Modelling an Effectual Glowworm Swarm Optimization Strategy for Feature Selection in Heart Disease Prediction

R.Gomathi, R.Ramprashath, P.Murugeswari, A.Jeyachitra

Abstract*: Heart disease is measured as a common disease all over the world. The ultimate target is to provide heart disease diagnosis with improved feature selection with Glow worm swarm optimization algorithm. The anticipated model comprises of optimization approach for feature selection and classifier for predicting heart disease. This system framework comprises of three stages: 1) data processing, 2) feature selection using IGWSO approach and classification with conventional machine learning classifiers. Here, C4.5 classifier is considered for performing the function. The benchmark dataset that has been attained from UCI database was cast off for performing computation. Maximal classification accuracy has been achieved based on cross validation strategy. Outcomes depicts that performance of anticipated model is superior in contrary to other models. Simulation has been done with MATLAB environment. Metrics like accuracy, sensitivity, specificity, F-measure and recall has been evaluated.

Keywords : heart disease, Glowworm swarm optimization, C4.5 classifier, feature selection, cross validation

I. INTRODUCTION

Heart failure is a state where heart is not capable to provide essential blood to fulfill basic requirements [1]. Coronary artery is considered as an integral factor which is essential for providing blood to heart [2]. This disease (blocked or narrowed arteries) is more relevant kind of heart disease and most general HF cause.

There are numerous imperiling factors that may outcomes in HF disease. These factors may come under two factors with initial category comprising of imperiling or risk factors that will not be altered, for instance, Patients' age, sex and family history [3]. This secondary classification may be altered that comprises of attributed to way of patients' life. For example, smoking habit, high cholesterol level, higher BP level and physical activities [4]. As well, appropriate HF symptoms comprises of (breath shortage) dyspnea, (swollen feet) edema, weakness and fatigue.

By analyzing various factors, HF management turns to be extremely complex and worsen, specifically in some nations that lack suitable diagnostic medical experts and instruments [5]. Moreover, diverse tests suggested by health practitioners to recognize disease. Certain tests are done with ECG, angiography, nuclear scan and echo-cardiogram [6]. With all these tests, ECG is a non-invasive approach. However, it is extremely complex as it may cause to undiagnosed HF disease symptoms. There are some factors that cause angiography, and some sort of diagnosis utilized to validate heart disease instances [7]. It is measured as superior model for HF disease prediction. Moreover, certain crisis is related with it like high cost, side effects and higher level requirements of technical [8]l expertise. Therefore, alternative modalities are required which may resolve this crisis [9]. Henceforth, it is essential to model an intelligent, effectual, medical based decision system with DM principle and machine learning.

With background studies, diverse decision making support systems based on decision tree, SVM, k-NN, ANN, fuzzy logic based approaches and ANN ensemble model are recommended for heart disease prediction. Author in [10], depicted gathering heart related data for Cleveland based heart disease prediction utilized for logistical regression to analyze heart disease based risk assessment, attaining classification precision with 78%, Newton cheung validated feasibility of diverse classifiers like Naive bayes, C4.5, BNNF and BNND procedures and attained 82%, 83%, 82% and 81% respectively as heart based risk prediction accuracy. Author in [11] utilized decision based support system that uses AIS and offered 85% accuracy.

Author in [12] modeled a modified AIS and acquired a heart disease risk prediction with 88%. This model of neural network ensemble was anticipated by author in [13] to improve classification precision and attained a percentage of 90% for improving accuracy prediction. Author in [14], anticipated an ensemble based decision model that assists in preliminary ANN and Fuzzy HF and acquired 92% heart risk classification accuracy. Author in [15] anticipated machine learning based approaches by optimizing and stacking two SVMs along with enhanced heart risk prediction and attained by heart failure prediction performance of 93%. Author in [16], recently determined an adaptive weighted fuzzy model based ensemble approaches. Author's suggestion leads to 93% heart rate accuracy prediction.

Revised Manuscript Received on April 16, 2020.

R.Gomathi, Assistant Professor, Department of Master of Computer Applications of Karpagam College of Engineering.

