An Efficient Disease Prediction in Big Data using Neuralnetwork based Optimization Method

K.Tamilselvi[a] and Dr. K.Ramesh Kumar[b]
[a] Research Scholar, Research and Development, Bharathiar University, Coimbatore - India
[b]Research Guide, Bharathiar University, Coimbatore – India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Generally, with Big Data concept large complex data are handled easily and in healthcare it helps to access data rapidly. Previously, Big Data in healthcare reached a complexity level but sometimes it is impossible to obtain the required data for use; thereby the growth in healthcare sector is slow. Big Data is much interesting when used to analyze healthcare data. Classifiers with cost-sensitive factors increases the stability of classification and reduces its computational costs when dealing with large scale, imbalanced and redundantdatasets like medical data. Moreover, the nature and growth of disease are unknown; thus, making prediction complex. This work predicts various diseases where the parameters of neural networkare optimized by using big data which includes data from social media. The efficiency of the TRAPezoidal Neural Network (TRAP-NN) and Improved Whale Optimization Algorithm (WOA) learning models named as (IWOA-TRAP-NN) were compared with the two standard methods such as Genetic Algorithm optimized Convolutional Neural Network (GA-CNN) and Ant ColonyOptimized Convolutional Neural Network (ACO-CNN). The proposed TRAP-NN achieves 86% accuracy, 74%of F1 score and 56.2% of kappa static. The results show that the TRAP-NN performs better than GA-CNN and ACO-CNN.

Keywords: Optimization, neural network, disease prediction, bigdata, healthcare.

1. Introduction

Several kinds of pressure come in life like aging and accelerating pace of life. Every factorincreases the frequency of diseases like diabetes and canceryear by year. Hence, preventing measures andproviding treatment to cure the diseases is of demand in medical as well as healthcare environment. Simultaneously, as medical informatization has been advanced continuously, public health sector of China has gathered a wealth of data resources which is similar to big data. The enormousresources of medical data generally contain more valuable information which includes diagnosed results of patients and the rules to provide treatment. As a distinctivecontinuing epidemic, the frequency of diabetes is high recentlywhich is an increasing trend [1].

At present, models related to machine learning are commonly employed for predicting the diseases [2]. Several research works have been carried out in diagnosing diabetes and providing treatmentwhere some results were obtained. When 16 patients were monitored, statistical methods were used for analyzing the risks of diabetes [3]. Simultaneously, few research works employed decision tree and multilayer perceptron approaches for comparative analysis. Machine learning techniques like random forest (RF), Support Vector Machine (SVM), and naive Bayes (NB) were used to observe the early symptoms for predicting diabetes. From the results, it was proved that decision tree (DT) and RF methods achieved better prediction results while dealing with diabetes data [4]. In the meanwhile, deep convolutional neural network (DCNN) was used in several applications due to its powerful ability of extracting features. It possibly extracted deeper features from a huge training data due to the hierarchical network structure which was not able to obtain using conventional classifiers[5]. Thus, this approach was extensively employed in applications like image recognition, speech recognition, text detection etc. [6]. It is well known that medical dataset has rich characteristics and featureswhich helps in discovering potential laws and valuable information fromwhen DCNN was applied to medical data [7]. Practically, it has several significant potential and social value [8]. Most of the existing works tried to predict disease using only the search query data via Internet. But there aroused a demand to consider different big data and environmental factors while predicting diseases [9]. Moreover, with models using deep learning approaches, performance of prediction rate was improved by optimizing the parametersof these approaches [10]. In view of this, the main objective is to classify the risk level of the patient by predicting the disease using efficient optimization technique which hence reduce the massive challenges.

The remaining part of this paper is presented as follows: Section 2 discusses the works in the literature relevant to the system proposed. Section 3 elaborates the designed disease prediction model withoptimization adopted at various stages. The results are presented in section 4. At last, section 5 concludes the work with closing remarks and future directions.

