ISSN (ONLINE) : 2045-8711
ISSN (PRINT) : 2045-869X

INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY & CREATIVE ENGINEERING

May 2023
Vol - 13 No - 5
@IJITCE Publication

Design by Hemharsit

UK: Managing Editor
International Journal of Innovative Technology and Creative Engineering
1a park lane,
Cranford
London
TW59WA
UK

USA: Editor
International Journal of Innovative Technology and Creative Engineering
Dr. Arumugam
Department of Chemistry
University of Georgia
GA-30602, USA.

India: Editor
International Journal of Innovative Technology & Creative Engineering
36/4 12th Avenue,
1st cross St,
Vaigai Colony
Ashok Nagar
Chennai, India 600083

Email: editor@ijitce.co.uk

www.ijitce.co.uk

IJITCE PUBLICATION

International Journal of Innovative Technology & Creative Engineering

Vol.13 No.05

May 2023

www.ijitce.co.uk

Dear Researcher,

Greetings!

Articles in this issue discusses about study on data mining concepts in computer science.

We look forward many more new technologies in the next month.

Thanks,
Editorial Team
IJITCE

Editorial Members

Review Board Members

Contents

A NOVEL MORTALITY PREDICTION APPROACH IN CONGESTIVE HEART FAILURE PATIENTS USING RANDOM FOREST (RF) WITH INTENSITY WEIGHTED FIREFLY OPTIMIZATION (IWFFO)

Dr. C. Sowmiya
Department of Computer Science,
Government Arts and Sciences College, Aravakurichi, Karur (DT),Tamil Nadu, India
Dr. A. Divya
Department of Computer Science,
Government Arts and Sciences College, Aravakurichi, Karur (DT),Tamil Nadu, India

Abstract – Heart related disease is the significant cause for short life of humans. People in large population country depend on healthcare industry, that's why people need accurate test result in short time. In healthcare industry, very huge amount of information is formed in daily large. A Random Forest algorithm with IWFFO proposed for the heart disease prediction. The overall performance of the proposed method was compared with the prior Support Vector Machine (SVM) with Recursive Feature Elimination. Total number of 14 attributes from Cleveland heart disease dataset are selected. They are age, sex, cp, trestbps, chol, fbs, restecg,thalach, exang, oldpeak, slope, ca, thal, target. IWFFO selectscp, trestbps, chol, fbs, thalach, exang, old peak, ca, thal, target. IWFFO feature selection method applies to RF algorithm which achieved high accuracy 98.7% when compare to SVM algorithm.

Keywords: Data mining, Heart Disease, Classification, Feature Selection, Dataset.

1. INTRODUCTION

The most important human part of the body is heart which circulated the blood throughout the body. The prominent causes of death happened due to cardiac disease. In a short oftime span the mortality rate has spiked. Cardiovascular diseases refer to these heart associated diseases. These diseases are seen more in the developing rather than developed countries as referred and suggested by. Inaccurate diagnosis of the disease may cause fatalities and hence precision and safety in diagnosing the heart disease would be the prime factor in healthcare practice. Hence, the diagnosis of the disease is a multifarious problem. It should not be ridden by false presumptions and dubious side effects. There are various conditions that affect the circulatory system. Hence, heart disease alone cannot be specified as a single disease.

2. SYMPTOMS

Smoking, Obesity, Depression, Hyper tension, High blood cholesterol, Poor diet, Family history and Physical inactivity

A. Types of Heart Disease

- *Coronary heart disease*
- *Angina pectoris*
- *Congestive heart failure*
- *Cardiomypathy*
- *Congenital heart disease*
- *Arrhythmias*
- *Myocarditis*
- *Heart attack*
- *Heart cancer etc.*

For detecting cardiovascular problems into the patients, the prediction methods like SVM, decision tree and neural networks are compared based on the metrics. It was found that in predicting cardio vascular disease the SVM performed better than other methods suggested by [73]. for forecasting the existence of heart disease, performance comparison was performed with Recursive Feature Elimination (RFE) feature selection with SVM classifier, Random Forest (RF) and Intensity Weighted Firefly Optimization (IWFFO).

3. LITERATURE REVIEW

Murthy & Meenakshi (2014) [36] proposed Neuro-genetic model development to predict coronary heart diseases. The new presented research task work was feature subset selection by multi-objective genetic algorithm deprived of previous ANN accuracy on the basis of heart

disease predictor. The intended designed Neuro-Genetic model pattern had been authenticated by means of 303 patient data sets attained for various age groups. The projected investigation had displayed prior heart disease detection with greater experimenting accuracy of 89.58% via reduced feature subset, hence minimizing the difficulty.