R.Ramprashath, Assistant Professor, Karpagam College of Engineering, Coimbatore

P.Murugeswari M.Sc., Assistant Professor, G.T.N Arts College, Dindigul

A.Jeyachitra M.C.A.,M.Phil ., Assistant professor, G.T.N Arts College, Dindigul

Inspired from diverse decision making support system anticipated in earlier approach that has been discussed above, this work also tries to model a novel decision system for heart based risk prediction to enhance prediction based classification accuracy and diminish complexity or computational cost [17]-[18]. Decision making system modeled based on Glow worm swarm optimization approach for feature selection. This model considered statistical analysis to examine feature based selection [19]-[20]. To acquire an optimal amount of ranking features in this work, we exploit C4.5 classification model for these searching approach. Performance of every classifier and feature selection model will be computed with classifier model that is considered as machine learning based classifiers. It is most essential to discuss classifier model uses simpler prediction model however provides superior efficiency than more superior prediction model like ANN and ensemble classifiers. With no prior investigation, to the best of knowledge have addressed classifier model with other background model, experimental findings of recommended approach are promising in heart based prediction accuracy.

The remainder of this paper is constructed as trails: section 2 demonstrates background study and drawbacks related to it. Section 3 provides detailed explanation regarding anticipated model for validation and evaluation. Section 4 offers findings and discussions. Section 5 shows final part of this work.

II. RELATED WORKS

Author in [21], anticipated feature selection approach with Binary GWO. Two models are utilized in feature selection procedure. Ultimate target is to examine classification prediction and reduce number of chosen features. These are performed in 18 different datasets from UCI machine learning repository that comprises clinical data. Mean function parameters of 0.030 and 0.150 are attained for breast cancer datasets correspondingly are compared superior with other values attained with GA and particle swarm optimization.

Author in [22], anticipated a classification structure by merging benefits of extreme learning machine and fuzzy set. Datasets are changed to fuzzy sets using trapezoidal member functions. Classification is done with FFNN with single hidden layers with extreme learning machine. Experiments are done with Cleveland disease with five class labels, Cleveland heart disease with two class labels, PID and SHD from UCI machine learning repository and provided accuracy of 75%, 94%, 95% and 93% respectively.

Author in [23] offered Meta-heuristic model with Ant-lion optimizer for feature selection. This optimizer based variants are examined by deploying transfer functions. Every function was utilized to map constant search space. Tree based V-shaped and S-shaped transfer functions were utilized in this investigation for experimentation with UCI ML repository and in contrary with PSO and two diverse ALO variants based algorithms. Experimental outcomes demonstrate superior accuracy in contrary to prevailing approaches. For Wisconsin diagnostic breast cancer dataset, ALO algorithm with V-shaped transfer function acquired an accuracy of 98%, ALO with V-shaped transfer function carries out superior function with S-shaped transfer function by eliminating local optima.

Author in [24] have offered hybrid bio-inspired heuristic procedure that merges Grey Wolf optimization and Ant-Lion optimizer procedures. In hybridization, convergence is attained with global optimizers by eliminating local optima and forcing up search procedure. In hybridization model that works individually may outperforms Grey Wolf Optimizer and ALO which has been experimented with Heart dataset that belongs to clinical domain. GWO-ALO procedure has provided exploration of search space and optimal solution exploitation in balanced manner [25]. Average fisher values with 0.070 and 0.760 are attained for Cleveland and Wisconsin dataset respectively. This work uses parallel distribution mode was recommended to improve classifiers' convergence time. Author in [26], anticipated a Meta-heuristic approach termed as wind driven swarm optimization model for medical diagnosis.

Author in [27], anticipated a novel evaluation metrics that are measured based on both classifier accuracy and rule set size that was initiated for constructing classifier model. Efficiency with other conventional PSO procedure are determined to be more appropriate. Experiments are performed on clinical datasets acquired from UCI dataset and Liver Disorder dataset. For liver disorder dataset, anticipated approach may provide accuracy of 65% and heart disease dataset that yields 78% accuracy.

Author in [28] anticipated feature selection approach with GA for detecting cancer diagnosis. Here, wrapper based approach with GA based feature selection is proposed. In case of classification, PS, ANN and GA classifier were utilized. The idea behind this was evaluated based on WDBC, WBC and WPBC dataset [29]. Outcomes from these experimentation demonstrate that anticipated feature selection may enhance classifier accuracy were these outcomes are compared with WDBC, WBC and WPBC with accuracy of about 96%, 96% and 78% correspondingly with GA classifier. PC classifier is utilized for attaining an accuracy of 96%, 97% and 97% correspondingly [30]. Accuracy of these datasets that are compared with ANN classifier was utilized as 97%, 97% and 80% correspondingly. Accuracy of anticipated approach was superior in contrary to prevailing related approaches.

