Received 05 March 2021; Accepted 01 April 2021.

A study on Heart disease prediction with ECGSFS and Genetic Optimized NN classifier

C.Sowmiya*, Dr.P. Sumitra , C.Sowmiya *
Ph.D Research Scholar, PG and Research Department of Computer Science and Applications
Vivekananda College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tiruchengode-637205,
Tamilnadu, India. sowmiyac83@gmail.com

Abstract

Heart diseases are a more prevalent issue in modern medical situations. Every year a massive number of people perish due to this cardiac discomfort. Malapropos medications without the guidance of clinicians and detection of diseases at a later stage are the leading cause of these fatalities. The number frequency of mortality rate increases every year. This study presents an innovative classification technique with the utilization of evolutionary correlated gravitational search feature selection (ECGSFS) and Genetic optimized Neural network (GONN). Real-time implementation results and observations are clearly described. The present study achieved great performance in terms of precision, recall, F-measure and accuracy. A comparison is made with prior approach to evaluate the proposed work.

Keywords

Data mining, prediction, heart disease classification, feature selection

Introduction

There exist a various types of cardiac discomfort such as myocardial infarction, cardiac attacks, heart arrhythmia, and atrial fibrillation. It is immensely prerequisite for the clinicians to handle with the cardiac problems if it identified at a later stage and it can even lead to morbidity [1]. Thus, earlier detection and diagnosis of cardiac diseases are the significant requirements of the critical condition patients and electronic health medications. At present, there exists a variety of machine learning algorithms, which, abet in medical data classification and prediction processes but accurate prediction and classification is a challenging task [2]. The fundamental requirement of the classifiers, it should be computationally efficient and cost-effective, it is clear that an efficient classification algorithm with improved accuracy measurement is the inevitable part of real-time heart disease prediction and management systems [3].

The present system of healthcare detects and treats diseases through the use of various tests and medications which, are expensive and time consuming for the patients. Further, the delay in test reports results in the delay of medications. The consequences of certain diseases such as cardiovascular diseases become highly severe with next prediction and medication measures. The application of data mining techniques across healthcare systems reduces both time and money to the users. However, selection of appropriate classification technique for various healthcare domains is highly tricky. The primary objective of this work is to define efficient techniques for earlier prediction and diagnosis of heart diseases. The use of GONN classifier provides efficient classification process and the feature selection techniques improve the prediction accuracy.

Related Works

Liu et al., 2016 [4] given an approach of privacy-preserving clinical decision support system using nave bayes classification technique. This decision support system provides a new approach to privacy-preservation of adequate diagnosis and medications to patients diseases. This proposal stores patients sensitive informations in a current trends on cloud computing environment. A training model is built using the naive bayes classifier in a more secure manner. The trained model is applied across the new coming instances to predict the risk of diseases in an effective manner. It allows the patients to retrieve informations about top diseases based on their preferences.

Afef Mdhaffar et al [5] applied the complex event processing technology in predicting the heart failure. The data is collected from the wearable device in the patient. This approach obtains the great result with respect to precision and recall. Gaetano Valenza et al [6] utilized the Multifractal Point-Process in predicting the heart disease which attain the accuracy of 79.11%. The real time data is collected from the patient to evaluate this approach.

Ashir Javeed et al [7] introduced a novel system using the random search algorithm and random forest for feature selection and classification respectively. With the help of feature selection process most significant 7 features is extracted from Cleveland dataset and applied for prediction which achieved the great result with 93.33% accuracy.

Yuanyuan Pan et al [8] Enhanced Deep learning assisted Convolutional Neural Network (EDCNN) has been proposed to assist and improve patient prognostics of heart disease. Test results show that a flexible design and subsequent tuning of EDCNN hyperparameters can achieve a precision of up to 99.1 %.

Proposed Methodology

The present prediction model is framed with efficient feature selection algorithm and classification model: ECGSFS and HKNN respectively. The proposed system is shown in Fig.1 Feature selection is the most crucial step because the extracted features are huge in dimension. So, the number of the feature is minimized by applying feature selection process that successfully selects the heart disease relevant feature. Throughout this study, assumptions made for the classification approach, system workflow, algorithm descriptions, and implementation procedures are discussed in detail. The first step associated with the proposed system is the process of data collection. Since the proposed algorithm mainly focuses on building an efficient classifier. In second step, the most relevant features are selected with proposed feature selection algorithm.

