Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

An Artificial Intelligence Algorithm to Optimize the Classification of the Hepatitis Type

Hiba Hathal Khalil (1)
University of Baghdad/ College of Administration and Economics/ Baghdad, Iraq
heba.hathal1201a@coadec.uobaghdad.edu.iq

Sabah Manfi Rada(2)
University of Baghdad/ College of Administration and Economics/ Baghdad, Iraq
drsabah@coadec.uobaghdad.edu.iq

Received:26/2/2023 **Accepted:26/3/2023** **Published: March / 2023**

Abstract

Hepatitis is one of the diseases that has become more developed in recent years in terms of the high number of infections. Hepatitis causes inflammation that destroys liver cells, and it occurs as a result of viruses, bacteria, blood transfusions, and others. There are five types of hepatitis viruses, which are (A, B, C, D, E) according to their severity. The disease varies by type. Accurate and early diagnosis is the best way to prevent disease, as it allows infected people to take preventive steps so that they do not transmit the difference to other people, and diagnosis using artificial intelligence gives an accurate and rapid diagnostic result. Where the analytical method of the data relied on the radial basis network to diagnose the disease, in addition to using the classification of the regression tree as well as the use of the genetic algorithm to improve the classification accuracy of both methods and by comparing the methods used to find out the most efficient methods of classification through criteria. Classification error, mean square root error, and average absolute relative error, and concluded that the experimental results are that the methods are good in terms of classification, as they gave results with less classification of error, and that the radial basis network was superior to the classification regression tree, and that the addition of the genetic algorithm led to an improvement classification accuracy.

Paper type: Research paper.

Keywords: Regression Tree Classification (CART) ,Radial Basis Function (RBF) ,Genetic Algorithm (GA)

(1)

(2)

1. Introduction

Hepatitis has become more improved in recent years in terms of the rise in the number of infections (Ismail, et al, 2019). Where doctors need to diagnose the disease with high accuracy, and the use of technology leads to an accurate diagnosis of the disease, especially the diagnosis using artificial intelligence, as it gives accurate and rapid diagnostic results. Therefore, artificial neural networks (ANN) were used, characterized by their ability to classify and process data, as well as their ability to deal with different types of data. The classification regression tree (CART) method, where its analysis strategies were used as a classification tool. In addition to developing a hybrid method of ANN-GA and CART-GA using the genetic algorithm, a number of studies dealt with the diagnosis of hepatitis, including. Bedossa and Poynard (1996) proposed a study to diagnose hepatitis by testing the accuracy of the simple algorithm, as they concluded that the agreement between the reference activity and the specific activity was great. Kurosaki et al (2010) studied a number of drugs for the treatment of viral hepatitis C through the use of classification and the CRAT regression tree in the recursive division method, and the researcher concluded that the decision tree is useful for predicting the response before treatment PEG-IFN As well as RBV. Singh and Pandey (2014) proposed a study to diagnose liver disorders, hepatitis and liver cancer using artificial neural network (ANN), data mining method (DM) and fuzzy logic (FL). Anto and Chandramathi (2015) proposed a study for the diagnosis of hepatitis based on the support vector machine (SVM) and hybrid genetic algorithm – simulated modeling (SA). Khatoon and Agarwal (2019) proposed a study to diagnose hepatitis using the decision tree algorithm and the genetic algorithm. Syafa'ah et al (2021) presented a study to diagnose hepatitis such as cirrhosis and liver cancer through liver enzymes, using machine learning diagnostic methods such as (random forests, KNN, naïve Bayes, neural network. They concluded that the method of neural networks accuracy is better than other methods. Obaido et al (2022) presented a study for the diagnosis of hepatitis B that uses a game-based theoretical approach to explain and visualize the predictions of machine learning models applied to hepatitis and includes decision tree and logistic regression algorithms.

Diagnosis is the process of determining the disease or determining the condition of the patient according to the symptoms he suffers from, and the use of artificial intelligence gives more accurate results, which helps to diagnose accurately, and there are many methods of artificial intelligence used diagnosis and among these methods is the radial basis function network and the classification regression tree.

