Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.6, No.1, June 2018, pp. 229~240
Available online at: http://pen.ius.edu.ba

Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network

B Sankara Babu[1], A Suneetha[2], G Charles Babu[3], Y.Jeevan Nagendra Kumar[4], G Karuna[5]
[1,5]Departement of Computer Science and Engineering, GRIET, JNTU Hyderabad, Telangana, India
[2] Departement of Computer Science and Engineering, KKR & KSRITS, JNTUK, AP, India
[3] Departement of Computer Science and Engineering, Mallareddy Engineering College, JNTU Hyderabad, Telangana, India
[4]Departement of Information Technology, GRIET, JNTU Hyderabad, Telangana, India
bsankarababu81@gmail.com

Article Info

Article history:

Received Dec 12th, 2017
Revised Mar 20th, 2018
Accepted May 26th, 2018

Keyword:

Auto Encoder
Grey Wolf Optimization
Neural Network
Recurrent Neural Network

ABSTRACT

Big data development in biomedical and medical service networks provides a research on medical data benefits, early ailment detection, patient care and network administrations.e-Health applications are particularly important for the patients who are unfit to see a specialist or any health expert. The objective is to encourage clinicians and families to predict disease using Machine Learning (ML) procedures. In addition, diverse regions show important qualities of certain provincial ailments, which may hinder the forecast of disease outbreaks. The objective of this work is to predict the different kinds of diseases using Grey Wolf optimization and auto encoder based Recurrent Neural Network (GWO+RNN). The features are selected using GWO and the diseases are predicted by using RNN method. Initially the GWO algorithm avoids the irrelevant and redundant attributes significantly, after the features are forwarded to the RNN classifier. The experimental result proved that the performance of GWO+RNN algorithm achieved better than existing method like Group Search Optimizer and Fuzzy Min-Max Neural Network (GFMMNN) approach. The GWO-RNN method used the medical UCI database based on various datasets such as Hungarian, Cleveland, PID, mammographic masses, Switzerland and performance was measured with the help of efficient metrics like accuracy, sensitivity and specificity. The proposed GWO+RNN method achieved 16.82% of improved prediction accuracy for Cleveland dataset.

Corresponding Author:

Dr. B. Sankara Babu,
Departement of Computer Science and Engineering,
GRIET, JNTU Hyderabad,
Telangana, India 500090, Taiwan, ROC.
Email: bsankarababu81@gmail.com

1. Introduction

The important aspect of e-Health is to predict the disease dynamics during an epidemic, which is useful to allocate resources and make a quick response in a public health event [1]. In such conditions, present networks are responsible for health risk and danger fluctuation with financial and statistical conditions [2]. The utilization of medical datasets has helped the analysts to predict the disease around the world. The use of detected information from medical repositories has been recognized by the World Health Organization (WHO) because it helps to find therapeutic information and prediction.The ML methods are used for prediction to identify the hidden patterns. The ML based prediction methods are classified into three groups such as supervised, unsupervised and semi-regulated learning systems [3]. Supervised machine learning

strategies are the arrangement of nominal, categorical, or persistent features (regularly a combination of every one of them) to their related result which can be present in any of those structures. The greatest achievement of machine learning over linear models is their capacity to learn relationship from training information and sum it up to testing of inconspicuous data and furthermore to defeat non-linearity and associations between features. In any case, this capacity should be deliberately figured out how to maintain a strategic distance from over-fitting. In machine learning, the design parameters cannot get any benefits from the models, these are considered as hyper parameters, for instance the quantity of concealed layers in neural systems. Hyper parameters must be improved by cross-approval or framework pursuit to make a harmony amongst variance and bias in expectation, known as the variance-bias exchange off [4], [5].

An existing ML based disease risk prediction system includes several methods suchas Logistic Regression (LR), Convolutional Neural Network (CNN),Support Vector Machine (SVM) and etc.[6], [7]. In testing process, patient's data are classified into the group of either normal or abnormal. Moreover,these plans have often accompanied with less attributes and imperfections. The information collection is regularly small, for patients and infections with particular conditions [8], the attributes are chosen through involvement. In any case, these pre-chosen quality features are not possibly fulfilling the changes in the disease and its influence factors. With the advancement of big data analytics innovation, more consideration has been paid to ailment forecast from the point of huge data investigation, different explores have been directed by choosing the attributes from an extensive number of information to enhance the risk classification accuracy [9], [10], instead of the other qualities.Better frameworks created by machine learning methods can be utilized to help doctors in diagnosing and forecasting diseases. To predict the disease dynamics, a few examinations have been directed to create techniques for their classification. Moreover, there is a huge difference between diseases in various regions, because of variations in the atmosphere and living propensities in the region. In this manner, risk classification based on big data analysis, makes a superior model by considering missing information for forecasting the disease dynamic. In this paper, an efficient medical disease prediction model named as GWO+RNNis proposed. The GWO algorithm is used for feature selection which removes the unrelated attributes and redundant attributes. It's majorly improves the performance of prediction. After feature selection, Auto Encoder (AE) based RNN method avoids the feature dimensionality problems. Also, predict the different kinds of diseases significantly using the UCI database.