Prediction process is done according to the classifier performance, across the training and the test datasets. The predicted outcomes are identified, and it is compared with various prior studies. The use of feature selection techniques improves the accuracy measures.

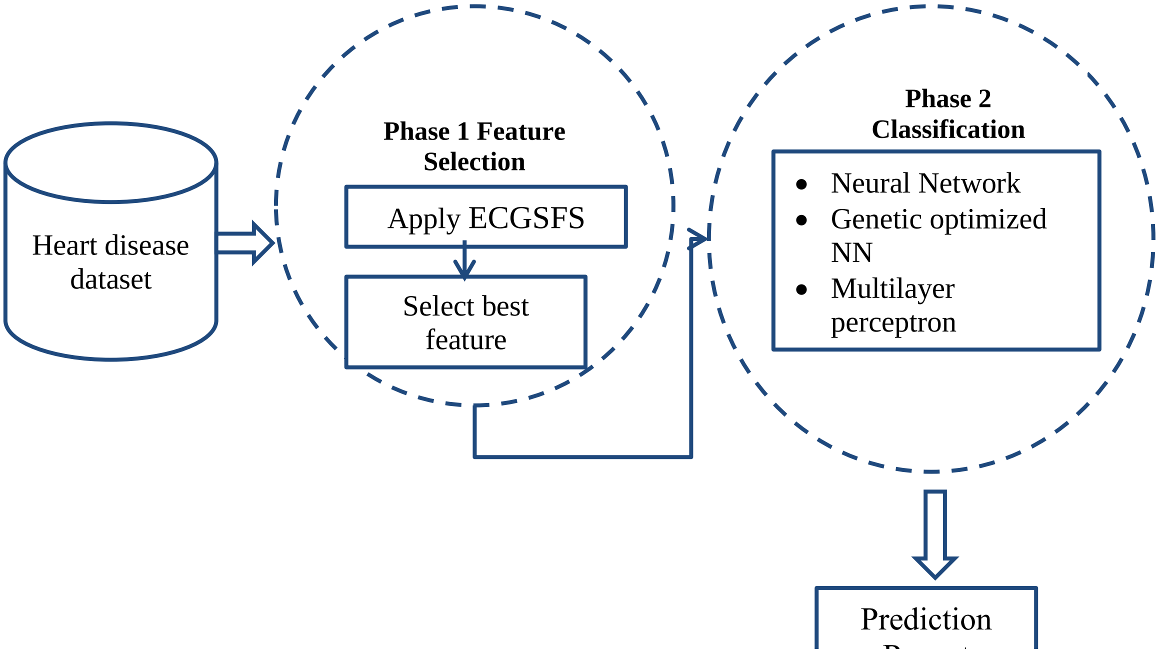

Figure 1. Proposed Heart disease prediction framework

Feature selection

In this work, ECGSFS is used to select the relevant heart disease features. Without analyzing the feature relation, correlation of the features is examined. By using the correlation feature, gravitational optimization technique is applied to get an optimal feature from features set. Due to the effective analysis of features, ECGSFS is used in this work for selecting heart disease features. First the extracted features are arranged in the feature space for analyzing the correlation between the features that helps to select optimized features. In addition to this, selected features reduce the training time, minimize data over fitting and eliminate curse dimensionality. Due to the importance of the feature selection,

At the time of this process, merit value of the features is computed for calculating feature weight. The feature merit value is estimated as

$$M_{S_k} = \frac{K_{\overline{r_{cf}}}}{\sqrt{\left(k + k(k-1)\overline{r_{ff}}\right)}} \qquad (1)$$

Where $\overline{r_{cf}}$ denoted as the mean value of features which are correlated for classification. $\overline{r_{ff}}$ is denoted as mean value of correlated features in the feature set. In addition to this merit value, feature set has been selected depending on the correlated criterion as

$$CFS = max_{sk}\left[\frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + \cdots r_{fifj} + \cdots + r_{fkf1})}}\right] \qquad (2)$$