III. METHODOLOGY

The anticipated system model comprises of three diverse sub-systems termed as pre-processing, feature selection and classification model. This pre-processing model may comprise of missing values interpretation and normalization phase. Feature selection is performed with Glow worm swarm optimization approach that chooses an optimal feature subset to construct classifier model. Here, feature selection model uses Meta-heuristic optimization model termed as Glowworm swarm optimization with an increased accuracy using fitness function. Classification is done with C4.5 classifier for training and testing system. Here, system model is provided in Fig. 2.

A. Dataset description

This framework has been tested with benchmark dataset that are attained with UCI machine learning repository termed as Heart disease dataset. This dataset may comprise of 155 instances with two diverse classes with 19 features and 168 missing values. Attributes are provided in table I that is given below, i.e. live/die were replaced as 'non-affected' and 'affected' respectively. Heart disease based dataset consists of diverse instances with various features and two class labels. Some missing values are in dataset. It is explained in Table given below:

B. Data Pre-processing

Noisy or missing values in heart disease dataset may influence classifier performance. The anticipated model utilizes heart disease based dataset for performing experimentation among all the available UCI machine learning repository comprises of 167 missing values where these dataset are free from noisy or missing values. Imputation is utilized for missing values. This imputation may handles by filling those values with same data points from features. Data is based on similar record and missing value is fulfilled with values provided in those records. As average missing values in heart disease is lesser than 35% missing values imputed from records that may not have missing value.

$$V' = \frac{v - \min_A}{\max_A - \min_A} (new\_\max_A - new\_\min A) + new\_\min A \quad (1)$$

Where $V'V'$ is needed for normalization value, vv is present variable value, $\max_A \; and \; \min_A \;\; \max_A \; and \; \min_A$ are maximal and minimal values of present range respectively and $new_{\max_A} \; and \; new_{\min_A} \;\; new_{\max_A} \; and \; new_{\min_A}$ are maximal and minimal values of normalization range specifically.

C. Feature Selection

Generally, pre-processed clinical data is provided to feature selection. This sub-system utilizes wrapper method utilizing three bio-inspired approaches termed as Gloworm swarm optimization with C4.5 classifier as classifier. Every bio-inspired procedure chooses feature sub-set providing three feature subsets. This feature selection is carried out to choose optimal features from three diverse feature sub-sets that may be diminishing feature set acquired from this selector to subject classification by providing superior classification accuracy.

D. Improved Glowworm swarm optimization (IGWSO)

This work is proposed by Ghose and Krishnanand is bio-inspired process based on collective characteristics of GW. Here, GWO chooses feature subset. Accuracy is measured as a fitness function. Steps included in this work are provided below:

Step 1: population is produced randomly with search space as every glowworm has $'n''n'$ number of features. All GWO features may consider value 1 or 0. If features are chosen, then it is specified as 0 else 1. Initially, every GW has equal luciferin level. Constant parameters are provided in table below.