The organization of the paper is as follows. Section 2 gives the description of the models analyzed by various researchers related to this study.Section 3 provides a description of the development of the proposed GWO+RNN methodology used for predicting the disease. Sections 4 present the results obtained by various experiments and the conclusions are made in Section 5.

2. Literature Review

Numerous methods have been proposed by researchers on the prediction of diseases. In this section, a brief review of some important contributions ofsome methodologies in the field of disease prediction is presented below.

Y. Chen, *et al.,* [11] introduced the Realistic Contact Networks (RCNs) to describe the disease progression in e-Health applications. The structure of such system powerfully changed during plague. Catching such sort of powerful structure was the primary importance of forecast. With the ubiquity of cell phones, it was conceivable to catch the dynamic difference in the system structure. The investigation assessed the effect of system structure on disease progression, by evaluating enormous spatiotemporal information gathered by cell phones. Depend on the consequences of this assessment, a model was intended to perceive the dynamic structure of RCNs. This paper implemented a prediction algorithm for disease elements that depends on the spatiotemporal information. The precision of disease expectations was evaluated by various experiment and the results stated that the algorithm provided better precision values. This strategy is not able to handle more number of information for forecasting the disease elements.

M. Chen, *et al.,* [12]presented an efficient disease risk forecasting system using CNN based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm. The proposed method used latent factor model for avoid the medical information loss. The investigations were carried on a territorial chronic disease of cerebral localized necrosis by using this technique. Here, no existing methods worked on the both structured and unstructured medical big data analytics, but the CNN-MDRP strategy focused on the both medical big

data analysis. The experimental outcomes demonstrated that the prediction accuracy of CNN-MDRP achieved 94.8% with a convergence speed which is faster than a few normal forecast algorithms. The technique is unable to handle the advanced features, such as fractal dimension, biorthogonal wavelet transforms to predict the chronic disease.

R. Prashanth, and S. D. Roy, [13] proposed a novel and enhanced method for Parkinson's Disease (PD) utilizing the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) features. The Adaboost Algorithm (AA) was efficiently predict the PD in standard medical UCI dataset. An experimental analysis demonstrated that AA achieved 97.46% of accuracy in disease prediction. In this literature, the proposed AA was difficult to detect the normal, early stage of PD, and propelled PD due to inaccessibility of information. Additionally had a limited number of direct stage PD tests. This is a case of extrinsic imbalance where the awkwardness is definitely not an immediate result of the information space.

M. Nilashi, *et al.,* [14] presented other information based framework for finding disease utilizing machine learning systems.The proposed Classification and Regression Trees (CART)strategy was construct the fuzzy standards for disease prediction.This method reduces the false prediction and able to perform in different medical datasets. In experimental analysis, UCI medical dataset was used for disease prediction. The prediction accuracy was determined by this strategy with the help of more techniques such as Principle Component Analysis (PCA), clustering and fuzzy rule based systems. Themethod had incomplete observations and imprecise which leads to poor performance in computation time of large data.

S. Sarkar, *et al.,* [15] applied improved machine learning algorithms to predict the outcomes of accident, such asinjury, close miss and damage related to accident data. Two prevalent machine learning algorithms SVM and Artificial NN (ANN) that have been utilized, whose parameters were improved by two intense streamlining algorithms, in particular genetic algorithm and Particle Swarm Optimization (PSO) algorithms are employed for accomplishing a high level of exactness and power. The experimental results stated that PSO-based SVM algorithm attained high exactness and robustness. Moreover, rules are separated by incorporating decision tree C5.0 algorithm with PSO-based SVM display. At last, an arrangement of nine valuable principles were extricated to recognize the underlying causes behind the damage, close miss and property harm cases. In information pre-processing task, a great deal of manual exertion was required to clean the information which is used for examination. Also, informational index utilized as a part of this examination has predetermined number of accident records.

This paper implemented a GWO-RNN method for predicting the different kinds of medical diseases and to overcome the above issues in the existing methodologies, using feature extraction and classification method.

3. Proposed Methodology

The Medical Disease Prediction System is the phenomena of analyzing medical data attributes of patient to recognize the presence of a particular disease or categorize the severity of a disease. Existing disease prediction systems have lower efficiency in processing the medical data due to the presence or processing of erroneous data samples. In this research, GWO and AE basedRNN methods are proposed for efficient medical disease prediction. Here, different kinds of disease based databases like Hungarian data,Cleveland data, Switzerland data, PID and Mammographic Masses used for disease prediction. The proposed architecture consists of several steps, those are data acquisition, preprocessing, attribute selection and classification. Figure 1 shows the proposed architecture.