Alshurafa et al. (2014) [37] examined consequences from a Remote Health Monitoring (RHM) arranged in a six month Women's Heart Health research of 90 patients, and adopted superior feature selection and machine learning algorithms. This approach detects the patients' key baseline related features and constructed efficient prediction patterns to assist in determining RHM results success. Consequences from the Women's Heart Health research revealed this heart disease threat, life quality, family history, stress factors, social support, and anxiety at baseline each assisted to predict patient RHM result success.

Gao et al. (2015) [38] introduced two novel features to predict cardiac arrest that are Approximate Entropy (ApEn) and Sample Entropy (SpEn). The dimensionality reduction concept, PCA, was adopted to overwhelm the dimensionality. The results proved that this prediction performance of appending ApEn and SpEn to the 24 parameters was enhanced considerably when evaluated with just 24 parameters. Dimensionality reduction had more positive impacts on enhancing the prediction consequences.

Kaya & Pehlivan (2015) [39] discussed Premature Ventricular Contractions (PVCs) heartbeats classification from ECG signals and especially, on assessment performance of particular features by Genetic.

Boshra Bahrami et al. [40]. The author discovered many approach to diagnosis the heart disease .The various classifier techniques are the Decision Tree, Naive Bayes(NB),J48 K Nearest Neighbours (KNN) and SMO are used to classify data set. The results are compared by the various Classification techniques by metrics such as the precision, specificity, area under ROC curve, accuracy, sensitivity3 and F-measure5. The results have been compared with the above algorithm and shows that J48 Decision tree is the best classifier for diagnosing the heart disease on the existing dataset.

4. MACHINE LEARNING APPROACH

Machine learning is an interdisciplinary study, which has focused on algorithm design and computers utilize these algorithm for learning purpose. Learning is nothing but learning from the feature dataset. Generally, machine learning techniques are designed and implemented in such way that, they have permitted the expert system to produce the solution for the diagnostic problem by using previous information.

The different learning methods available for the classification task are supervised classification, unsupervised classification and reinforcement learning. The important application of machine learning is data mining. Classification is one of the supervised learning processes and the classification tool is used to predict the target class.

A. CLASSIFICATION

Classification is a common decision making job of human activities. Classification issues occur when objects are to be designated into pre-specified groups or classes on the basis of the quantity of noted features relating to that object. Several industrial issues are realized as classification issues. For instance, stock market predictions, weather forecasts, bankruptcy predictions, speech recognitions, character recognitions and so on and so forth. The classification issues may be resolved both in a mathematical as well as non-linear manner.

B. FEATURE SELECTION

One of the significant preprocessing options is feature selection, which is mainly used in data mining applications. In medical field, information are becoming very huge that increase the difficulty in decision making. This, decrease the performance of the prediction result in disease forecasting process.

Feature selection plays an important role in solving the scalability problems and improves the overall classification accuracy through removing the irrelevant features. It is the effective tool to choose the relevant feature from the medical dataset. Therefore, reducing unwanted features in

the dataset is a significant process in data mining to achieve a better accuracy.

In this research, the dataset used for experimental setup was collected from UCI repository Cleveland database is selected for this research. 297 patients' data is taken this research and 14 attributes also included. The proposed system is developed in Java language. The Net Beans IDE is utilized for front end design. MYSQL is used for database access.

Table 1 Feature information and class

Attribute	Type of Data	Description
Age	Numeric	age in years
Sex	Numeric	sex (1 = male; 0 = female)
Cp	Numeric	chest pain type -Value 1: typical angina Value 2: atypical angina -Value 3: non-anginal pain Value 4: asymptomatic
Trestbps	Numeric	resting blood pressure (in mm Hg on admission to the hospital)
Chol	Numeric	serum cholestoral in mg/dl
Fbs	Numeric	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	Numeric	resting electrocardiographic results
thalach	Numeric	maximum heart rate achieved
exang	Numeric	exercise induced angina (1 = yes; 0 = no)
oldpeak	Numeric	ST depression induced by exercise relative to rest
slope	Numeric	the slope of the peak exercise ST segment -Value 1: up sloping - Value 2: flat - Value 3: down sloping
ca	Numeric	number of major vessels (0-3) colored by fluoroscopy
thal	Numeric	3 = normal; 6 = fixed defect; 7 = reversable defect
num	Numeric	Diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing

5. RECURSIVE FEATURE ELIMINATION (RFE)

In RFE process, entire dataset is split into various slots, and necessary protein is calculated for the given slot. The values which are duplicate are pruned first using RFE. Originally 2 slots are taken by dumping from one slot to the other. If any of the slots has more elements or values, then both the slot is equalized. Equation is used to measure side by side, the average for a given feature. The value that is nearestfrom the first slot to the mean average is known as the criticalgene for that slot.