2. Materials and Methods

2.1 Classification and regression tree (CART)

It is considered a non-parametric statistical methodology that incorporates all categorical and numerical variables into the analysis. It is used in the problem of classification and regression, and the general goal of CART analysis is to form models that use expectation variables to predict the values of the response variable in a non-parametric manner in the basis, and CART analysis is division process binary, is repeated successively for a data set based on the division criteria that determines (Lewis, 2000).

i. Regression Tree Generation

The model expresses the regression tree as follows (Ahmed, 2016):

$$f(x) = \sum_{m=1}^{n} C_m I(x \in R_m) \quad (1)$$

Cm= node means where

$$\hat{C}_m = \frac{1}{N_m} \sum_{X_i \in R_m} Y_i \quad (2)$$

It divides the origin node into binary sections in succession, depending on the scale of genetic impurities (Karthikeyan and Thangaraju, 2013):

$$G(t) = 1 - \sum_{i=1}^{k} (P_i)^2 \quad (3)$$

As the independent variable that resulted in the greatest improvement in the genetic standard for dividing the parent node into child nodes.

ii. Classification tree (CT)

After the division through the impurity scale, where the parent node is divided into two sub-nodes, after which the binary division begins in succession according to the level (Lewis, 2000).

iii. Tree growth

The construction starts from the root node, by binary division repeatedly and in a successive manner from groups at each level of tree growth, the parent node produces sub-nodes, which are at least two sub-nodes.

That is, the parent node is divided into two sub-nodes, and each branch in turn is divided into other sub-nodes (Ma, 2018), and the division occurs under a statistical test called the impurity scale. One of the common measures in which impurities are measured is the genetic index scale using Pi to indicate the percentage of cases that belong to the category (i) of the dependent variable in node (t) (Breiman et al, 2017)

$$G(t) = 1 - \sum_{i=1}^{k} (P_i)^2 \quad (3)$$

Where Pi represents the percentage in the category (value) (i) of the dependent variable in the node or the percentage of cases that fall into that category in node (t)

The degree of reduction of arranged impurities can be calculated by dividing the parent node into two sub-nodes, through the following formula (Saeed et al, 2019):

$$\Delta G(t) = G(t) - G(t_L)\frac{N_L}{N} - G(t_R)\frac{N_R}{N} \quad (4)$$

Since G(t) is an impurity of the parent node, it can be considered as related child nodes and the possibility that the case will move from $G(t_L)$ the left child node to $G(t_R)$ the right child node, respectively.

iv. Stop building the tree

The building process continues until it becomes impossible to continue growing the tree, and the tree stops building (Loh, 2008) .

1 -When there is only one observation in the child nodes

2- All observations in sub-nodes have a similar distribution of prediction variables

3- Using the chi-square χ^2 statistic to measure the degree of deviation in the number of cases between the independent variable section, which differs significantly from the random (weighted) section (Ma, 2018).

v. Pruning CART

Pruning consists of building a large tree T_{max} , producing a sub-tree T_{max} , calculating R(t) for each node t∈ (T_{max})and pruning T_{max} up to its roots and each node large becomes node small (Lewis, 2000) .

$T_{max}, T_1, T_2, \cdots\cdots, \{t_1\}$

The cost-complexity pruning method is used, which depends on the complexity denoted by α, which is gradually increased during the pruning process. The sub-nodes are pruned using the minimum complex error pruning, and the measure of complexity and cost is defined according to the function (Ma, 2018)

$$R_\alpha(T) = R(T) - \alpha\left|\tilde{T}\right| \quad (5)$$

where R(T) is a linear combination between the tree cost and its complexity, α is the complexity cost of the child nodes, and $R_\alpha(T)$ is formed by adding the misclassification of the tree to the node cost.

vi. Choosing the best tree

The goal of choosing the optimal tree is related to the expected performance on the original data set. Where the final tree is selected according to one standard error rule (SE1), and the standard error standard for the risk standard is calculated according to the following formula (Ma, 2018)

$$SE = \sqrt{\frac{R(T)\big(1 - R(T)\big)}{N}} \qquad (6)$$

vii. Verification via Authentication:

The cross-validation procedure in CART is used to choose the optimal tree size for each run. The validation procedure depends on the optimal ratio between tree complexity and classification error as the tree size increases . When the classification error decreases, it becomes 0. And make R(T) the substitution estimate for the misclassification rate, T and |T| It is the leaf node of the tree, and this is achieved through the cost and complexity function (Lewis, 2000) .

$$R_\alpha(T) = R(T) - \alpha\left|\tilde{T}\right| \to \min_T \qquad (7)$$

Where R(T) is tree classification error T: $\alpha(\tilde{T})$ is a measure of complexity that depends on T˜ the total number of leaf nodes in a node.

2.2 Radial Basis Function (RBF)

Network consists of three layers, and it is a network with either one layer or several layers, depending on the data that is entered into the network. Each node in the hidden layer uses the radial basis function (RBF), and it is symbolized by the symbol $\phi(x)$, where the hidden layer converts the data that is entered into non-linear data through the function (RBF), to produce an output layer that is linear **(**Wu et al, 2012 **)** .

2.2.1 Steps network

First: data entry

Second: It is formed by activating the radial basis function to convert the data to non-linear using the radial basis function, and the hidden layer outputs are through the radial activation function ϕ_j using the following Gaussian equation(Liu ., 2013) :

$$\phi_j(x) = \exp\left(-\frac{\|x - c_j\|^2}{2\sigma_j^2}\right) \qquad (8) \qquad where\ j = 1,2,3,\ldots.m\ , \sigma > 0$$

Where , $\phi_j(x)$: The results are for cell j and the output value is between 0-1 , X: the input vector of the network , c_j : central matrix j , σ_j *:* A positive number is called the width, where $\sigma = [\sigma_1, \ldots\ldots, \sigma_m]^T$ m : The number of hidden nodes .

The third step: The network outputs are represented by the following equation (Amin et al, 1998):

$$y_m(x) = \sum_{j=1}^{n} w_{jk}\phi\|x - c_j\| + w_0 \qquad (9) \qquad where \quad j = 1,2\ldots,m$$

Where w_j The weight between the hidden layer and the output layer , where $w = [w_1, \ldots, w_m]^T$ Amount of bias :w_0

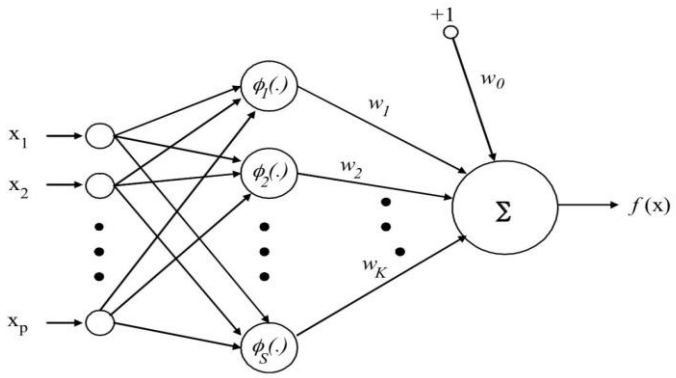

Figure (1) radial basis function

2.2.2 Training in the radial basis function network (RBF)

The training algorithm consists of two stages, the first stage is to find the center cj and the width σ, and the first stage is of the unsupervised learning type, and the second stage is characterized by adjusting the weights wj and is of the supervised learning type, and this is achieved by reducing the cost function, and the error function MSE It can be calculated by the following formula (Wu et al, 2012) :

$$E = \frac{1}{N}\sum_{k=1}^{N}(y_m - y_m(x))^2 \qquad (10)$$

First: The unsupervised learning of the radial basis function

This type of learning is used in the hidden layer, through which the network is trained, the radial basis of the resulting hidden layer, and the main objective of this type of learning is to modify the positions of the center Cj and width σj .

i. Method of Clustering

Clustering is a tool for data analysis and for characterizing the distribution of the data set, and it is used to determine the RBF centers, where the clustering group can be determined through the clustering algorithm. The K-means algorithm is the most used to determine the Cj center (Aik et al, 2019) .

$$c_j = \frac{1}{|m_i|}\sum_{k,x_k \in m_i} x_k \qquad (11)$$ Where $|m_i|$ size mi

ii. Calculate the width of σ

It becomes easy to calculate σ after determining the center cj, and to determine the width of the nucleus we use the nearest neighbor method and then apply the following law to find the width of σ (Benoudjit et al . , 2002).

$$\boldsymbol{\sigma}_j = \left(\frac{1}{P}\sum_{i=1}^{p}\|c_J - c_I\|^2\right)^{\frac{1}{2}} \qquad (12)$$

The closest neighbors to the center Cj .