2. Literature survey

In[11], generic structure-based assessment method was predicted which included non-imaging and imaging data termed as Graphic Convolutional Networks (GCNs) and was employed in application dealing with brain. In. [12] a framework was designed to examine the efficiency of various classifiers and classifiers developed based on Ensemble Classifier (EC). In [13], a novel methodological regime was formulated for diabetes and the classification was based on fuzzy rule to provide treatment for the disease. In[14], heart disease prediction approach was developed using Machine Learning techniques like Particle Swarm Optimization(PSO) and Ant Colony Optimization(ACO) techniques. In [15], five different classifiers were examined which included Artificial Neural Network (ANN), SVM, RF, DT, K-Nearest Neighbor (KNN) and observed that RF classifier produced the highest accuracy among all. In [16], Decision tree based Neural Fuzzy System (DTNFS) approach was introduced for analyzing and predictingdifferent sort of heart diseases. This work developed a cost-effective intelligent system which outperformed the existing systems. Particularly, data mining approaches were employed for improving the prediction of heart disease. It was found that SVM and NN produced higher prediction rate for heart disease. Yet, data mining techniques are not suitable for the prediction of heart disease. In [17],for heart disease prediction employed Genetic algorithm and generated optimum set of attributes which as useful for prediction.

3. Proposed Methodology

The major objective of this proposed model is to classify the risk level of the patient by predicting the disease. Fig. 1 demonstrates the general architecture for the proposed framework. Once the symptoms areidentified in a patient efficiently, an optimal approach is applied to the model to improve the performance of the system thereby determining an optimumseries of diseases. This prediction model is designed using TRAP-NN as higher accuracy rates for prediction can be produced.

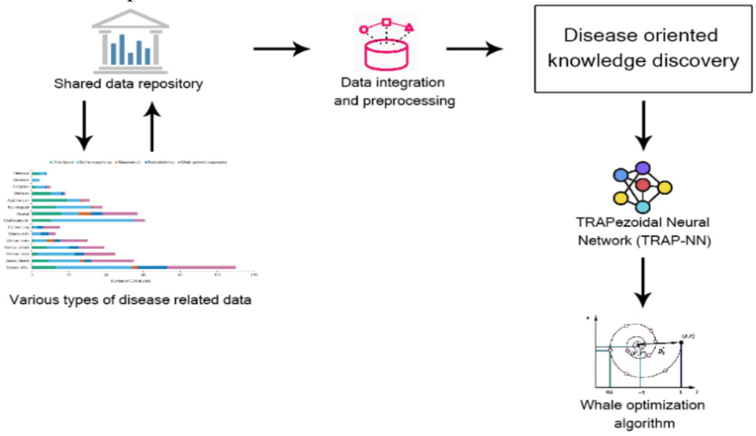

Figure-1 System architecture for disease prediction

4. Dataset description

UCI archive data is used which consists of 270 cases with complete characteristics. UCI is one of the open-source online dataset with huge several sicknesses, area speculations and information inventors utilized by specialists.

Table-1 Dataset description

Type of Data	Items	Description
Structured	Patient's Demographics	Details like Gender, height, weight, age etc.,
	Living habits	Smoking or not
	Examination items and results	Physical check
	Diseases	Cerebral infraction and other heavy disease
Unstructured	Readme illness of the patient	Health history (Illness or not)
	Medical records maintained by doctors	Integration records

5. Construction of network

To classify various diseases automatically, a TRAPezoidal Neural Network (TRAP-NN) was designed and implemented with unifying 3D bounding box estimation and in-region where features and weights are shared among various tasks. Basically, TRAP-NN uses the principle of FPN, where various feature map levels are combined for promoting discriminative feature extraction. The responsibility of the encoder is to focuson the 3D bounding box estimation. Decoder classifies VoI in-region of shared features with various levels of pyramid network. Convolutional 3D kernels are more expensive than 2D variants. Further, 3D framework has numerous trainable parameters, where every layer of the model adds

ClCl-1 $\prod i = \{x, y, z\} kl(i)$weight. Cl indicates the number of feature maps in layer l, and k { x,y,z} represent the kernel size related to the spatial dimension. Due to this, the network is increasingly inclined to overfitting thereby drastically increasing the GPU memory.