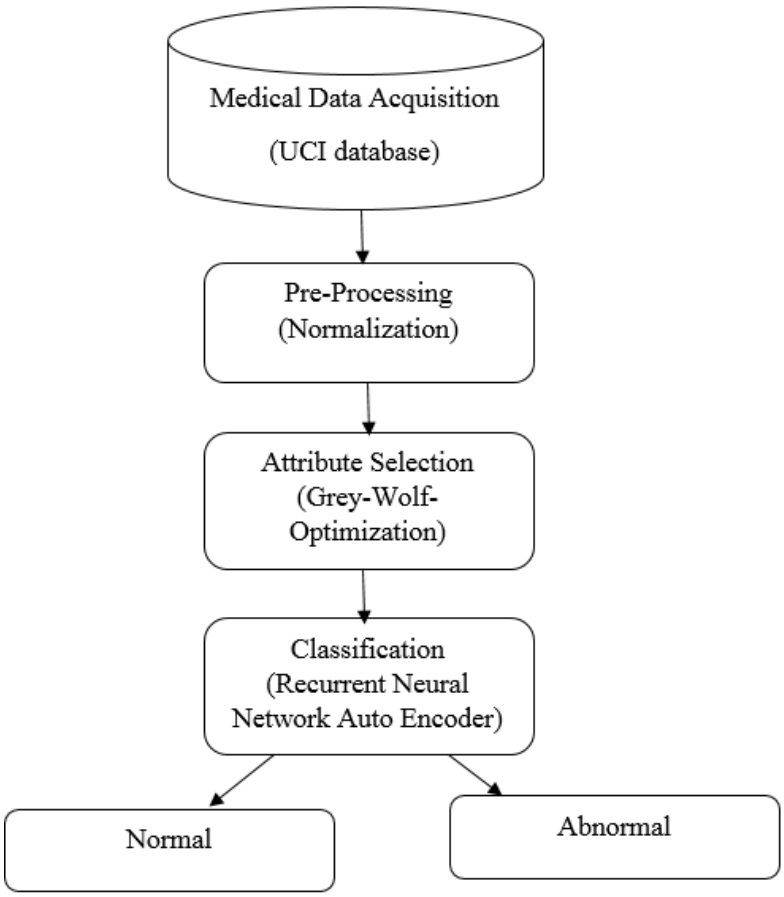

Figure 1 Proposed Architecture of medical data prediction

3.1. Medical Data Acquisition

The databases used in the experiment are different types.Those are Cleveland, Hungarian and Switzerland databases available in the UCI machine learning repository.

(i) Cleveland data

The Cleveland database contains 76 features and all data samples are listed under 14 categories. It represents the value from 0 to 4 stages of heart disease. For privacy purpose some of the patient's name and their social security number are removed from the data and replaced with the dummy values. Six of the 76 characteristics are removed due to incomplete tests performed among the values. This database consists of 56% of sample heart disease data and 46% are not belongs to heart disease.

(ii) Hungarian data

There are lots of data missing in this database and it consists of 123 data samples. It has about 14 features regarding heart disease. Among them 6.5 % of data samples indicate absence of heart disease and 93.5 % of data samples indicate presence of heart disease.

(iii) Switzerland data

This database is similar to the Cleveland database and it contains 261 heart disease data samples. Three or four data samples of the dataset are neglected due to an incompletion of data.The class distribution of this database is 62.5 % of data samples that have an absence of heart diseases data and 37.5 % of data samples have heart disease.

(iv)Pima Indians Diabetes (PID)

PID dataset stands for Pima Indians Diabetes. This database comprises of eight attributes and 768 instances, from National Institute of Diabetes, Digestive and Kidney disease. Where 0 determines negative result and 1 indicates positive result.

(v) Mammographic Masses

This database is collected from the UCI machine learning repository. Six attributes are present in this dataset. Thus, using the age and BI-RADS attributes, it detects mammographic mass lesions which could be either benign or malignant.

The input data is taken from these five databases, which consists of different kinds of diseases such as heart disease, diabetes disease, mammogram related diseases,etc. These raw data are forwarded to the preprocessing step.

3.2. Preprocessing

The preprocessing step fill the missing value, identify or remove the outliers and resolve inconsistencies of data. The raw data have some noise or errors; it is very important to mine the data in order to get better outcomes from the given data set. With a specific end goal to improve the nature of information and resultant information is pre-handled in order to enhance the proficiency and simplicity of mining procedure. In this research work, Normalization method is used for preprocessing because all the attributes are different ranges in the dataset so it's converted into [0, 1] range. In such a way all the attributes are in the uniform range of values. The range value of an attribute will be determined as in equation (1)

$$Range, r = max - min \quad (1)$$

Where, $max \rightarrow maximum\ value of attribute$

$$min \rightarrow minimum\ value of attribute$$

The normalized value of an attribute can be found with the aid of the following equation (2)

value= t- min

t ⟶ Test Sample

$$Normalized\ value, N = value/r \quad (2)$$

Where $value \rightarrow Value\ of\ the\ attribute$

$$r \rightarrow Range\ of\ the\ attribute$$

This conversion of attribute value to [0, 1] depends upon the proportion of a particular attribute covering the range of the corresponding attribute. Then the normalized values are stored in the collection instead of the original values and proceeded for further processes.

3.3. Attribute Selection using Grey Wolf Optimization Algorithm

In this section, GWO algorithm is used for the relevant attribute selection process to improve the efficiency of the medical disease prediction system. The GWO is the meta-heuristic and bio-inspired techniques from nature of grey wolves. In a grey wolf community there are four categories of grey wolves,those are alpha, beta, delta, and omega. Among them alpha is considered to be the leader of the group. Beta wolves assist alpha in decision making and hunting which are considered to the next candidate eligible to be alpha if alpha attains the stage of retirement or death while hunting. Delta wolves are elder wolves or former alpha wolves or sentinels or scout that protects the boundaries of their group. Omega wolves are the least prioritized wolves because it needs to submit all other dominant wolves and follow all other category wolves.