Where Ci :

Second: Learning is supervised

This type of learning is in the output class, where weights are trained .

i. Weight training

After the outputs of the hidden layer to train the weights after finding the center and the width or the covariance matrix σ is the radial basis function in the hidden layer, and the gradual descent method is used to train the weights for the purpose of reducing the error (Wu, 2012) .

2.2.3 The radial basis network training algorithm

i. Generate initial values of network weights between (0-1)
ii. It is represented by finding the radial basis function by finding its parameters by defining the center Cj and the width σ of the radial activity function
iii. Determine the center through K-means and determine the width of the nucleus σ by finding the nearest neighbor to the center using the P-Nearest Neighbors.
iv. Through which the weights that link the hidden layer and the output layer are trained through the gradual descent method.
v. Make representation by the output layer, and the network outputs are (-1_1)
vi. Calculating the average error MSE for the training sample, and if the training did not reach the lowest MSE, the weight is updated according to the gradual descent method .
vii. Execute the process for a number of iterations.

2.3 Genetic algorithm for training classification and regression tree CRAT-GA

To generate high precision trees, we develop the genetic algorithm for that. Through the structure of the new chromosome and the initial population as well as the function of fitness, selection, hybridization and stopping algorithm. To improve the classification and regression tree structure by hybridizing it with the genetic algorithm , this is done through the following stages (Fu et al, 2003; Ersoy et al, 2020):

Repeat steps for the total number of runs:

i. Establishing an initial population of size N (first generation) through the regression tree.
ii. Determine the parameters of the genetic algorithm, which include population size, crossbreeding and mutation ratio, as well as parental selection and maximum iterations.
iii. Create a random community of size N (CART=N), as well as coding for each tree.
iv. Using the training sample as an initial regression tree
v. Activating the crossbreeding factor: by selecting the parents to produce a new generation.
vi. Adding more of the new generation through crossbreeding with other parents until we have N from the new community
vii. Selection of the chromosome with the least fitness by selecting the greatest probability. And by using one of the selection methods, the roulette wheel generates a random number T_m that is confined between [0,1], so if T_m is achieved, $T_m < m_1 or\, T_m \geq m_{max}$). Each time, one chromosome is identified according to the target function .
viii. Choosing two individuals from the chromosomes for crossbreeding by selecting the parents according to crossbreeding at a point to form the new generation, and repeating for a certain number.
ix. Replacing or changing the new generation if the fitness values are not satisfactory, and binary coding is used for the occurrence of a different mutation, so that a new generation is formed through it J = J + 1, where P = New P is the new community.
Number of genes = total genes * population
x. End the algorithm.
Repeating these steps with the number of activation times to reduce the effect of random distribution, as well as choosing the final solution as the best tree found in the total iterations.

3.4 The Genetic algorithm used to train the radial basis function network RBFN-GA

The main factor in using the genetic algorithm to improve the RBF network is chromosome coding, knowing the fitness function, and building genetic factors. and to determine the optimal RBF-GA of the adaptive regime. To straighten the structure of its network through its weights, and to make it possible to hybridize the genetic algorithm with the neural network, as in the following stages (Jia, 2014; Al-Akidi, 2020):