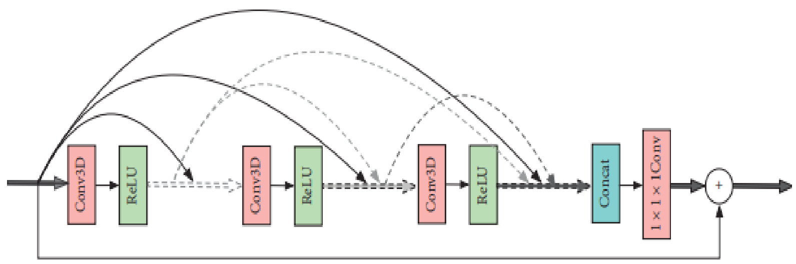

Figure 2- Residual dense block as the building module

TRAPezoidal Neural Network is promoted by fusing the features at various levelsusing skip connection for obtaining the ability of advanced learning. To classify the disease in a better way, multilevel feature fusion layers in the network are designed where the estimated 3D bounding box is directly cropped without resampling and then are embedded into decoder.

Improved Whale Optimization algorithm(IWOA)

There are three main stages in the whale optimization algorithm, which are searching prey, encircling prey and Logarithmic spiral prey. Among them, searching for prey is the exploration stage of the algorithm. Moreover, encircling prey and Logarithmic spiral prey is the exploitation stage of the algorithm.

Step 1: Obtaining the objective function value of all points, and getting the best point as the X g , and then getting the second-best point as theX b , supposing that X s is the one should be substituted. f (X) s , f (X) b and f (X) g show the objective function values.

Step-2: Obtaining the objective function value of the middle point Xc between point Xg and point Xb

$$Xc=\frac{Xg+Xb}{2}$$

Step 3: Obtaining the reflection point Xr by using the following formula. α is the reflection coefficient, which is set as 1:

Xs–（Xc α+ Xc =Xr ）

Step 4: If f (Xr)<f(Xg) getting the expansion point from formula

Xe=Xc+μ(Xr-Xc)

where the expansion coefficient is μ, which usually set to 2. If f (Xe) <f (Xg),Xs will be replaced by Xe, otherwise, Xs will substitute Xr .

Step 5: If f (X) < f (X) rs , the compression point can be acquired

Xe=Ac+Ω(Xs-Xc)

Step-6: I f(Xg)<f(Xr)<f(Xs), shrink point is Xw and the shrink coefficient is Ω

Xw=Xc-α(Xs-Xc)

IWOA IN TRAP-NN

The representation between slices is learnt using slice-wise NN by using cuboid kernels of size 1×1×n. The aim is to wrap the 3D input into a 2D feature map using 1D slice-wise convolution without considering the channels and this operation enables the network to concentrate on slices. Therefore, initially, 2D convolutional kernels are set to 1, where the third one n, depends on the number of expected sets,a convolution can stackbefore deriving a 2D feature map,

$$n = \frac{D-1}{t}+1$$

here D denotes the third dimension. The feature maps of various scales obtained for the same view are combined using element-wise addition for strengthening the unique patterns

y(v)=f(x{wi}1+fv(x,wi)2+Fv(x(wi}3)

F v (x, {Wi}j) is a function which has to be learned for transforming input x to different feature maps, where j ∈ {1, 2, 3} represent various scales.