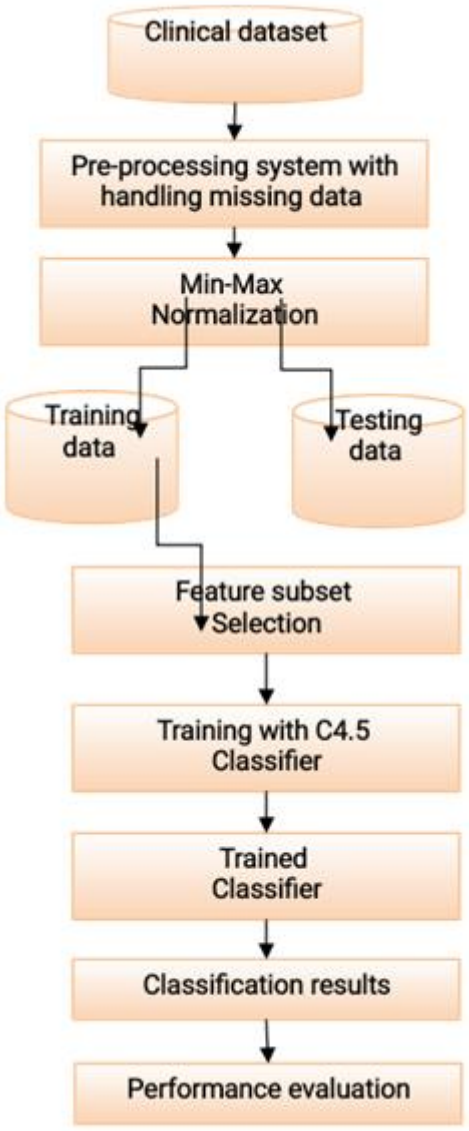

Fig. 1. Flow diagram of proposed model

Step 2: Luciferin based on fitness function at every GW position. Accuracy of classifier is based on fitness function. Every GW during updates will be added to prior luciferin level as in Eq. (2):

$$l_i(t+1) = (1-\rho)\, l_i(t) + \gamma(x_i(t+1)) \quad (2)$$

Where $l_i(t)l_i(t)$ is luciferin level related to GW $'i''i'$ at time $t, \rho t, \rho$ is luciferin decay constant, $\gamma\gamma$ is luciferion constant improvement, $J\big(x_i(t+1)\big)J\big(x_i(t+1)\big)$ fitness function value of $i^{th}i^{th}$ GW at time $'t'.'t'.$

3) Every $i^{th}i^{th}$ GW that determines move towards lighter GW that has finest luciferin value. It chooses superior GW with probabilistic method as in following Eq. (2):

$$P_{ij}(t) = \frac{I_j(t) - l_i(t)}{\sum_{k\varepsilon\, n_1(t)} I_k(t) - l_i(t)} \quad (2)$$

Where $j\,\varepsilon\, N_i(t), N_i(t) = \{j: d_{ij}(t) < r_d^i(t); l_i(t) < l_j(t)\}$

$j \varepsilon N_i(t), N_i(t) = \{j: d_{ij}(t) < r_d^i(t); l_i(t) < l_j(t)\}$ is neighbourhood set of GW at time $'t','t'$, $d_{ij}(t) d_{ij}(t)$ specifies Euclidean distance among GW $'i''i'$ and $'j''j'$ at time $'t''t'$ and $r_d^i(t) r_d^i(t)$ is variable neighbourhood range related to GW $'i''i'$ at time $'t'.'t'$.

Step 4: GW movement is provided as in Eq. (3):

$$x_i(t+1) = x_i(t) + s\left(\frac{x_j(t) - x_i(t)}{||x_j(t) - x_i(t)}\right) \quad (3)$$

Where $x_i(t) x_i(t)$ is GW location $'i''i'$ at time $'t','t'$, $||x_j(t) - x_i(t)|||x_j(t) - x_i(t)||$ is Euclidean distance among GW $'i''i'$ and $'j''j'$ where $'s''s'$ is step size.

$$r_d^i(t+1) = min\{r_s, \max\{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\} \quad (4)$$

Where $r_o r_o$ is initial neighbourhood GW range, $\beta\,\beta$ is constant parameter, $n_t n_t$ is parameter utilized to manage number of neighbourhoods.

Step 5: repeat 2, 3 and 4 for maximal of 100 iterations. GWO may set maximal luciferin is considered as feature set of GWO.

A. Classification system

In classification system, dataset is partitioned into training sets and testing sets. DT is measured as a non-parametric learning approach that does not require searching for optimal factors in training and therefore it is utilized as weak learning classification. Here, classifier utilizes C4.5 as baseline model. This work uses booting approach to model these classifiers. CV is utilized to enhance amount of data for testing outcomes. Optimal outcome may attains superior classification accuracy.

IV. NUMERICAL RESULTS AND DISCUSSIONS

The anticipated model works on heart disease dataset that has been executed with MATLAB environment. Feature significance of this approach is computed with information gain. The anticipated model chooses appropriate attributes with feature selection based bio-inspired termed as Glowowrm swarm optimization that maintains classification accuracy and based on classifier performance. It will not depend on statistical class values by separability measure. Feature selection acts as a significant role in medical application for effectual classification. These feature selection may provides the extraction process with appropriate information from raw data to enhance classification performance. It provides a clear objective of data processing and data visualization to enhance prediction performance.