The first step: Adjusting the network RBF, according to the maximum number of neurons in the hidden layer, by calculating the baseline center Cj by K-means aggregation method, and calculating the width by the nearest neighbor method, to obtain the radial basis function RBF.
The second step: Determine the GA parameters, which include population size, crossbreeding ratio, mutation rate, parental selection mechanism, crossbreeding factor and mutation factor, error target function, maximum replication.
The third step: Randomly initializing the community of size N (network number RBF = N), and encoding the network for each individual.
The fourth step: Using the initial training sample that was configured for the RBF network, which has a size of N.
The fifth step: Calculate the amount of error mse and the number of grid cells for the hidden layer, followed by using the fitness function to know the fitness of each chromosome that corresponds to the entire grid.
The sixth step: According to the fitness function, sorting the chromosome to choose the best fitness for the community, which is denoted by F_b . Check $< E_{min}\ or\ G \geq G_{max}$. If yes, move to the ninth step. If not, move to the seventh step.
The seventh step: Selecting many of the best individuals to keep for the new generation P directly through the fitness function.
The eighth step: Selecting two individuals from the chromosomes for crossbreeding at one point, to form a new generation, repeating this step to reach the maximum number of community size, and the coding will be in this case separately.
The ninth step: The Change of the community from the new generation, using binary coding and part of the real number coding for the occurrence of a different mutation, so that we have a new community P=New P, G=G+1, back to the fourth step
The tenth step: Obtaining the optimal neural network, the end of the repetition of the genetic algorithm, i.e. the optimization stops.

3. Discussion of Results

Simulation is the process of imitating or building a virtual system that is similar to the real world. Simulation is considered an efficient methodology to solve many problems in the real world. Simulation is flexible as we can through testing and training, in addition to generating a large number of samples of different sizes that are difficult to obtain in the applied field of real data then simulation allows control and control over the results (Al-Rudaini, 2019). Monet Carlo method is one of the most simulation methods used because it depends on the generation of random numbers, therefore it is used to compare among the methods using different sizes, where the sizes relied upon are (100,200,400) to generate the initial values, where the data was generated on the basis of restricted generation by relying on the limit The lowest and highest of the original data and generate it randomly within the permissible limit, through the following formula (Rashid, 2014):
U=a+(b-a)*rand(n,m)
Where: a: the minimum limit of the values of variable , b: the upper value of variable m : the number of variables to be generated within the limits , n : the size of the sample.

3.1 Comparison criteria

Criteria by which classification methods are compared (Ahmed, 2016)
First: Error of Classifications

$$Error\ of\ Classification = \frac{(N - sum(diag(cm)))}{N} \qquad (13)$$

Second: Root Mean Square Error

$$RMSE = \sqrt{\frac{(y - \hat{y})^2}{N}} \qquad (14)$$

Third: the mean absolute relative error (AlAkidi, 2020) .

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}|}{y}}{n} \quad (15)$$

3.2 The Analysis and Interpretation

To apply the comparison methods, the infection rate (dependent variable) was classified into infected and non-infected through the following table:

Table Classification (1)

Classic	Code
Y<0.50	0
y>=0.50	1

Accordingly, the following results were obtained:

Table (2) comparison criteria CART

Training Data=80% of Data	Test Data = 20% of Data			
N	method	classification of error	RMSE	MAPE
100	CART	0.4500	0.6333	0.6098
	CART – GA	0.4500	0.108	0.0386
200	CART	0.4750	0.6779	0.4677
	CART – GA	0.3500	0.1017	0.0170
400	CART	0.4500	0.6045	0.5199
	CART – GA	0.4250	0.0641	0.0099
Training Data=70% of Data	Test Data = 30% of Data			
N	method	classification of error	RMSE	MAPE
100	CART	0.5000	0.6546	0.5646
	CART – GA	0.4000	0.1105	0.0204
200	CART	0.4500	0.6673	0.7697
	CART – GA	0.3167	0.0850	0.0123
400	CART	0.4667	0.6553	0.7542
	CART – GA	0.4167	0.0457	0.0053

Table (3) Comparison Criteria RBF and RBF –GA

N	method	Classification of Error	RMSE	MAPE
100	RBFN	0.4400	0.5845	0.5943
	RBFN – GA	0.4400	0.0824	0.000135
200	RBFN	0.4400	0.5999	0.4128
	RBFN – GA	0.4150	0.0833	0.00062
400	RBFN	0.4750	0.7396	0.1757
	RBFN – GA	0.3875	0.0803	0.000963

We notice from the above tables when dividing the data, 80% training data and 20% test data in the regression tree method compared to the radial basis function network method. The results were better in RBF than CART at size (100, 200) through the comparison criteria used.