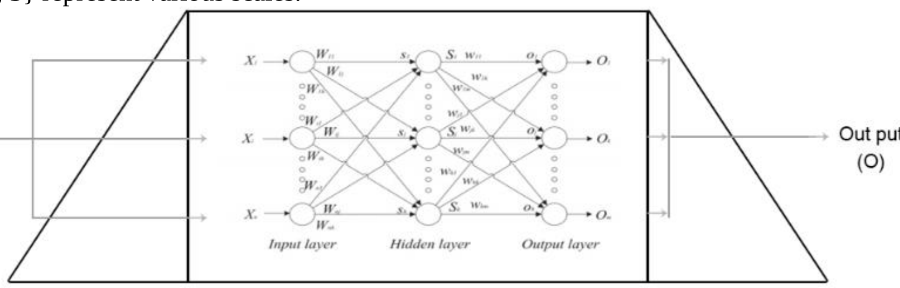

Figure-3 Trapezoidal Neural Network

Before optimizing perceptron, initially, neural network structure is obtained. The neurons of the input as well as output layer is adopted using classified datasets, while the neurons of the hidden layer is decided by the Kolmogorov theorem

Hidden =1+ Input× 2

when IWOA is adopted to obtain the values ofconnection weights and bias, D represents the dimension of the candidate solution

D=(input×hidden)+(hidden×output)+hidden[bias]+output[bias]

where the number of neurons of the input, hidden and output layers are expressed by Input, Hidden and Output. For the hidden and output layer, number of bias is shown by Hidden [bias] and Output[bias]

The mathematical models of the IWOA algorithm are denoted by vectors, so vectors N is used to represent agents in the population. Each agent in population is denoted, X n KX1 , X 2 , X3 ,=X the weights of the input and hidden layer are denotedK==,i 1,2, ,n, i }iw,hw,hb,ob {by X byiwand hw, the biases of the hidden layer and the output layer are shown by hb and ob . Generally, after agents are defined by vectors, the objective function of NN defines the fitness function of TSWOA algorithm. Hence, the fitness function is considered as the difference between the theoretical and actual output in the neural network

MSE= $\sum_{I=0}^{n}(\mathrm{Ok} - \mathrm{Dk})$

where the total number of output is shown bym, the desired output and the actual output of ithinput are exhibited by k di and k O

Algorithm: Improved Whale Optimized Algorithmin TRAPezoidal Neural Network (IWOA-TRAP-NN)
Input: Dataset (D), weight matrix W *, hidden and visible layer element bias* B *and* A *respectively*
Output: Categorized disease
Start
θ= *{W, A,* B *}.*
Initialize global search
Input parameters- max_ iter, s, n2, η
Construct the population M
Mi = {Mi1, Mi2, ... , MiD}
Evaluate the fitness f(t)
τi {t} ← *D*
D+1 ← *If τi {t−1}*
Update spiral (p), p= τi {t}
Update the position
Xi(t)=xi(t-1)+Rw(τixi(t-1))
Concurretn factor rescursion in network by
*W={wij€R(n*m)}*
If w<D then
Initialize the bias threshold
Else
Go for global search
Hidden layer (H)={H1,H2,H3….Hn}
K(n<nk)
P(xk<m), ni(n1,n2.n3….nm)
MSE= $\sum_{I=0}^{n}(Ok - Dk)$
Nm ← *MS*

6. Performance analysis

For experiment, standard datasets including20 types of disease are used which can be accessed from secondary database. The data available in the datasets were captured from various scenarios in health sector environment, which broadly evaluates the proposed TRAPezoidal Neural Network (TRAP-NN) and Improved Whale Optimization Algorithm (IWOA) learning models which is named as (IWOA-TRAP-NN) were compared with two standard methods such as Genetic Algorithm optimized Convolutional Neural Network (GA-CNN) and Ant ColonyOptimized Convolutional Neural Network (ACO-CNN) The tool used for obtaining the result is PHYTHON.

Accuracyis the ability of prediction obtained by proposed deep learning model. True positive (TP) and true negative (TN) are the predictions made by the model indicating the presence and absence of attack. False positive (FP) and false negative (FN) are the false predictions made.