Assume that every wolf as searching solution in the search space. The $w_i = \langle w_{i1}, w_{i2}, \dots \dots w_n \rangle$ is represented as position vectors in search space, whereas, the dimension of the problem is shown as n. The fitness function (based on problem definition) is employed to estimate the position of the wolves. Based on the fitness value the best wolves are classified into three groups such as first solution is represented as α, second is β, and third is δ respectively. In the best solution searching process, the wolves update their position according to the position of α, β, and δ. In the starting stage, the wolves' population is generated and the position of every wolf is initialized. The co-efficient vectors of $\vec{A}$ and $\vec{C}$are described in equation (3) and (4).

$$\vec{A} = 2\vec{a}.\overrightarrow{r1} - \vec{a} \quad (3)$$

$$\vec{C} = 2\overrightarrow{r2} \quad (4)$$

The $\overrightarrow{A}$takes random values in the range of$[-a, a]$, $\vec{C}$ with a random value in the range [0, 2] and it avoids the trap of local optimal. In the traditional GWO algorithm, the vectors linearly decrease from 2 to 0 during the execution of every iteration. Once the coefficients are initialized, every wolf (search agent) fitness value is estimated. After that, best fitness solutions are selected as first, second and third such as α, β, and δ, respectively.

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1}.\overrightarrow{X_\alpha} - \vec{X}\right|, \overrightarrow{D_\beta} = \left|\overrightarrow{C_2}.\overrightarrow{X_\beta} - \vec{X}\right|, \overrightarrow{D_\alpha} = \left|\overrightarrow{C_3}.\overrightarrow{X_\delta} - \vec{X}\right| \quad (5)$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1}.\left(\overrightarrow{D_\alpha}\right), \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2}.\left(\overrightarrow{D_\beta}\right), \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3}.\left(\overrightarrow{D_\delta}\right), \quad (6)$$

$$\overrightarrow{x(t+1)} = \frac{\overrightarrow{X_1}+\overrightarrow{X_2}+\overrightarrow{X_3}}{3} \quad (7)$$

In each iteration of the algorithm, the wolves' positions are updated depends on the position of wolves α, β, and δ according to Equations (3), (4) and (5). In addition, values of vectors $\overrightarrow{A}, \vec{C}$ and $\vec{a}$ are updated. On the basis of new positions, the value of fitness function of wolves iscalculated and α, β, and δ will be selected. The GWO algorithm selects the relevant attributes for medical data prediction.

Pseudocode for GWO algorithm

1. Initialization of grey wolf population $X_i(i = 1,2, \ldots . n)$
2. Initialization of $\overrightarrow{A}, \vec{C}$ and $\vec{a}$ parameters
3. every agent or wolf fitness value
4. X_α $best\ search\ agent$
5. X_β $second\ best\ search\ agent$
6. X_δ $third\ best\ search\ agent$
7. while $t <$ max $number\ of\ iterations$
8. for each search agent
9. update the position of current search agent
10. end for
11. update a, A and C
12. calculate fitness of all search agents
13. update X_α, X_β, and X_δ
14. $t = t + 1$
15. end while
16. return X_α

Furthermore, the swarm intelligent methods are usually used to solve the optimization problem which doesn't have the leader to monitor the entire proceeding period. This limitation is resolved in GWO method; the grey wolves have individual leadership capacity. This algorithm recognizes the different related attributes such as heart disease, breast cancer, diabetes and etc. Also, employs minimum parameters for example, blood pressure, cholesterol, and etc. This algorithm selects the relevant diseases related attributes and forwarded to the RNN.

3.4. Auto Encoder and Recurrent Neural Network using medical decease prediction

The RNN classifier is used for various kinds of medical disease prediction. RNN classifier is a dynamical system that arrange the information in the input sequence and more suitable in data classification as well as prediction.An AE method is to learn an approximation of the identity function from set of unlabeled training sets. Generally, AE is helps to reduce the dimensionality of the data and it's consists of single hidden layer. In encode process the input data is indicated as $x \in R^{d_x}$ to mapping the hidden layers is shown in the equation (8),

$$y = \sigma(Wx + b) \quad (8)$$

Whereas, weight matrix is indicated as $W \in R^{d_y d_x}$, bias vector is represented as $b \in R^{d_y}$and σ is indicated as non-linear activation function in the input layer. Then the target out is represented in equation (9),

$$Z = \sigma'(W'y + b') \quad (9)$$

Whereas, weight matrix and bias vector of output layer is represented as $W' \in R^{d_z d_y}$ and $b' \in R^{d_y}$. In training process of AE, estimates the parameters $\theta = (W, W', b, b')$to reduce the sum of reconstruction cost every training set. The cost calculation is mathematically shown in the equation (10).

$$J = \sum_{x \in D_x} L(x, z) + \lambda \sum_{i,j} {W_{i,j}}^2 \quad (10)$$

Where,

$L(x, z)$ – Estimation of square error

D_x – Training samples in dataset

λ- Hyper parameter of regularization strength

The RNN method able to handle the sequence of input and output data stored in its internal states of the network. Similarly, internal states are holds the previous inputs information and act like a memory. The RNN method is mapped the sequence of fixed vectors and adopt the single fixed output vectors. The general architecture of RNN is shown in the figure.2.