In Eqn. 4.4, r_{cf1}, $rfifj$ are correlated variables, k is specific feature. According to equations 1 and 2, selection criteria, features are chosen from feature set. In addition to this selection criterion, gravitational search algorithm related objective function is used to select heart disease optimized features. The gravitational search algorithm was introduced in 2009 for solving optimization problem according to the concept of mass interaction and law of gravity. The algorithm consists of various agents which are interacting with each other based on gravity force. The interaction and performance of agents are computed using their mass value. The chosen features are gathered in gravitational search space that is defined as

$$Y_i = \left(y_i^1, \cdots y_i^d \cdots y_i^n\right), \qquad i = 1,2, \cdots N \qquad (3)$$

In Eqn. 4.5, y_i^d is the ith position of feature in d dimension value belongs to 0 and 1. From the feature position, mass value is estimated according to fitness value as

$$ma_i(t) = \frac{(fitnessi(t) - worst(t))}{(best(t) - worst(t))} \qquad (4)$$

Along with mass value (t), direction of features are estimated as

$$M_i(t) = \frac{(mai(t))}{\left(\sum_{j=1}^{N} mai(t)\right)} \qquad (5)$$

In Eqn. 5, $fitness_i(t)$ features fitness function at t time. By utilizing this, most noticeably awful and best features are characterized at *t* time as

$$best(t) = min\, fitnessj(t)j \in 1, \ldots N \qquad (6)$$

$$worst(t) = max\, fitnessj(t)j \in 1, \ldots N \qquad (7)$$

Using direction change of features, distance has to be computed as

$$F_ij^d = G(t)\frac{\left(mai(t) * maj(t)\right)}{\left(D_{i,j}(t)^n + \epsilon\right)} \qquad (8)$$

In Eqn. 8, F_ij^d is magnitude of features respect to feature mass of I and j in dth measurement. $G(t)$ gravitational power at t time. $M_i(t)$, $M_j(t)$ is represented as two features mass value. $G(t)$ is calculated as

$$G(t) = G(G0 * t) \tag{9}$$

In Eqn. 9, G0 is the first iteration gravitational value. Then local optimum trapping value should be minimized to compute best fitness value using

$$F_i^d = \sum (j \in k - bestj \neq i) randjFijd(t)) \tag{10}$$

In Eqn. 4.12$, rand_j$ is a arbitrary number somewhere in the range of 0 and 1. In view of the above procedure features subset is selected. The algorithm for ECGS algorithm is shown below.

Step 1: Collect quality improved heart disease features in feature space.

Step 2: Calculate merit value for each feature using Eqn. 1

Step 3: with the help of merit value, correlation features selection criteria is estimated as using Eqn. 2

Step 4: Arrange criteria based selected features in feature space,

Step 5: Calculate feature mass, direction change and fitness value to get an optimized feature

Step 6: Feature direction is calculated with fitness value

Step 7: from step 1 to step 6, heart disease features are analyzed and optimal features are chosen to classification process.

According to algorithm, heart disease features are analyzed, and then optimized features are selected that helps to reduce the feature subset effectively.

Classification

The classification is performed with neural network, genetic optimized neural network and multilayer perceptron which is discussed in this section.

Neural network

The network is just the variation of the radial basis neural network that was developed in 1991 by D.F. Specht. The network works according to the nonparametric regression process where the training samples are denoted as the mean of the radial neurons. The neural network computes the random function into the input and output. The structure of NN which comprises of three layers, for example, input, covered up and yield layer. The info layer gets the information from the feature selection phase, which is fed into the input and the computed output is processed by the output layer to get the output. Each layer performance is defined as

$$E[y|x] = \frac{\int_{-\infty}^{\infty} y \, * \, f(x,y).dy}{\int_{-\infty}^{\infty} f(x,y).dy} \qquad (11)$$

Where *y* denotes the estimator variable for the output, *x* represents the input, *E*[*y*|*x*] represents the expected value of output given the input vector *x*, *f*(*x*, *y*) represents the joint Probability Density Function (PDF) of *x* and *y*.