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

Table 2 shows the comparison of accuracy between existingGA-CNN, ACO-CNN and proposed IWOA-TRAP-NN.

Table-2 Comparison for Accuracy

Number of Datasets	GA-CNN	ACO-CNN	IWOA-TRAP-NN
100	45	78	81
200	59	80	83
300	61	84	85

400	72	86	89
500	85	89	94

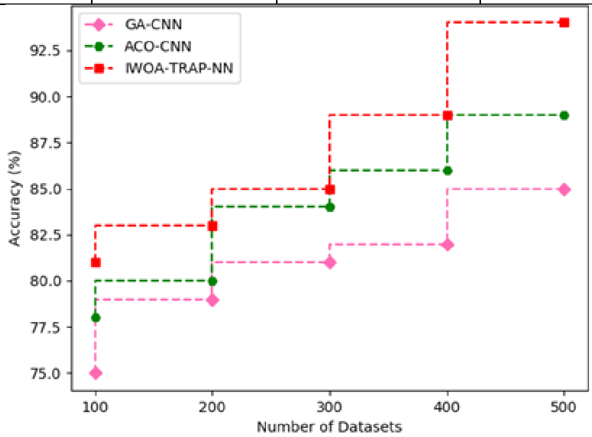

Figure 4: Comparison of accuracy

The figure 4 shows the comparison of accuracy between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NNmethods. X axis represents the number of datasets used for analysis and Y axis indicates the obtained accuracy values in percentage. When compared, existing method achieves 81% and 80% while the proposed method achieves 4% better than GA-CNNand 2.1 better than ACO-CNN.

Precision and sensitivitygive the success of the attack and classification model accordingly. Precision describes the positive predictions made by the classifier in the presence of disease. It is given by:

$$Precision\ (P) = \frac{TP}{TP + FP}$$

Specificitygives the negative prediction of the classifier in the absence of the disease and is estimated by:

$$Specificity(S) = \frac{TP}{TP + FN}$$

Table 3 shows the comparison of specificity between existingGA-CNN, ACO-CNNand proposed IWOA-TRAP-NN

Table-3 Comparison for specificity

Specificity	**GA-CNN**	**ACO-CNN**	**IWOA-TRAP-NN**
50	62	66	70
75	66	70	72
100	70	72	75
125	72	75	79
150	74	79	85

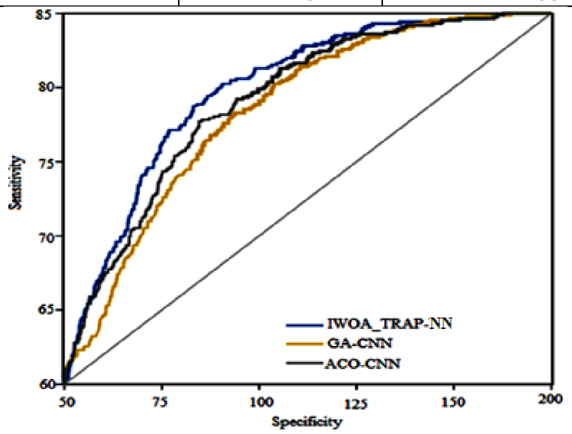

Figure 5 Comparison of specificity

The figure 5 shows the comparison of specificity between existing GA-CNN, ACO-CNN and proposed IWOA-TRAP-NNmethods whereX axis represents the specificitywhile Y axis represents the sensitivity in percentage.