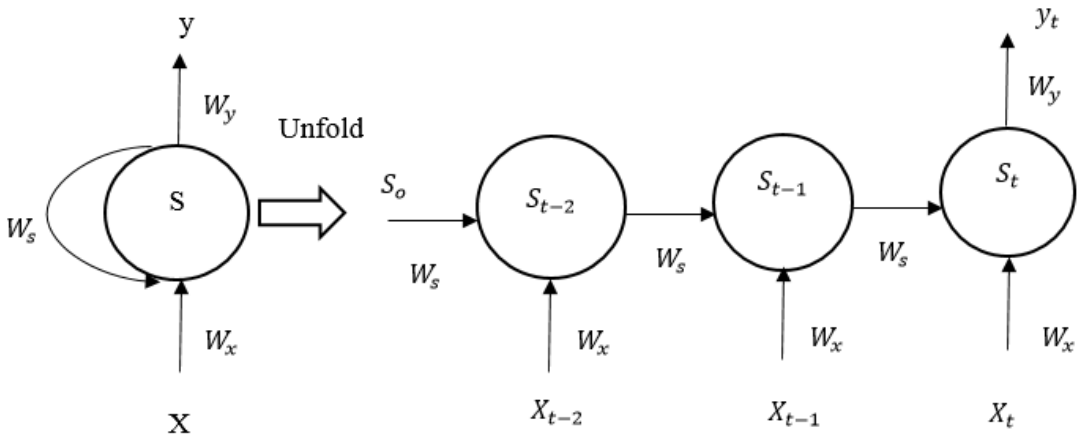

Figure 2 General Architecture of RNN

Let's assume that$x$ is the each input vector of the sequence and internal state is indicated as S which holds the previous state value. Hence,the variable S_t is indicated ashidden state and $t$is represented as time. The calculation of S_t is shown in the equation (11),

$$S_t = \sigma(W_x x_t + W_s S_{t-1}) \quad (11)$$

Whereas, weight matrices are indicated as $W_x$andW_s, input vector is denoted as x_t. The output is indicated as y_t and calculated in equation (12).

$$y_t = \sigma'(W_y S_t) \quad (12)$$

Whereas, σ' is depicted as output activation function and W_y is the weight matrix. The biases terms are omitted from the above equation (12). The AE based RNN algorithm receives relevant disease related attributes from GWO. The RNN classifier quickly predict the different kinds of diseases such as heart disease, breast cancer, diabetes and etc. This algorithm reduces the feature dimensionality problem. The performance of proposed GWO+RNN is evaluated using effective evaluation metrics and demonstrated in following sections.

4. Experimental Results and Discussion

For experimental simulation, PyCharm software was employed in the PC with 3.2 GHz with i5 processor.The proposed medical disease prediction system performance compared to the existing method namely GFMMNN [16]. The performance of the proposed methodology was evaluated by means of accuracy, sensitivity and specificity.

4.1. Performance measure

An evaluation metrics are measure the relationship between the input and output variables of a system. This section describes the different performance measure such as sensitivity, specificity and accuracy. The equation (13) and (14) represents the mathematical description of the specificity and sensitivity.

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (14)$$

Accuracy parameter is estimates the correctly predicted medical disease and wrongly predicted disease. It's mathematically shown in the equation (15).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (15)$$

Where, TP is represented as true positive, FP is denoted as false negative, TN is represented as true negative and FN is stated as a false negative.

4.2. Experimental Analysis using UCI dataset

In this section, UCIdatabase is employed for comparing the performance evaluation of proposed approach and the existing method GFMMNN. In table1, the performance of proposed and existing methods is validated by means of accuracy, sensitivity and specificity.The following table1 tabulatedresult of the proposed and existing method medical data classification performance with respect to five different datasets such as Hungarian, Cleveland, Switzerland, PID and Mammographic Masses.

Table.1 Performance evaluation of different datasets

Dataset	GWO+RNN		
	Accuracy	**Specificity**	**Sensitivity**
Hungarian	95.12	90.00	96.47
Cleveland	98.23	97.6	99.10
Switzerland	91.35	93.45	96.36
PID	96.21	96.78	97.45
Mammographic Masses	95.56	94.57	98.47
Dataset	**GFMMNN [16]**		
	Accuracy	**Specificity**	**Sensitivity**
Hungarian	83.33	69.81	95.2
Cleveland	85.14	85.14	85.14
Switzerland	84.01	70.75	94.68
PID	95.30	93.36	97.03
Mammographic Masses	94.21	91.58	97.48

Table 1 represents the experimental results of proposed and existing methods. According to the table the proposed GWO+RNN method classification performanceis analyzed with respect to different medical datasets and compared it with the existing GFMMNN. The classification performance of GWO+RNN method is measured using efficient parameters such as Accuracy, Specificity and sensitivity. According to the table 1, existing GFMMNN method shows maximum performance variations with respect to different dataset, but proposed GWO+RNN method approximately constant. The graphical representation of accuracy performance is shown in the Figure 3.