CART was better at size 400 through the criteria of classification error and the mean square root. As for the average relative absolute error ,RBF better than, classificion regression tree. We also note that the use of the genetic algorithm led to an improvement in the results in both methods.

- when dividing the data in the CART method into 70% training data and 30% test data, we notice that at size 100, 200)) RBF was better through the three criteria, while at size 400 the CART classification error was less than RBF, either when using the algorithm To improve the genetic results of CART, the classification error was better at size 200,100 in CART than RBF-GA, and when improving the results of RBF using the genetic algorithm, the classification error at size 400 was less than CART-GA.

We will classify the dependent variable (infection rate) into three levels (uninfected, early infected, infected) for comparison between the criteria. The following table shows the classification:

Table (4) Classification

Classification	Code
y>=0.7	2
0.4<=y>0.70	1
Y<0.4	0

After applying the programming on the methods, we get the following results:

Table (5) comparison criteria CART

Training Data=80% of Data Test Data = 20% of Data				
N	method	classification of errr	RMSE	MAPE
100	CART	0.5000	0.7294	0.4306
	CART – GA	0.4500	0.1488	0.0392
200	CART	0.6000	0.9941	0.4226
	CART – GA	0.4000	0.1411	0.0143
400	CART	0.6375	0.9109	0.5688
	CART – GA	0.5875	0.0892	0.0091
Training Data=70% of Data Test Data = 30% of Data				
N	method	Classification of error	RMSE	MAPE
100	CART	0.5000	0.9874	0.4760
	CART – GA	0.4333	0.1717	0.0226
200	CART	0.6500	1.1382	0.8065
	CART – GA	0.3833	0.1451	0.0132
400	CART	0.5833	0.9963	0.6010
	CART – GA	0.5500	0.0759	0.0052

Table (6) comparison criteria RBF and RBF-GA

N	method	Error of classification	RMSE	MAPE
100	RBFN	.0.5300	0.7980	0.6276
	RBFN – GA	0.5300	0.0824	0.000135
200	RBFN	0.5500	0.8022	0.5179
	RBFN – GA	0.5100	0.8340	0.000612
400	RBFN	0.6450	0.8003	0.3221
	RBFN – GA	0.6275	0.0803	0.00096

We notice through the table after dividing the response variable into three levels when the training data is 80% and 20% test data in the regression tree, the classification error at the size 100 and 400 is better than it is in the radial basis function network, but when the size is 200 the classification error is In RBF it is better than in CART. As for the other criteria, we note its superiority at size 100 in CART, but at size 200,400 it is lower in RBF than in the classification regression tree. Using the genetic algorithm for both methods gave lower results than it was, as its use led to an improvement in classification.

When dividing the data of the classification regression tree method, 70% training data and 30% test data, we notice at sizes 100 and 400 that the classification error in CART is better than it is in the radial basis function network, but when the size is 200, the classification error in RBF is better than it is in CART. As for the other criteria we notice that at the size of 200 , 400 100, it is lower in the RBF than in the classification regression tree, and using the genetic algorithm for both methods gave lower results than it was, as its use led to improving the classification.

4. Conclusion

Through the results obtained through simulation, we conclude that the experimental results of the methods are good in terms of classification, as it gives few results for classification error and that the radial basis function network is superior to the classification regression tree at certain sizes, as for the other two criteria, we note that the basis network radial is better than regression tree. Added the genetic algorithm to the methods leads to improved results for both methods.

References

1-Ahmed, Suhad Ahmed ., 2016, The use of classification trees and logistic regression in estimating an aggregate model and comparing them with a practical application, PhD thesis, University of Baghdad - College of Administration and Economics.

2-Aik, L. E., Hong, T. W., and Junoh ., A. K. 2019. Distance weighted K-Means algorithm for center selection in training radial basis function networks. International Journal of Artifical Intelligence,vol 8 , no 1, pp 54-62.