The computed value is passed to the output layer to get the output value which is defined as follows The estimator variable for the output is defined by,

$$y_i = \sum_{i=1}^{n} h_i w_i j \left(\sum_{i=1}^{n} h_i \right) \qquad (12)$$

Where w_{ij} represents the target yield relating to include preparing vector xi, $hi = e((1 - D2i)/(2 * \sigma 2))$ is the output of a hidden layer neuron, $D_i^2 = (x - ui)T\ (x - ui)$ is the squared distance between the input vector x and the training vector *u*, *x* is the input vector, *ui* is training vector, smoothing factor *σ* = *a* constant controlling the size of the receptive region. The constant value for the smoothing factor *σ* is varied from 0.10 to 1.

Genetic optimized neural network

John Hopfield is introduced neural network; it is form of recurrent neural network. The network helps to analyze false patterns from the set of features by covering the local minimum. The network structure consists of set of unit, normally represented in binary threshold unit that does not exceed the threshold value. The units are represented as 1 or -1. In general, the network has weight value between node i and j. The connection between nodes is formed as an undirected graph and the weight value have following restriction.

$$w_{ii} = 0 \forall i (itself\ connection\ does\ not\ have\ unit) \qquad (13)$$

$$wij = 1, \forall i, j (network\ have\ unit\ value, when\ it\ has\ symmetric\ connection \qquad (14)$$

The weight value, is used to compute estimated output for particular pattern recognition problem. During the output computation process, unit value is updated as follows,

$$S_i = \begin{cases} +1\ if\ \sum j w_{ij} s_j \geq 0 \\ -1 \qquad otherwise \end{cases} \qquad (15)$$

where w_{ij} is denoted as the connection between two node or units, *Si* is state of the unit and θ_i is threshold of unit i.

The incoming heart disease features are fed into the neural network in n-dimensional space in terms binary components -1 and 1. The binary components represent the output of the neural network. During the input process, each node has specific weights which is mentioned as

$$w = \frac{1}{n}\left(\sum_{i=1}^{D} \sigma_i^T \, \omega_i\right) \qquad (16)$$

Where D is class patterns, ω_i is the vector. According to the feature weight, network output is computed as

$$x(t + 1) = sign(W * x(t)) \qquad (17)$$

where, $x(t)$ is the input feature state time. This process is repeated continuously to get the output of every incoming patient features. At the time of the above classification procedure, network has error rate which completely reduces the efficiency of the heart disease prediction system and also maximizes miss-classification rate. So the performance of the system is enhanced by the optimizing network with a genetic algorithm. It is one of the evolutionary approaches which improve network efficiency in terms of using natural selection process. This algorithm is introduced by John Holland in 1960's based on the Darwin evolutionary theory. GA calculation uses the three administrators' determination, hybrid and transformation to optimize neural network structure.

Based on the working flow, genetic algorithm optimizes the discussed neural network. First network weight values are analyzed and selected based on the genetic fitness function. In this analysis, fuzzy membership function is used as the fitness function that is mentioned as

$$fitness\, function = U_i(x_i) \in 0, 1 \qquad (18)$$

Based on the fitness function, weight value is estimated and arranged in the terms of descending order.

The chosen value is applied to the mutation and crossover process to get an optimized value because it reduces miss-classification rate. In mutation process, one weight value is replaced with the other weight value that is different from previous solution. After performing mutation process, voting process is applied to perform a crossover operation.

This process estimates relation of the node weight value. Maximum voting values are treated as connectivity with the other value and remaining values are eliminated effectively. The chosen values help to classify whether the given heart disease features are normal or abnormal. According to the algorithm steps and discussion, heart disease features are classified with a minimum time and also reduces the misclassification rate. Then the efficiency of the system is evaluated using experimental results and discussions.

Multi-Layer Perceptron (MLP)

A multilayer perceptron (MLP) [10] is a class of the neural system which is a well-connected network of artificial neuron units. It pursues feed forward manner that maps input information onto relating yields. MLP comprises numerous layers, where layers are completely associated in the form of input, hidden and output. The neurons in MLP acts as a processing component with a activation function. MLP pursues standard linear perceptron to recognize information that are not straightly distinguishable.