In the present heart disease prediction system, the RFE calculates the $\|w\|^2$ value to found the each attribute contributions. The w denotes the weight vector. The $\|w\|^2$ is considered as the ranking variable where the higher value of $\|w\|2$ is considered as significant features by RFE.

$\|w\|2 = |\sum n\ \alpha\ y\ \emptyset(x)|$, j=1, 2...m

Where

$\emptyset(xij)$ - Occurrence of feature vector.

This feature selection method comes under backward procedure. This follows the iterative practice to remove the unwanted attributes in the dataset.

The working flow of the PFE is discussed below. Initially the model considered all the attributes/ features from the inputdataset. The ranking factor is computed for each attributes and selects the attributes with high rank which improves the performance of the model and the remaining attributes are considered as unwanted and removed. These processes are continuous to construct the final model that fit the final feature selection result. The procedure get closed when it reaches thestopping criterion

Support Vector Machines (SVM)

SVM are typically utilized in patterns recognition as wellas objects recognition originally, given sets of points that are part of one of two classes; linear SVMs discover hyperplanes leaving biggest possible fraction of points of the same classes on the same side, and performing maximization of distance of either class from hyperplanes SVMs as classification methods have yielded improved classification outcomes when compared with the other typically utilized pattern techniques like maximum likelihood and neural network classifiers. It is typically beneficial for classifying remotely sensed information.

SVM and other kernel methods were studied thus forming an optimized parameter bound by the number of training samples rather than the dimension of the feature space. Thus, kernel methods solve the dimensionality problem to a certain extent correlating with the samples; besides neuro imaging data put forth a high correlation between the feature needing a low dimensionality than the actual voxel and samples.

Moreover the advantages of feature selection outweigh the disadvantages.

The advantages are

a) It speeds up the testing process.
b) It makes interpretation easy.
c) It is simple to evaluate the results of smaller subsets sufficient to maintain classification accuracies.

6. PROPOSED METHODOLOGY

The proposed methodology for heart disease using Data Mining process is described with the architecture as shown in Figure 3.8. The first step is the selection of the best features is performed by Intensity Weighted Firefly Optimization (IWFFO) Then the classification process is performed to predict Congestive Heart Failure.

A. Intensity Weighted Firefly Optimization (IWFFO)

Feature selection is process of selecting useful feature from dataset. Firefly is a kind of a flash light which tries to communicate with the other members of their nature. As the intensity of light vanishes with respect to distant locations, its accuracy can be defined at local horizons for finding the best solution for any function. The fireflies are the particles or the extracted features from peak estimations. Each extracted feature (firefly) is assigned by light intensity and out of all the extracted features the distinct features which have common species are selected as the best one. This is best explained by the contours in which random regions are created based on the nature of features extracted and the particles of similar species are attracted towards the centre of the regions of the contours.

The random regions are created based on the feature categories and the particles of similar nature will follow their own regions. Out of all the particles, some are in the centre of the regions and these are defined as the best features for better classification of disease. Hence fire fly optimization will serve the purpose of feature reduction technique by considering similar natured particles and neglecting the others. Let us consider T_F as the feature vector or feature matrix. On selecting a training feature, T_r ' Define $\propto$, β and γ with some random values (here 0.2, l.0 and l.0 are considered respectively).

Let $X = X_i$ $(i = 1,2,3 \ldots\ldots n)$

Where 'n' is the number of particles and 'X' is the population of fire flies.

DefineI = rand (T_F)

Where I denotes light intensity. Updating the observation coefficient as

$$y_i = \sqrt{(x_i - x_j)^2 + (y_i - y_i)^2}$$

Where i= I,2, 3, ... n, j=I,2, 3,..... m. Final updates are expressed as

$$x_n = x_n(i) * (1 - \beta) + x_n(j) * \beta + \alpha(\text{rand} - 0.5)$$
$$y_n = y_n(i) * (1 - \beta) + y_n(j) * \beta + \alpha(\text{rand} - 0.5)$$

When the light intensity gets updated after some iteration, the final values are indicated as

$f_t = I$ (x, y) ⟶ Exact fitness value
$id = \min$ (f_t) ⟶ Exact best fitness value
$T_r = T_F$ (idx) ⟶ Selected best feature

7. RANDOM FORESTS

Classification is a common decision making job of human activities. Classification issues occur when objects are to be designated into pre-specified groups or classes on the basis of the quantity of noted features relating to that object. Several industrial issues are realized as classification issues. For instance, stock market predictions, weather forecasts, bankruptcy predictions, speech recognitions, character recognitions and so on and so forth. The classification issues may be resolved both in a mathematical as well as non-linear manner.

