset splitting is equal for the testing and training of back propagation perceptron.

**Fig. 1. Neural Network back propagation perceptron**

<table>
<thead>
<tr>
<th>Thal</th>
<th>Desert Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fbs</td>
<td>Fasting Blood Sugar</td>
</tr>
<tr>
<td>Thalach</td>
<td>Greatest heart speed reached</td>
</tr>
<tr>
<td>Ex ang</td>
<td>Exercise induced angina</td>
</tr>
<tr>
<td>Old peak</td>
<td>ST despair induced by apply virtual to relax</td>
</tr>
<tr>
<td>Slope</td>
<td>Slope of the height effect ST section</td>
</tr>
<tr>
<td>Ca</td>
<td>Number of key vessels painted by fluoroscopy</td>
</tr>
<tr>
<td>Thal</td>
<td>Desert Category</td>
</tr>
</tbody>
</table>
features are reasonable to prepare the data and train a smaller network. This is can be optimized using the cost function. In Back Propagation, cost function to measure how close the predicted values are to their corresponding real values has follows.

\[
\text{Cost}(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[ y_i \ln(h_{\theta}(x_i)) + (1 - y_i) \ln(1 - h_{\theta}(x_i)) \right] + \frac{1}{2m} \sum_{i=1}^{m} \left( \theta^T W \right)^2
\]

Fig 2. Initial Count Of Colours

where \( h_{\theta}(x) \) is the activation function. The activation function can be calculated using the formula given below

\[ h_{\theta}(x) = \sigma \left( \sum_{i} w^i l^{i-1} + b^i \right) \]

where,

- \( w^i \) - Weight matrix for each layer \( i \)
- \( l^i \) - Input matrix for each layer \( i \)
- \( b^i \) - Bias vector

The gradient helps in optimizing the weights in order to minimize the value of the cost function.

V. EVALUATION METRICS

We compare the performance of diseased samples taken by Logistic Regression, Naïve Bayes, K–means clusters, KNN and back propagation algorithms. This section examines the methods chosen to assess performance, stability and scalability.

A. PERFORMANCE

In any setup of binary classification, all the test data points are classified as one of two classes, namely the positive and the negative classes. The corresponding labels (the "correct" or "true" label) are also included in the test set. Under such a framework, we define the following terms -

1. True positives (TP) - These are the data points which actually belong to the positive class and are classified as such
2. True negatives (TN) - The are the points which belong to the negative class and are classified as such
3. False positives (FP) - These are the points which actually belong to the negative class, but have been incorrectly classified as positive
4. False negatives (FN) - These are the points which actually belong to the positive class, but have been incorrectly classified as negative

The true positives and true negatives constitute the set of correctly classified data whereas false positives and false negatives are the set of points which were incorrectly classified.

All the above metrics are expressed in terms of absolute numbers.

To have comparability across datasets of different sizes, we express them in the form of ratios :

1. True positive rate (TPR) - This is the fraction of all positive class data points which were classified correctly
2. True negative rate (TNR) - This is the fraction of all negative data points which were correctly classified
3. False positive rate (FPR) - Since false positives are essentially negative data points in our test set which got incorrectly classified, we define the FPR as the proportion of all negative data points which got incorrectly classified.
4. False negative rate (FNR) - Similar to (3), FNR is the proportion of positive data points in our test set which got incorrectly classified.

The above definition also leads to the following result. Consider the set of all positive class data points in our test set. Each of them will get classified either as "positive", or as "negative". The positive points which get classified as "positive" end up in the set of "true positives". The positive points which get incorrectly classified as "negative" end up in the pile of "false negative" points.

The performance measurements shall be used for accuracy calculation and efficiency tests on True Positive, False Negative, True Negative and False Positive algorithms. The accuracy rate is the ratio of the true positive disease rate correctly identified over the total sample quantity as shown in the equation (1). The accuracy rate is calculated by the correctly identified disease ratio, as shown in the equation (2) by the total total amount of PT and FP. The recall sample shall be computed with the accuracy with the difference between the ratio of TP and the total sum of TP and FN as indicated in equation (3).

\[
\text{Accuracy} = \frac{TP}{(TP + TN + FP + FN)} \quad \cdots \quad (1)
\]

\[
\text{Precision (P)} = \frac{TP}{(TP + FP)} \quad \cdots \quad (2)
\]

\[
\text{Recall (r)} = \frac{TP}{(TP + FN)} \quad \cdots \quad (3)
\]

\[
F1 \text{ Score} = \frac{2 \times PR}{(P + R)} \quad \cdots \quad (4)
\]

Five different types of machine learning algorithm (Logistic regression, Naïve Bayes, Fuzzy K nearest neighboring, K–clustering and back propagation) are used in the experimental analysis of cardiac conditions. The technique of 10-fold cross-validation is used. Table 2 provides the best performance measurements for the diseased and not as 1 and 0. Cross-validation has been done to achieve the accuracy of each algorithm implementation. The neural network shows maximum precision among all five algorithms (Fig 2).

![Fig 2. Comparison of Various Machine Learning Algorithms](image-url)
VI. CONCLUSION AND FUTURE WORK

In the present lifestyle, cardiovascular disease is a major cause of death and heart disease. This paper proposes a model to recognize cardiac diseases using a machine learning approach based on regularized logistic regression, Naive Bayes, KNN, Clustering, Back Propagation (deep learning). The assessment compares several approaches to machine learning. The results obtained confirm that the algorithm of neural networks competes strongly against these algorithms in the detection of diseases according to the data set presented. In our model, the accuracy of our neural back propagation network is 98.2 percent, while the other algorithms are less exact than the other algorithms used.

REFERENCES

3. Ralph B. D’Agostino, Sr, PhD; Ramachandran S. Vasan, MD; Michael J. Pencina, PhD, Philip A. Wolf, MD; Mark Cobain, PhD;Joseph M. Massaro, PhD, William B. Kannel, “General Cardiovascular Risk Profile for Use in Primary Care”, Vol:117, pp: 743–753, (2008)
7. Takanari Hondaa, Daigo Yoshidaa,b, Jun Hataa,b,c, Yoichiro Hirakawa,b,c, Yuki Ishidaa,b, Mao Shibataa,b, Satoko Sakataa,c, Takanari Kitazonoc, Toshiharu Ninomiya,b, Development and validation of modified Risk Prediction models forcardiovascular disease and its subtypes: The Hisayama Study, Vol:279, pp:38–44,(2018)