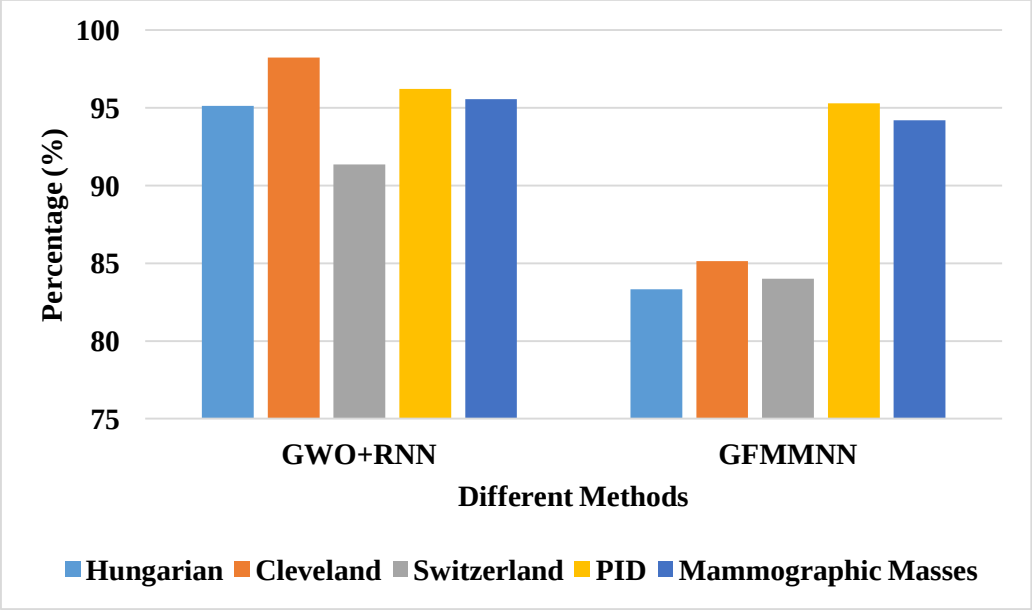

Figure 3 Performance of Accuracy

The Figure 3 shows the performance of prediction accuracy of GWO+RNN and GFMMNN approach. The GFMMNN method achieved 95.30% of accuracy in PID database. Moreover, performance of proposed GWO+RNN method achieved 95.12%, 98.23%, 91.35%, 96.21% and 95.56% of accuracy with respect to five different datasets respectively. Compared to the existing methods the proposed method showed superior results.

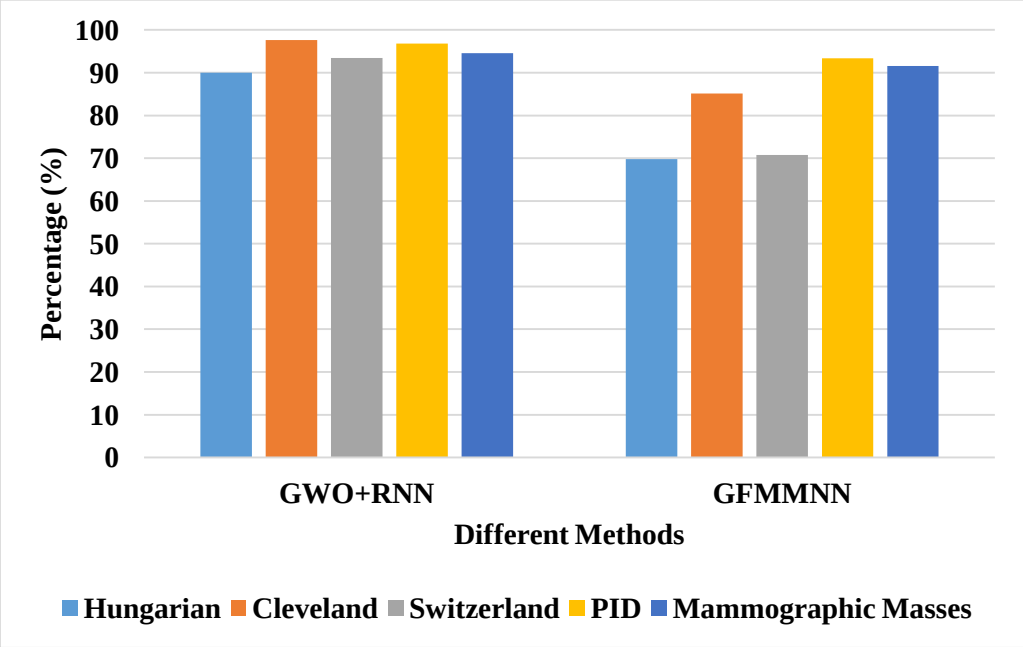

Figure 4 Performance of Specificity

The Figure 4 indicates the performance of specificity in disease prediction system. The traditional GFMMNN method achieved 69.81%, 85.14%, 70.75%, 93.36%, and 91.58% of specificity with respect to different UCI datasets. The proposed GWO+RNN classifier attained 90%, 97.6%, 93.45%, 96.78%, and 94.575 of specificity respectively.