The Random Forest Tree (RFT) classifier is an ensemble classification approach and it is the type of the nearest neighbor classifier method. Breiman stated that, the RFT develops more number of trees based on the random selection of the variables. During the learning stage of the classifier, tree nodes are separated using a random subset of data features. The RFT classifier works based on bagging concept, in which each successive tree is generated independently using bootstrap sample of the data items and classification of data items is done, which is based on majority vote.

The drawback in traditional decision tree approach was overcomes in RF. The RF overcomes the problems faced in the prior decision tree approaches. In RF, the 10 fold cross validation is considered as the default value.

8. EXPERIMENTAL RESULT

The original Cleveland dataset with 303 instances is given as input to the SVM-RFE Algorithm. The results obtained have been tabulated in Figure 3.9. The dataset yields the improved weighted average values of Precision 90%, Recall 89%, F- measure 94% and achieved the highest accuracy of 94%.

TABLE II. TABLE EXPERIMENTAL RESULT OF PROPOSED METHOD RF-IWFFO

Algorithm	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
SVM-RFE	90.3%	89.4%	94.5%	94.5%
RF-IWFFO	93.2%	90.1%	98.7%	98.7%

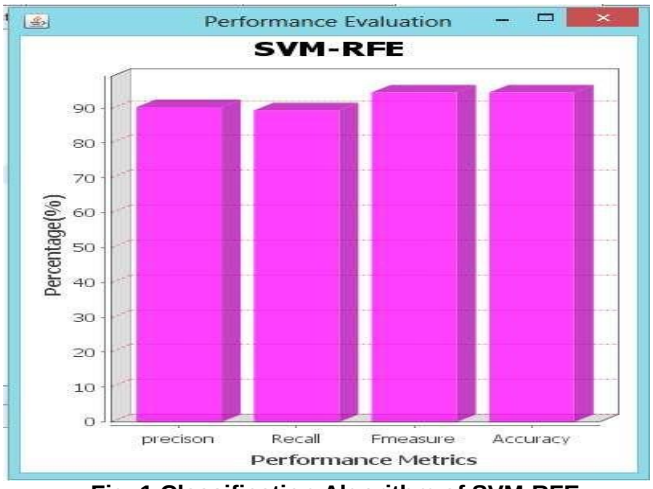

Fig. 1 Classification Algorithm of SVM-RFE

The original Cleveland dataset with 303 instances is given as input to the RF-IWFFO Algorithm. The results obtained have been tabulated in Figure 3.10. The dataset yields the improved weighted average values of Precision 93%, Recall 90%, F-measure 98% and achieved the highest accuracy of 98%. Compared to SVM-RFE classifier, RF-IWFFO classifier obtains the highest accuracy of 98.7% in heart disease prediction.

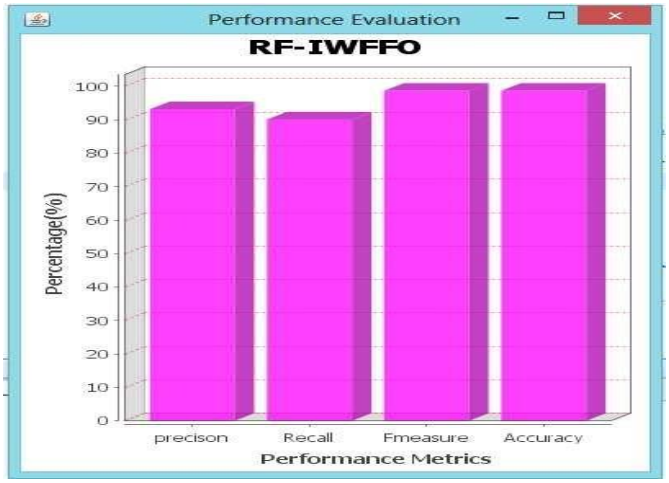

Fig. 2.Classification Algorithm of RF-IWFFO

9. CONCLUSION

Cardiac diseases pose a serious threat. When arteries thatsupply oxygen and blood to the heart are completely blockedor narrowed, cardiac issue happens. Although huge amount ofdata is produced by the healthcare organizations, the data is not appropriately utilized. Among others, in making business decisions, classification problems of allocating various observations into distinct group play a significant role. SVM- RFE algorithm gives 94.5% predictive accuracy, whereas Random forest-IWFFO algorithm gives 98.7% predictive accuracy. These results were obtained with less number ofiteration and it shows improvement from SVM-RFE paper.