3- Al-Rudaini, Sarah Adel Mazloum ., 2019. Using the Genetic Algorithm to Estimate the Parameters of the Binary Logistic Regression Model with Application, Master Thesis, University of Baghdad - College of Administration and Economics.

4-Al-Akidi, Mounir Shehab Ahmed., 2020 "Data mining in time series with practical application" Master's thesis - University of Baghdad - College of Administration and Economics. Hidden class - genetic algorithm

5-Amin, S. M., Rodin, E. Y., Liu, A. P., Rink, K., and García-Ortiz, A., 1998. Traffic prediction and management via RBF neural nets and semantic control. Computer-Aided Civil and Infrastructure Engineering, 13(5), 315-327.

6-Anto, S., and Chandramathi , S ., 2015., An expert system based on SVM and hybrid GA-SA optimization for hepatitis diagnosis . International Journal of Computer Engineering in Research Trends ,vol 2 no7, pp 437-443

7 -Bedossa, P., and Poynard, T., 1996. An algorithm for the grading of activity in chronic hepatitis C. Hepatology, vol 24 no2, pp 289-293.

8- Benoudjit, N., Archambeau, C., Lendasse, A., Lee, J. A., and Verleysen, M., 2002. Width optimization of the Gaussian kernels in Radial Basis Function Networks. In ESANN Vol. 2, pp. 425-432.

9- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., 2017. Classification and regression trees. Routledge. Printed in the United States of America .

10- Loh, W. Y. (2008). Classification and regression tree methods . Encyclopedia of statistics in quality and reliability, vol 1, pp315-323.

11-Ersoy, E., Albey, E., and Kayış, E., 2020. A CART-based genetic algorithm for constructing higher accuracy decision trees. In DATA 2020-Proceedings of the 9th International Conference on Data Science, Technology and Applications. SciTePress

12 -Fu, Z., Golden, B. L., Lele, S., Raghavan, S., and Wasil, E. A., 2003. A genetic algorithm-based approach for building accurate decision trees. INFORMS Journal on Computing, vol 15 no 1, pp3-22 .

13 -Ismail , Ibtisam Adel., Mahmoud, Abdul Karim Abdullah, and Al-Hasnawi, Jawad Kazem., 2019., Temporal and demographic trends of viral hepatitis disease in Najaf province. Geographical Research Journal, vol 1 no 29 , pp 137-170 .

14-Jia, W., Zhao, D., Shen, T., Su, C., Hu, C., and Zhao, Y., 2014. A new optimized GA-RBF neural network algorithm. Computational intelligence and neuroscience, vol 2014, pp 44-44.

15- Karthikeyan, T., and Thangaraju, P., 2013. Analysis of classification algorithms applied to hepatitis patients. International Journal of Computer Applications, vol 62, no15 .

16- Khatoon, M., and Agarwal, A. K .,2019. Diagnosis of Hepatitis Virus using Decision Tree and Genetic Algorithm. International Journal of Computer Applications,0975, 8887,vol 117,no13 , pp70-72

17- Kurosaki, M., Matsunaga, K., Hirayama, I., Tanaka, T., Sato, M., Yasui, Y., ... and Izumi, N., 2010., A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis. Hepatology research, vol 40 no 3, pp 251-260.

18- Lewis, R. J., 2000,. An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California Vol. 14. California: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

19- Liu, Jinkun ., 2013 " Radial Basis Function (RBF) Neural Network Control for Mechanical Systems " Springer Science &Business Media , Springer Heidelberg New York Dordrecht London .

20- Ma,Xin .,2018,. Using Classification and Regression Trees A Practical Primer , INFORMATION AGE PUBLISHING, INC , Printed in the United States of America .

21-Mahdieh, A., and Hassan, Z., 2011. Automatic disease diagnosis systems using pattern recognition based genetic algorithm and neural networks. International Journal of Physical Sciences, vol 6 no25, pp 6076-6081.

22- Obaido, G., Ogbuokiri, B., Swart, T. G., Ayawei, N., Kasongo, S. M., Aruleba, K., ... and Esenogho, E., 2022. An interpretable machine learning approach for hepatitis b diagnosis. Applied Sciences, vol 12 no 21, pp 11127.