Results and Discussion

Initially, the brilliance of the ECGSFS is based on component choice procedure. In this work, the size of data is minimized by ECGSFS method; in addition to this the effective examination of features properties, characters helps to identify the relevant heart disease features from the set of Cleveland heart disease dataset information. The successful selection of choosing criteria, and fitness function used to reduces the complexity while getting the exact heart disease relevant features. Along with this, the ECGSFS method chooses minimum number of features collated with the traditional algorithms. Then the efficiency of sample selected number of features is shown in table 1. The efficiency of ECGSFS method is collated with different algorithms such as Genetic Algorithm neural network, genetic optimized neural network and multilayer perceptron.

Table Feature selection Result

Feature count	Selected Feature
7	trestbps, chol, fbs, rest ecg, thalach, oldpeak

The Table 2 demonstrates that due to the effective analyze of genetic algorithm based weight selection, optimization process reduces the error rate and also satisfies expected output value. The result is shown in Fig. 2.

Table 2 Classifier Result

Models	**Accuracy (%)**	**precision**	**Recall**	**F-measure**
Neural Network	97.21	96.50	97.64	97.21
Genetic neural network	98.90	97.28	98.55	98.90
Multilayer perceptron	98.46	96.58	98.23	98.46

The Fig. 2 illustrates that performance measure of applied classifier. The presented GONN predicts strange heart disease features with the high effectiveness. The attained outcome is compared to the neural network, genetic optimized neural network and multilayer perceptron.