F1- Score is utilized to determine the prediction performance. It is the harmonic mean (or weighted average) of both the precision as well as recall. If the score is 1, the model is said to be the best else if 0 it is worst. F1-Score is estimated by:

$$F1 - Score = \frac{2 * P * R}{P + R}$$

Table 4comparestheF1-score between existingGA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN

Table-4 Comparison for f1-score

Number of Datasets	GA-CNN	ACO-CNN	IWOA-TRAP-NN
100	65	68	71
200	69	70	73
300	71	74	75
400	72	76	79
500	75	79	84

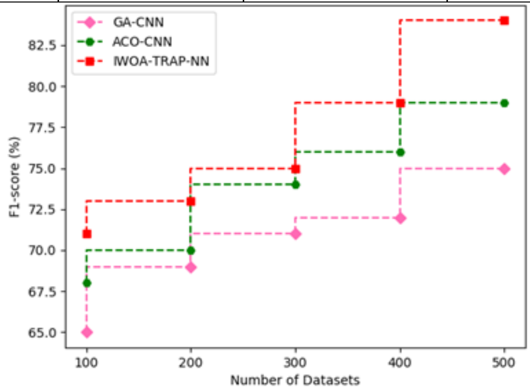

Figure 6Comparison of f1-score

The figure 6illustrates the comparison of f1-score between existing GA-CNN, ACO-CNNand proposed IWOA-TRAP-NN whereas X axis shows the number of datasets and Y axis shows f1-score in percentage.

The **kappa static**generally test the interrater reliability whose importance represents the correctness of the data collected.

Table 5compares the kappa static between existingGA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods.

Table-5 Comparison for kappa static

Number of Datasets	GA-CNN	ACO-CNN	IWOA-TRAP-NN
100	60	64	68
200	64	68	70
300	69	70	73
400	70	72	75
500	72	75	83

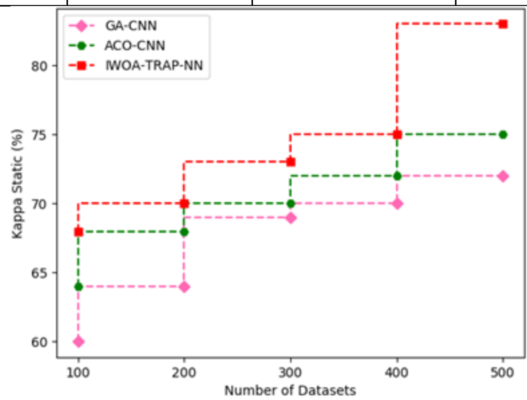

Figure 7Comparison of kappa static

The figure 7 shows the comparison of kappa static between existing GA-CNN, ACO-CNN and proposed IWOA-TRAP-NNmethods. X axis represents the number of datasets and Y axis provides the obtained kappa staticvalues in percentage .

Table-8 Overall comparative analysis

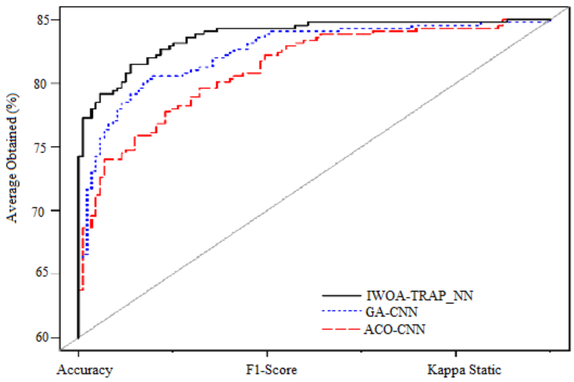

Figure 7 Overall analysis

Figure-8 shows the comparison of various parameters between existingGA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN. As a result the proposed method achieves 86% accuracy, 74%of f1 score and 56.2% of kappa static.

7. Conclusion

In this paper,a Improved Whale Optimized Algorithm TRAPezoidal Neural Network (IWOA-TRAP-NN) that is based on the optimization process is proposed to speed up the entire computational process of hospital sector big data. This proposed method consists of three modules such as, classification, prediction and learning process which usesfew characteristics, like persist/cache strategies, fault tolerance, for faster decomposing than the existing algorithms. As a result, several experiments were carried out to test the efficiency of IWOA-TRAP-NN method for medical big data disease prediction and achieved86% of accuracy, 74%of F1-score and 56.2% of kappa static.