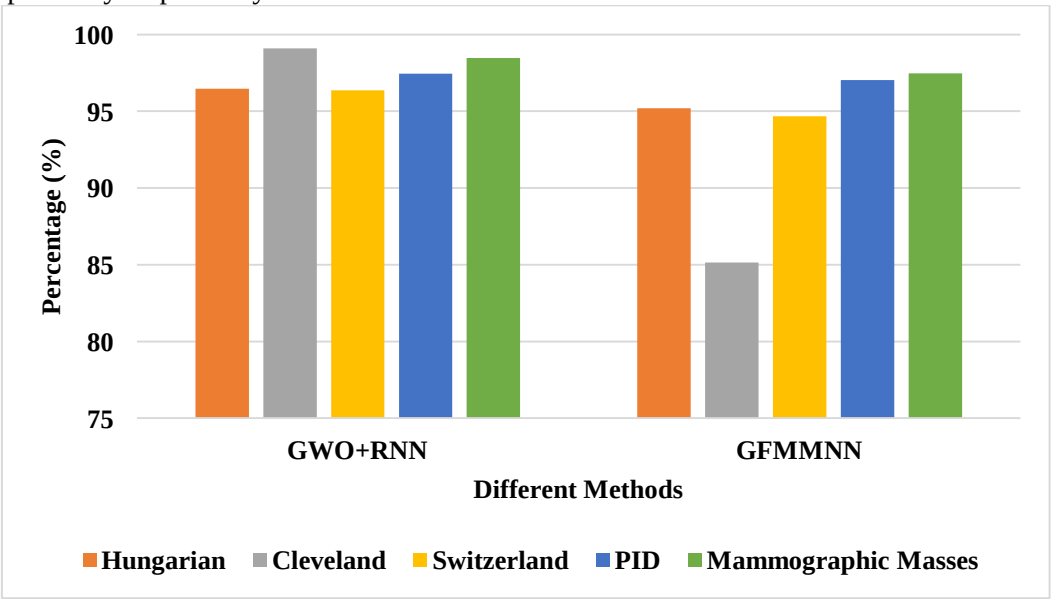

Figure 5 Performance of Sensitivity

The Figure 5 represents the sensitivity performance of medical disease prediction using proposed GWO+RNN method. The traditional GFMMNN method showed better results in Mammographic masses dataset because it achieved 97.48% of sensitivity. The GWO+RNN method attained 96.47%, 99.10%, 96.36%, 97.45%, and 98.47% of sensitivity.

An existing GFMMNN method uses irrelevant and redundant data because it randomly takes the features for classification which degrades the disease prediction accuracy. In order to rectify these problem, an efficient attribute selection method is used namely GWO and AE based RNN classifier that decreases the dimensionality of the features and predict the disease efficiently.

4.3. Comparitive Study

In this section, the table 2 represents the comparative study of existing and the proposed method of medical disease prediction. N. G. Hedeshi, and M. S. Abadeh [17] presented Fuzzy and PSO (Fuzzy+PSO) method for detecting the presence and absence of coronary artery disease in a patient. This algorithm performed faster if suitable values and parameters are set. The major limitation of PSO algorithm is maximum computation time because each time it executes the entire fuzzy rules. M. C. Tu, *et al.* [18] presented Bagging algorithm to identify the warning sign of heart disease. The bagging algorithm combines a series of learned models to generate an improved model. In this paper, the performance of Bagging algorithm was compared with the Decision Tree approach. Also,performance of the medical data classificationmeasured with respect to different parameters such as Accuracy, specificity and Sensitivity.

Table.2 Comparative study of different methods and proposed method

Methodology	Database	Accuracy	Specificity	Sensitivity
Fuzzy +PSO[17]	UCI	85.7	91.08	90.02
Decision Tree [18]	UCI	78.91	84.48	72.01
Bagging Algorithm [18]	UCI	81.41	84.48	74.93
ANN+FNN [19]	UCI	84.2	87.3	80.3
BPNN [20]	UCI	94.51	98.31	87.37
Proposed GWO+RNN	Hungarian	95.12	90.0	96.47
	Cleveland	98.23	97.6	99.10
	Switzerland	91.35	93.45	96.36
	PID	96.21	96.78	97.45
	Mammographic Masses	95.56	94.57	98.47

H. Kahramanli, and N. Allahverdi [19] presented Artificial Neural Network and Fuzzy Neural Network (ANN+FNN) method used for diabetes and heart disease prediction. This method is not able to perform on multiple databases. A. Marcano-Cedeño, *et al.* [20] presented Artificial met plasticity Multilayer Perceptron (AMMLP) algorithm for breast cancer prediction using UCI database. Here, proposed algorithm performance was compared with the Back Propagation Neural Network (BPNN). Finally, medical data classification performance of GWO+RNN method with respect to five different UCI datasets such as Hungarian, Cleveland, Switzerland, PID, and Mammographic Masses is shown in the table 2. GWO+RNN method achieved 98.23% of prediction accuracy with respect to Cleveland dataset that is higher compared to other datasets.