REFERENCES

[1] Han, J & Kamber, M 2012, “Data Mining Concepts and Techniques”, Morgan Kaufman Publishers, USA.

[2] Sathyadevi, G 2011, “Application of CART algorithm in Hepatitis disease diagnosis”, IEEE International Conference on Recent Trends in Information Technology, India, pp.1283-1288.

[3] Krishnaiah, V, Narsimha, G & Chandra, NS 2013, “Diagnosis of LungCancer Prediction System Using Data Mining Classification Techniques”, International Journal of Computer Science and Information Technologies, Vol. 4, No. 1, pp. 39-45.

[4] Nasser, HS, Tharwat, AA & Moniem, NKA 2010, “Support vector machine for diagnosis cancer disease: A comparative study”, EgyptianInformatics Journal, Vol.11, No. 2, pp. 81-92.

[5] Navid, KG & Saheb, A 2015, "Combination of PSO Algorithm and Naive Bayesian Classification for Parkinson Disease Diagnosis", Advances in Computer Science: an International Journal, Vol. 4, No. 4, pp. 119-125.

[6] Chaves, R, Ramirz, J, Gorriz, JM & Puntonet, CG 2012, "Association rule-based feature selection method for Alzheimers disease diagnosis", Vol. 39, No. 14, pp. 11766-11774.

[7] Venkatalakshmi, B & Shivsankar, MV 2014, "Heart disease diagnosis using predictive data mining", 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14), Tamil Nadu, India, Vol. 3, No. 3, pp. 1873-1877.

[8] Cherian, V & Bindu, MS 2017,"Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique", Vol. 5, No. 2, pp. 68-73.

[9] Shinde, SB & AmritPriyadarshi 2015, "Diagnosis of heart disease using data mining technique", Vol. 4, No. 2, pp. 2301-2303.

[10] Palaniappan, S & Awang, R 2008,"Intelligent heart disease prediction system using data mining techniques", Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on IEEE, pp. 108-115.

[11] Shouman, M, Turner, T & Stocker, R 2012,"Using Data Mining techniques in heart disease diagnosis and treatment", Electronics Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on IEEE, pp. 173-177.

[12] Deekshatulu, BL & Chandra, P 2013,"Classification of heart disease using artificial neural network and feature subset selection", Global Journal of Computer Science and Technology, Vol. 13, No. 3, pp. 4- 14.

[13] Sonawane, JS & Patil, DR 2014, "Prediction of heart disease using learning vector quantization algorithm", IT in Business, Industry and Government (CSIBIG), Conference on IEEE, pp. 1-5.

[14] Tugce Kantar, Ovul Koseoglu, Aykut Erdamar, 2014,"Analysis of Heart Diseases from ECG signal", National Biomedical Engineering Meeting (BIYOMUT), 2014 18th National IEEE, pp. 1-4.

[15] Guyon and A. Elisseeff, 2003 "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, Vol. 3, No. 3, pp. 1157–1182.

[16] Sanz, Hector, Clarissa Valim, Esteban Vegas, Josep M. Oller, and Ferran Reverter, 2018, "SVM-RFE selection and visualization of the most relevant features through non-linear kernels", BMC bioinformatics 19, No. 1 pp. 1-18.

[17] Breiman L, Random Forests, Machine Learning, 45, pages 5-32, (2001)

[18] Fang K N, Jian-Bina W U, Zhu J P, et al, 2011, " A Review of Technologies on Random Forests", Statistics & Information Forum,

[19] Guo, Yang, Guohua Bai, and Yan Hu , 2012, "Using bayes network for prediction of type-2 diabetes", International Conference for Internet Technology and Secured Transactions, pp. 471-472.IEEE.

[20] Koklu, Murat, and Yavuz Unal, 2013, "Analysis of a population of diabetic patients databases with classifiers", International Journal of Biomedical and Biological Engineering Vol.7, No. 8, pp. 481-483.

[21] Leo Breiman, L., Jerome Friedman, J. H., Richard A Olshen, & Stone, C. J. (1984), "Classification and regression trees, Wadsworth & Brooks. Monterey, CRC press.

May 2023

Vol - 13 No - 5

@IJITCE Publication

Design by Hemaharsith