IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111207

Diagnosis of Liver Fibrosis using RBF Neural Network and Artificial Bee Colony Algorithm

Mohammad Ordouei[1], Touraj Banirostam[2]

Computer Engineering Dep, Islamic Azad Univerity, South Tehran Branch, Tehran, Iran[1]

Computer Engineering Dep, Islamic Azad Univerity, Central Tehran Branch, Tehran, Iran[2]

Abstract: Liver turquoise is one of the silent and dangerous diseases. If it can be detected in the early stages, the lives of many affected people can be saved. Providing smart methods to identify and diagnose this disease can save patients' lives in addition to reducing medical costs and overheads. In this research, an innovative method using a three-layer radial basis neural network is proposed as a multi-class method for diagnosing liver fibrosis. To increase the accuracy and efficiency of the pre-processed data, the data are balanced using the SMOT method. Also, feature selection is done with the bee algorithm. In this way, the desired features are first reduced using the bee algorithm. For this purpose, a mapping of features is done using the bee algorithm. Then the data with reduced features are applied to the proposed RBF network. The simulation results show that the proposed method is 5% more accurate than similar methods.

Keywords: Liver Fibrosis Diagnosis - Feature Selection – Artificial Bee Colony - Radial Basis Function (RBF) neural network.

I. Introduction

Long-term liver damage or inflammation can cause excessive amounts of scar tissue in the liver and liver fibrosis. In addition, most chronic liver diseases can eventually lead to fibrosis. Unlike healthy liver cells that can repair themselves, scar tissue cells do not have this ability and cannot function properly [1].

Liver fibrosis treatments include treating infections, making lifestyle changes, and taking certain medications, which can often reverse the damage caused by mild to moderate liver fibrosis. Fibrosis is the first stage of liver injury. Later, if most of the liver becomes scarred, it is known as liver cirrhosis [2]. Early diagnosis of fibrosis is the most important factor in preventing mortality from chronic liver diseases. Therefore, medical decisions in the field of liver diseases are largely based on the diagnosis of the exact stages of fibrosis, and doctors usually consider early predictions to arrive at a personalized management algorithm [3]. Diagnosing fibrosis using biopsy is associated with problems, such as its high cost, painful complications, and errors in sampling [4]. Another approach to diagnose the stage of fibrosis is to use machine learning methods.

In this research, the diagnosis of liver fibrosis using the RBF neural network and the bee colony algorithm has been discussed. In the second part of this article, related works are reviewed. In the third part, the approach of the proposed method is described. In the fourth part, the used data set described in the fifth part, the simulation results of the proposed approach and its comparison with other methods are made. The sixth part provides the final conclusion.

II. Related Work

In [5], fatty liver disease was predicted using machine learning algorithms on 577 patients whose primary fatty liver screening was done in Taipei city hospital. The goal was to create a machine learning model to predict fatty liver, which prediction accuracy was obtained for random forest method 87.84, Bayesian 65.82, artificial neural network 81.85 and logistic regression 96.76%.

In 2019, a new diagnostic method for cirrhosis based on X-ray absorption spectroscopy was investigated which used the liver of healthy and cirrhotic mice as samples and used support vector machine and neural networks for diagnosis and had a diagnostic power of 5.99% [6].

In another study in 2021, on the same data set, with two approaches without SMOTE and after applying it, the accuracy of the algorithms was checked, and the SMOTE method was used to equate categories from unbalanced data sets. According to the results, the application of SMOTE in the machine learning models used except for the Bayesian network has been associated with an increase in the accuracy of diagnosis [7].

In [8], he used the digital images of patients' biopsies and the severity of the difference in fibrosis as data. These biopsies covered the spectrum and severity of fatty liver and stages of fibrosis. The result was a model that could identify the zero, third and fourth stages of fibrosis with an accuracy of more than 90% and the first and second stages of fibrosis with an accuracy of 78 to 86%.