23-Rashid, Hossam Abdel-Razzaq ., 2014. Non-parametric smoothing of the variable and partial variable coefficients model. PhD thesis, University of Baghdad - College of Administration and Economics .

24- Saeed, H. M., Mhamad, A. J., and Ahmed, R. A., 2019. Using CART Model to Classify Family's Monthly Expenditure According to Urban and Rural Area in Sulaimani Governorate You are here , The Scientific Journal of Cihan University- sulaimanyia , vol (3) ,issue(1),pp130-150 .

25 -Siahaan, W. F., Sitompul, O. S., and Situmorang, Z., 2020., The Accuracies of ANFIS and Genetic Algorithm with Tournament Selection on Classifying Hepatitis Data. In Journal of Physics: Conference Series Vol. 1566, No. 1, p. 012121. IOP Publishing.

26- Singh, A., and Pandey, B., 2014., Intelligent techniques and applications in liver disorders: a survey. International Journal of Biomedical Engineering and Technology, vol16 no1, pp27-70.

27 - Syafa'ah, L., Zulfatman, Z., Pakaya, I., and Lestandy, M., 2021., Comparison of machine learning classification methods in hepatitis C virus. Jurnal Online Informatika, vol 6 no 1, pp 73-78.

28- Wu, Y., Wang, H., Zhang, B., and Du, K. L., 2012., Using radial basis function networks for function approximation and classification. International Scholarly Research Notices, vol 2012.

خوارزمية الذكاء الاصطناعي لتحسين دقة التشخيص لنوع التهاب الكبد

الباحث /هبه هذال خليل [1]
جامعة بغداد/ كلية الإدارة والاقتصاد/ بغداد، العراق
heba.hathal1201a@coadec.uobaghdad.edu.iq

أ.د. صباح منفي رضا [2]
جامعة بغداد / كلية الإدارة والاقتصاد / بغداد ، العراق
drsabah@coadec.uobaghdad.edu.iq

Received:26/2/2023 **Accepted:26/3/2023** **Published: March / 2023**

مستخلص البحث

يعد التهاب الكبد من الأمراض التي أصبحت أكثر تطورًا في السنوات الأخيرة من حيث ارتفاع عدد الإصابات. التهاب الكبد يسبب التهاباً يدمر خلايا الكبد ، ويحدث نتيجة للفيروسات والبكتيريا وعمليات نقل الدم وغيرها. هناك خمسة أنواع من فيروسات التهاب الكبد وهي (A,B, C,D ,E) وشدتها. يختلف المرض حسب النوع. التشخيص الدقيق والمبكر هو أفضل وسيلة للوقاية من المرض ، حيث يتيح للمصابين اتخاذ خطوات وقائية حتى لا ينقل المرض إلى الآخرين ، والتشخيص باستعمال الذكاء الاصطناعي يعطي نتيجة تشخيصية دقيقة وسريعة.

حيث اعتمدت الطريقة التحليلية للبيانات على شبكة الأساس الشعاعي لتشخيص المرض ، بالإضافة إلى استعمال التصنيف و شجرة الانحدار وفضلاً عن استعمال الخوارزمية الجينية لتحسين دقة التصنيف لكلتا الطريقتين , ومقارنة الأساليب المستخدمة لاكتشاف أكثر طرق التصنيف كفاءة من خلال معايير. خطأ في التصنيف ، ومتوسط خطأ الجذر التربيعي ، ومتوسط الخطأ النسبي المطلق ، واستنتجت النتائج التجريبية أن الطرق جيدة من حيث التصنيف ، وأن شبكة الأساس الشعاعي تفوقت على شجرة انحدار التصنيف ، حيث كان خطأ التصنيف أقل . وأن إضافة الخوارزمية الجينية أدت إلى تحسين دقة التصنيف.

نوع البحث: ورقة بحثية.

المصطلحات الرئيسة للبحث: شجرة الانحدار التصنيفية , شبكة دالة الاساس الشعاعي , الخوارزمية الجينية

[1] بحث مستل من رسالة ماجستير