References

1. T. F. Blaschke, M. Lumpkin, and D. Hartman, "The World Health Organization prequalification program and clinical pharmacology in 2030," *Clinical Pharmacology and Therapeutics*, vol. 107, no. 1, pp. 68–71, 2019.View at: Google Scholar

A. M. Carracher, P. H. Marathe, and K. L. Close, "International Diabetes Federation 2017," *Journal of Diabetes*, vol. 10, no. 5, pp. 353–356, 2018.View at: Publisher Site | Google Scholar

2. P. Qian, H. Friel, M. S. Traughber et al., "Transforming UTE-mDixon MR abdomen-pelvis images into CT by jointly leveraging prior knowledge and partial supervision," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2020.View at: Publisher Site | Google Scholar

3. K. Xia, H. Yin, P. Qian, Y. Jiang, and S. Wang, "Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal CT images," *IEEE Access*, vol. 7, pp. 96349–96358, 2019.

4. K. Xia, H.-s. Yin, and Y.-d. Zhang, "Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-Flow algorithm," *Journal of Medical Systems*, vol. 43, no. 1, 2019, 2:1-2:12.View at: Publisher Site | Google Scholar

5. Y. Jiang, D. Wu, Z. Deng et al., "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 12, pp. 2270–2284, 2017.

6. Noden B.H., Kent M.D., Beier J.C. The impact of variations in temperature on early Plasmodium falciparum development in Anopheles stephensi. *Parasitology.* 1995;**111**:539–545.

7. 41. Liang W., Gu X., Li X., Zhang K., Wu K., Pang M., Dong J., Merrill H.R., Hu T., Liu K., et al. Mapping the epidemic changes and risks of hemorrhagic fever with renal syndrome in Shaanxi Province, China, 2005–2016. *Sci. Rep.* 2018;**8**:749. doi: 10.1038/s41598-017-18819-4.

8. 42. Huang X., Clements A.C.A., Williams G., Milinovich G., Hu W. A threshold analysis of dengue transmission in terms of weather variables and imported dengue cases in Australia. *Emerg. Microbes Amp Infect.* 2013;**2**:e87. doi: 10.1038/emi.2013.85

9. Li Q., Guo N.N., Han Z.Y., Zhang Y.B., Qi S.X., Xu Y.G., Wei Y.M., Han X., Liu Y.Y. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. *Am. J. Trop. Med. Hyg.* 2012;**87**:364–370.

10. Parisot S., Ktena S.I., Ferrante E., Lee M., Guerrero R., Glocker B. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimers disease. *Med Image Anal.* 2018;48:117–130.

11. Weng C.-H., Huang T.C.-K., Han R.-P. Disease prediction with different types of neural network classifiers. *Telemat Informat.* 2016;33(2):277–292.

12. 16. Kumar P.M., Lokesh S., Varatharajan R., Babu G.C., Parthasarathy P. Cloud and iot based disease prediction and diagnosis system for healthcare using fuzzy neural classifier. *Future Generat Comput Syst.* 2018;86:527–534
13. V. V Ramalingam, Ayantan Dandapath and M. Karthik Raja, "Heart Disease Prediction using Machine Learning techniques: a survey International Journal of Engineering & Technology, 7 (28), pp 684 – 687, 2018
14. Youness Khourdifi and Mohamed Bahaj, "Heart Disease Prediction and Classification using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering & Systems, Volume 12, No 1, pp 242-251, 2019
15. Jun Ni, Ying Chen, Jie Sha, and Minghuan Zhang, "Hadoop-based Distributed Computing Algorithms for Healthcare and Clinic Data Processing", IEEE, 2015, pp. 188-193, DOI: 10.1109/ICICSE.2015.41
16. Akhil Jabbar M , Priti Chandra, Deekshatulu B L. Heart Disease Prediction System using Associative Classification and Genetic Algorithm. Soft Computing inData Analytics, part of AISC volume 78,pp, 735-742, 2018