Table 1: Heart based features from dataset

Feature	Description	Domain
Age	-	29-77
Sex	Male, Femalee	0,1
Chest pain	Angine	1,2,3,4
BP	Asymptomatic	94-200
Serum cholesterol	Abnormal	126-564
Fasting blood pressure	-	0,1
Electrocardiograhic results	-	0,1,2
Maximum heart rate	-	71-202
Exercise induced angina	Norm, abnormal, hyper	0,1,2
Old peak=depression induced by exercise	-	0-6.2
Slope of peak exercise	-	1,2,3
Major vessels	-	0,1,2,3,4
Thal	Defect, reversible defect	3,6,7

Precision, accuracy, sensitivity, specificity are utilized to compute performance of classifiers which is provided in Eq. (5), Eq. (6) :

$$Acc = \frac{samples\ classified\ appropriately}{total\ classified\ samples} \quad (5)$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{samples\ classified\ appropriately\ as\ positive}{total\ samples\ positively\ classified} \quad (6)$$

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{samples\ classified\ appropriately\ as\ positive}{total\ positive\ samples\ in\ dataset} \quad (7)$$

$$Sensitivity = \frac{TN}{TN + FP}$$

Where $TP,TN,FPTP,TN,FP$ and $FNFN$ are attained from confusion matrix.

Table 2: Statistical analysis

Parameter	Value
ρ	0.5
γ	0.7
β	0.09
n_t	6
s	0.02
l_o	6

Table 3: Feature weight

Feature	Weight
C_2	0.17
C_{13}	0.14
C_7	0.12
C_{12}	0.122
C_9	0.10
C_3	0.09
C_{11}	0.05

Table 4: Performance metrics

S. No	Measure	Heart disease dataset (%)
1	Accuracy	98%
2	Precision	98%
3	Sensitivity	99%
98%4	Specificity	98%

Table 5: Confusion matrix

	Patients with disease	Healthy person prediction
Actual prediction with heart disease	TP	FN
Actual prediction	FP	TN