5. Conclusion

In medical field, many researchers focused onprediction of different disease using various medical applications. In this research, GWO+RNN technique is used for medical disease prediction. The GWO method is used for feature selection, which removes the redundant and irrelevant attributes. An AE based RNN classifier predicts various diseases and avoids the feature dimensionality issues. This GWO+RNN method is tested using five benchmarks of UCI dataset namely Hungarian, PID, Mammographic masses, Cleveland and Switzerland. In experimental analysis, classification performance of the proposed method is compared with the existing method namely GFMMNN, ANN+FNN, BPNN, and DT. The performance of GWO+RNN method is calculated in terms of different evaluation metrics like specificity, sensitivity and accuracy. The GWO+RNN method achieved 16.825%of improved accuracy in Cleveland dataset for disease prediction. In future, research work can be extended as an efficient hybrid technique for improving the efficiency of different medical disease classification.

References

[1] M. Niksic, B. Rachet, S. W. Duffy, M. Quaresma, H. Moller, L. J. Forbes, "Is cancer survival associated with cancer symptom awareness and barriers to seeking medical help in England? An ecological study," *British journal of cancer*, vol. 115, no. 7, pp. 876, 2016.

[2] M. Scatà, A. Di Stefano, P. Liò, and A. La Corte, "The impact of heterogeneity and awareness in modeling epidemic spreading on multiplex networks," *Scientific reports*, vol. 6, pp. 37105, 2016.

[3] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand, "A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques," *Biocybernetics and Biomedical Engineering*, vol. *38*, no. 1, pp. 1-15, 2018.

[4] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A Manufacturing Big Data Solution for Active Preventive Maintenance", *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2039-2047, 2017.

[5] W. Yin and H. Schutze, "Convolutional neural network for paraphrase-identification." in HLT-NAACL, pp. 901–911, 2015.

[6] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," *IEEE Transactions on Industrial Informatics*, vol. 99, no. 1, 2016.

[7] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 3, pp. 1294–1307, 2016.

[8] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1033–1069, 2015.

[9] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015.

[10] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," *Journal of biomedical informatics*, vol. 53, pp. 220–228, 2015.

[11] Y. Chen, N. Crespi, A. M. Ortiz, and L. Shu, "Reality mining: A prediction algorithm for disease dynamics based on mobile big data," *Information Sciences*, vol. 379, pp. 82-93, 2017.

[12] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869-8879, 2017.

[13] R. Prashanth, and S. D. Roy, "Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning." *Neurocomputing*, vol. 305, pp. 78-103, 2018.

[14] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Computers & Chemical Engineering*, vol. 106, pp. 212-223, 2017.

[15] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Computers & Operations Research*, 2018.

[16] D. M. Rafi, and C. R. Bharathi, "Optimal Fuzzy Min-Max Neural Network (FMMNN) for Medical Data Classification Using Modified Group Search Optimizer Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol.*9, Issue.*3, pp.1-10. 2016.

[17] N. G.Hedeshi, and M. S. Abadeh, "Coronary artery disease detection using a fuzzy-boosting PSO approach", *Computational intelligence and neuroscience*, p.6. 2014.

[18] M. C. Tu, D. Shin, and D. Shin, "Effective diagnosis of heart disease through bagging approach", In *Biomedical Engineering and Informatics, BMEI'09. 2nd International Conference on*, pp. 1-4, IEEE, 2009.

[19] H. Kahramanli, and N.Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", *Expert systems with applications*, vol.*35, Issue.* 1-2, pp.82-89, 2008.
[20] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network", *Expert Systems with Applications*, vol.*38, Issue*.8, pp.9573-9579. 2011.

BIBLIOGRAPHY OF AUTHORS

Dr. B. SANKARABABU was a doctorate from Aacharya Nagarjuna University, Guntur, Andhrapradesh, India, and completed PhD in 2016. Currently working as a Professor in Department of Computer Science and Engineering at Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad .He has done his research on Data Mining. He has published several research papers on Data mining concepts.

G.Charles Babu , Presently working as a Professor in Dept. of CSE in Malla Reddy Engineering College(Autonomous), Secunderabad, Telangana Since 5 Years and Total Teaching experience of 20 Years. Completed B.Tech (CSE) in 1997 from KLCE, M.Tech(SE) in 1999 from JNTUH and Ph.D(Data Mining) from ANU. Published more than 50 Research Papers in Data Mining, Cloud Computing, Image Processing.

Dr. Y. Jeevan Nagendra Kumar, obtained his Ph.Din Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP in 2017 and M.Tech Computer Science Technology from Andhra University in 2005. He is working as Professor and Dean - Technology and Innovation Cell in GRIET since 2005. He has about 9 Research Papers in International / National Conferences and Journals and also attended many FDP Programs to enhance his knowledge. With his technical knowledge he guided the students in developing the useful Web applications and data mining related products.

Dr. G.Karuna, Professor in Computer Science and Engineering, completed her Ph.D from JNTUH, Hyderabad and she has over twelve years of academic and research experience. Currently working as a professor in CSE Department and prior to this, she worked for Malla Reddy Institute of Technology & Science andJ.B.Institute of Engineering & Technology. Her Ph.D. thesis was titled as "An Effective Representation of Shape for Object Recognition and Classification based on Statistical and Structural Features" in the area of Image Processing. She has published25 papers in referred International Journals and Conferences and written two books for SCDE, JNTUH, and Hyderabad.