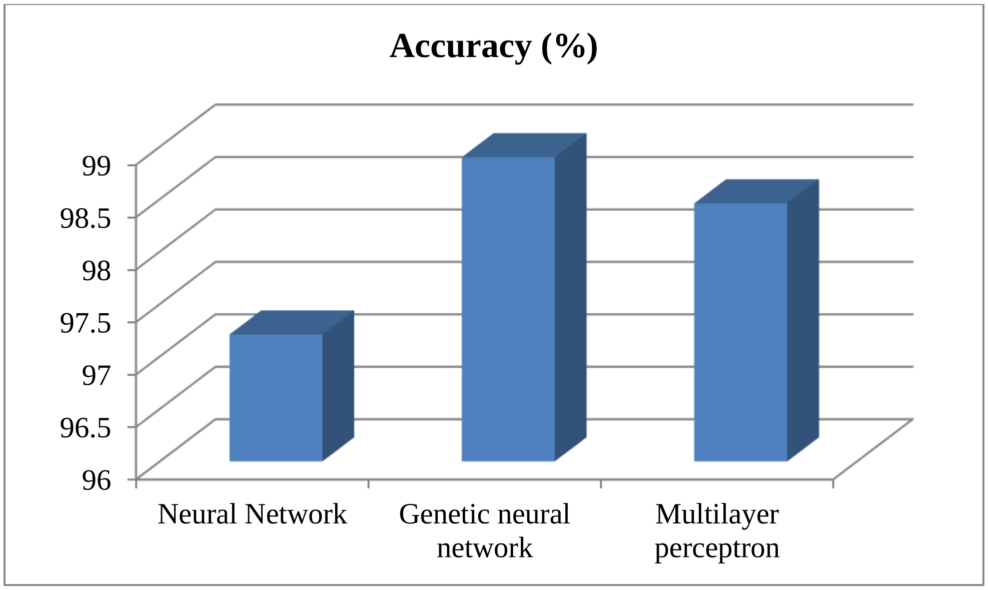

Figure 2. Accuracy of applied Classifiers

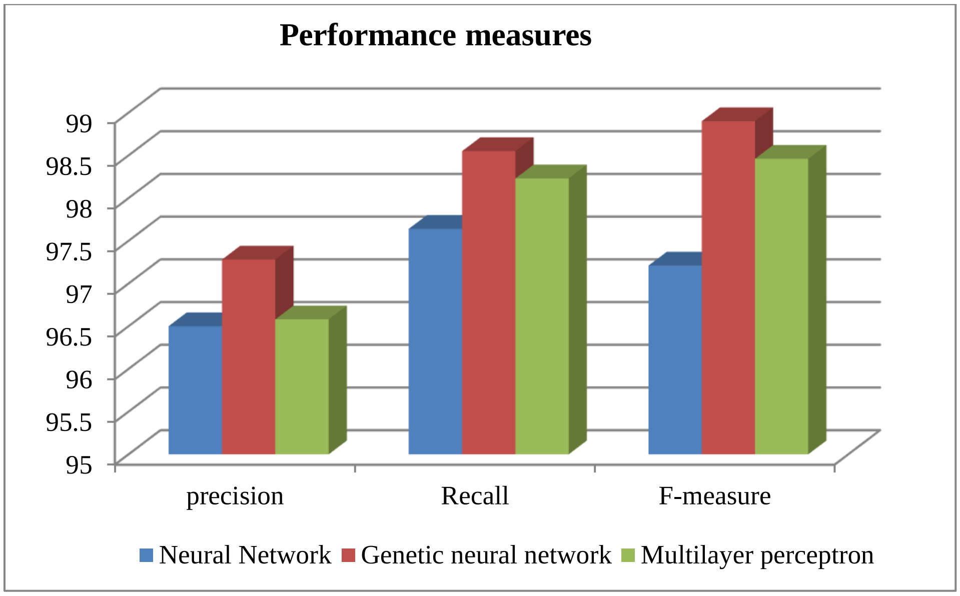

Figure 3. Precision, recall and F-measures of classifier

Figure 2 and 3 demonstrates that GONN method attains the high accuracy rate due to the successful usage of the selected features in every node of processing. In addition to this, the network has optimized weights, bias and learning rule concept while classifying output features. Excellence of GONN classification process is evaluated using precision, Recall and f-measure collated with other methods such as neural network and multilayer perceptron.

Conclusion

This study provides the need for machine learning techniques in health data mining and management processes. In particular, the research provides special attention to heart disease prediction and diagnosis measures. This due to the reason that heart diseases are one of the major concern across several countries and can even lead to death if not treated at an earlier stage. The present feature selection approach improves the accuracy of the classification measure. The experiments are conducted using netbeans with heart disease dataset from UCI repository. The efficiency of ECGSFS method is collated with different algorithms such as neural network, genetic optimized neural network and multilayer perceptron. This algorithm is mainly designed for heart disease prediction systems and in future, this can be extended to various other critical disease.

References:

[1]. Long, Nguyen Cong, M. H. (2015), ‘A highly accurate firefly based algorithm for heart disease prediction’, Journal of Expert Systems with Applications 42(21), 8221–8231.

[2]. Patil, J. (2016), ‘Heart disease prediction using machine learning and data mining technique’, International journal of computer science 7(1), 129–137.

[3]. Manimekalai, M. (2014), ‘Study of heart disease prediction using data mining’, International journal of advanced research in computer science and software engineering 4(1), 121–128.

[4]. Liu, Wang, M., Moran, A. E., Liu, J. and Coxson (2016), ‘Projected impact of salt restriction on prevention of cardiovascular disease in china: a modeling study’, Journal of plos one 11(2), 1–16.

[5]. Mdhaffar, I. Bouassida Rodriguez, K. Charfi, L. Abid and B. Freisleben, "CEP4HFP: Complex Event Processing for Heart Failure Prediction," in *IEEE Transactions on NanoBioscience*, vol. 16, no. 8, pp. 708-717, Dec. 2017.

[6]. G. Valenza et al., "Mortality Prediction in Severe Congestive Heart Failure Patients With Multifractal Point-Process Modeling of Heartbeat Dynamics," in IEEE Transactions on Biomedical Engineering, vol. 65, no. 10, pp. 2345-2354, Oct. 2018.

[7]. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in *IEEE Access*, vol. 7, pp. 180235-180243, 2019

[8]. Y. Pan, M. Fu, B. Cheng, X. Tao and J. Guo, "Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform," in *IEEE Access*, vol. 8, pp. 189503-189512, 2020.

[9]. Rashedi, E., Nezamabadi-pour, H. & Saryazdi, S. BGSA: binary gravitational search algorithm. *Nat Comput* **9,** 727–745 (2010).

[10]. Popescu, Marius-Constantin, Valentina E. Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. "Multilayer perceptron and neural networks." *WSEAS Transactions on Circuits and Systems* 8, no. 7 (2009): 579-588.