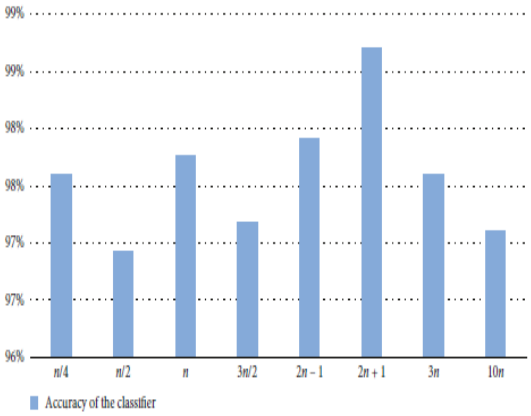

Fig. 2. Accuracy computation

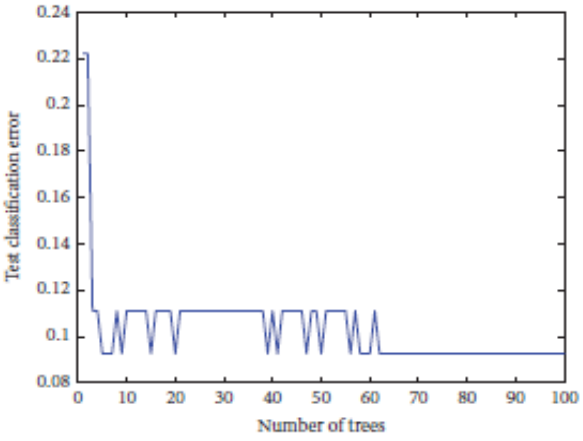

Fig. 3. Classification error

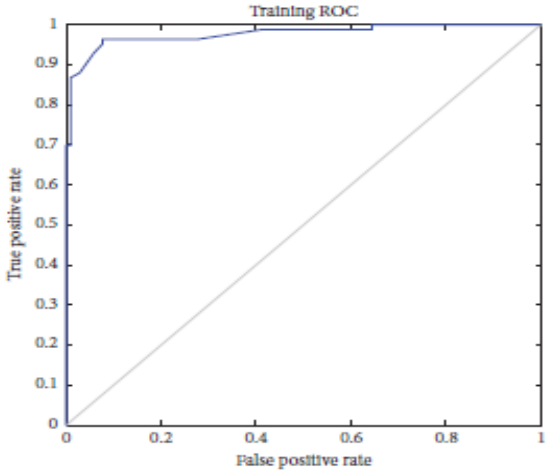

Fig. 4. ROC computation based on training

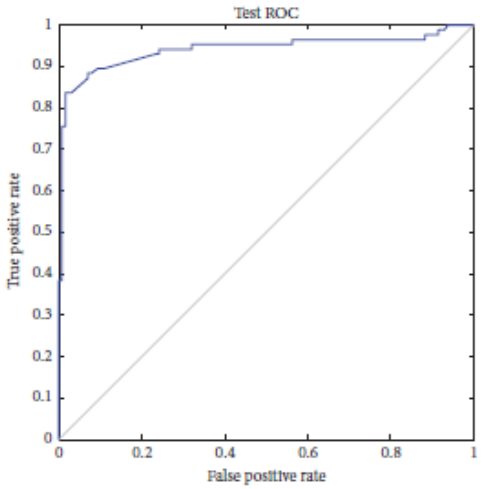

Fig. 5. ROC computation based on testing

Table 1 provides attributes that are related to heart disease, Table 2 provides statistical analysis and Table 3 explains about feature weight, Table 4 explains about performance metrics with accuracy, sensitivity, specificity and F-measure. Table 5 provides confusion matrix for predicting heart disease. Fig. 2 depicts accuracy of anticipated model, Fig. 3 provides classification error while Fig. 4 and Fig. 5 explains about ROC computation attained during training and testing.

V. CONCLUSION

Here, a novel feature selection and classification is anticipated for heart disease prediction. The significant novelty of this work relies over this anticipated approach; combination of feature selection to categorize heart disease in effectual and superior manner. C4.5 classification system comprises of two factors: GWO and classification sub-system. UCI ML dataset was chosen to test this system. Experimental outcomes demonstrate that reduced features may attain superior classification accuracy using C4.5 decision tree. Results are demonstrated that IGWO approach has superior performance with ACC, specificity and sensitivity. As well, performance of anticipated model is superior to prevailing models.

Based on empirical analysis, outcomes may specify that classification system is utilized as a tool for making decision for heart based disease prediction.

REFERENCES

1. D. Tomar, "Feature selection based least square twin support vector machine for diagnosis of heart disease," iNT. j *of Bio-Science and Bio-Technology*, 2014.
2. Robnik, "Theoretical and empirical analysis of ReliefF and RReliefF,"*Machine Learning*, 2003.
3. M. Buscema, "Training with Input Selection and Testing (TWIST) algorithm: a significant advance in pattern recognition performance of machine learning," j. *of Intelligent Learning Systems and Applications*, 2013.
4. K. Srinivas, "Applications of data mining techniques in healthcare and prediction of heart attacks,"Int J *on Computer Science and Engineering*, 2010
5. T. Helmy, "Multi-category bioinformatics dataset classification using extreme learning machine," Proc. *of IEEE Congress on Evolutionary Computation*, May 2009
6. G¨unes¸, "Effect of feature-type in selecting distance measure for an artificial immune system as a pattern recognizer," *Digital Signal Processing*, 2008.
7. S¸ahan, "The medical applications of attribute weighted artificial immune system: diagnosis of heart and diabetes diseases," *Artificial Immune Systems*, 2005.
8. F. Nie, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," in *Advances in Neural Information Processing Systems*, MIT Press, 2010.
9. Z Zhao, "Efficient spectral feature selection with minimum redundancy," in *Proc of 24th AAAI Conference on Artificial Intelligence*, 2010.
10. Randa,"Feature analysis of coronary artery heart disease data sets",Procedia Comp sci,Elsevier, 2015
11. Jabbar, "Computaional intelligence technique for early diagnosis of heart disease' IEEE, 2015
12. PK Anooj," Clinical decision support system: Risk level prediction of heart disease using Weighted fuzzy rules", J. of king saud university, 2012
13. Mai Shouman, " Using decision tree for diagnosing heart disease patients", ACM , 2011.
14. Tu," Effective diagnosis of heart disease through bagging approach" Biomedical Engineering and approach, IEEE 2009.
15. Liao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, 2012.
16. Ngai, "The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature," Decision Support Systems, 2011.
17. Y. Li, "A distributed ensemble approach for mining healthcare data under privacy constraints," Information Sciences, 2016.
18. A. Malik, "Extreme learning machine based approach for diagnosis and analysis of breast cancer," J. of Chinese Institute of Engineers, 2016.
19. M. Hariharan, "A new hybrid intelligent system for accurate detection of Parkinson's disease," Computer Methods and Programs in Biomedicine, 2014.
20. I. Castelli "Combination of supervised and unsupervised learning for training the activation functions of neural networks," Letters, 2014
21. I. Castelli, "Combination of supervised and unsupervised learning for training the activation functions of neural networks," Letters, 2014
22. S. Bashir, "HMV: a medical decision support framework using multi-layer classifiers for disease prediction," J. of Computational Science, 2016.
23. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," Expert Systems with Applications, 2009.
24. R. Chitra, "Heart disease prediction system using supervised learning classifier," Int. J. of Software Engineering and Soft Computing, 2013.
25. Masethe, "Prediction of heart disease using classification algorithms," in Proc of World Congress on Engg and Computer Science, 2014.
26. Masethe, "Prediction of heart disease using classification algorithms," in Proc of World Congress on Engg and Computer Science, 2014.
27. Çalişir, "A new intelligent hepatitis diagnosis system: PCA–LSSVM," Expert Systems with Applications, 2011.
28. Taneja, "Heart disease prediction system using data mining techniques," Oriental J. of Computer Science and Technology, 2013.
29. Shouman, "Using data mining techniques in heart disease diagnosis and treatment," Japan-Egypt Conf on Electronics, Communications and Computers, 2012.
30. Yan, "Development of a decision support system for heart disease diagnosis using multilayer perceptron," in Proc of the Int S. on Circuits and Systems, 2003.

AUTHORS PROFILE

R.Gomathi is an Assistant Professor in Department of Master of Computer Applications of Karpagam College of Engineering. Working from December 2013 in Karpagam College of Engineering College situated in Coimbatore. During the mentioned tenure handled classes on C, C++, Java programming, Python Programming, Database Management Systems, Operating Systems and Unix & Shell Programming. Also NPTEL certified candidate in Database management System, Operating Systems and Python for Data Science and Oracle certified candidate. Handled facilitation skills for prefinal year engineering under graduate students in collaboration with Infosys Soft Skills. Couple of times silver awarded partner faculty under Inspire in Infosys Campus Connect Faculty Partnership Model in the year 2017-2018 and 2018 - 2019. Also published a paper in national conference and three papers in international journals.

R.Ramprashath MCA.,MPhil.,MBA(HR) Currently working in Karpagam College of Engineering, Coimbatore for the past 5 Years 2 Months.In the UGC NET June 2019 (Computer Science and Applications) qualified for Assistant Professor. Now training students in Oracle, MySQL, Operating systems in the placement and Training division of the college. Worked as an Assistant Professor in Kalasalingam University, Krishnankovil with the experience of three years. Then experienced as Software analyst and as consultant for1 year,3 months. I have experience as Operating Systems/Database Technologies Team Lead of Training academy, Karpagam College of Engineering for the past 2 years. Currently i am handling students training in Operating Systems,DBMS Technologies MYSQL and Oracle for product companies. Pl SQL Programming Trained under Oracle University Training programme. Published four research papers in International Journal. Silver awarded partner faculty under Inspire in Infosys Campus Connect Faculty Partnership Model in the year 2018 - 2019. Also NPTEL certified candidate in Database

P.Murugeswari M.Sc., M.Phil., currently working in G.T.N Arts College, Dindigul for the past 9 years 10 Months. In the TNSET March 2018(Computer Science & Applications) qualified for Assistant Professor. Published a research paper in International Journal.

A.Jeyachitra M.C.A.,M.Phil .,SET., Currently working as Assistant professor for the past 9 years. Qualified TNSET in the year 2016 .Published three journals research paper. Appointed as nptel and swayam coordinator. One of the editorial board member in department magazine in the 2017 Awarded NPTEL elite certificate by Successfully completed database management system course. Nominated as IQAC member of my department and working for it. Received a Certified Microsoft innovative educator certificate.