In a study in 2020, research was conducted on a dataset of patients to predict hepatocellular carcinoma (HCC)1, which is considered one of the most malignant liver diseases and requires non-invasive investigation. The result of the investigation was the alternating decision tree algorithm (ADTree) with an accuracy of 6.95% and the linear regression algorithm with an accuracy of 2.93% [9].

In 2021, [10] used k-nearest neighbor, simple Bayes, neural network and random forest. The aim of this study is to evaluate the level of accuracy using machine learning classification algorithm to diagnose HCV disease. This study was conducted by researching a dataset of more than 600 patients with 14 characteristics. Prediction using neural network in this research also had an accuracy of 24.90%. The predictive power of random forest was equal to 31.94%.

In [11], a further investigation was conducted on the blood test results. The investigation on the data includes the SGDText1 approach of random gradient descent with the best result of 38.88%, compared to other models, it has a higher accuracy. The simple linear regression2 method gives the lowest accuracy compared to other models, which is 60.02%.Other algorithms such as simple logistics3, sequential minimum optimization (SMO)4, SMOreg5 were also investigated, which had an accuracy of 93.60%, 23.69% and 14.80%, respectively.

III.PROPOSED METHOD

In this research, in the first stage, with the help of pre-processing methods, the data is prepared to enter the desired algorithm. Then the training and testing data are separated and then the data is applied to the classification algorithm used in the research (Radial Basis Function (RBF) neural network) and the results are checked.

The steps of the research are according to figure (1):

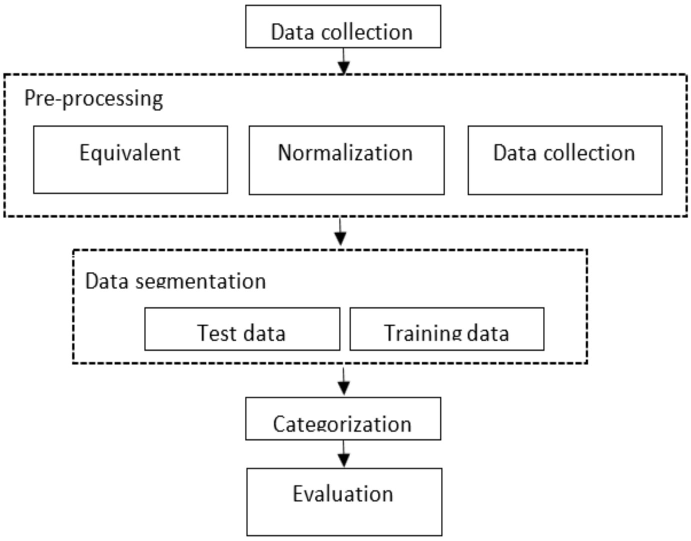

Fig. 1 Research States

Also, in order to validate the results, the accuracy criterion1 is checked with formula (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

(TP, TN, FN and FP represent true positive, true negative, false negative and false positive, respectively.)

A. Pre-processing

In this step, the data are examined from several aspects, which are fully addressed. Before connecting the data set to the proposed model, pre-processing includes several steps to check the comprehensive suitability of the data set. These steps include replacing unknown data, normalization, equalization and feature selection.The data are replaced by the average value of the characteristic of that data according to equation 2.

$$mean = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2)$$

The min-max method is also used for normalization based on to equation 3.

$$\acute{v} = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \qquad (3)$$

One of the challenges of using data mining methods is the imbalance of the data set. To solve this problem, they increase the number of samples of the minority class or decrease the number of samples of the majority class [12].To solve this problem, the SMOTE method was used in this research. Feature selection means that less important features that have less impact than other features on the output of the algorithm are removed.

In this research, the Artificial Bee Colony (ABC) algorithm was used to select the feature. In the ABC, the position of each food source in the N-dimensional space expresses a possible solution of the problem space, and the amount of nectar of each food source determines the quality and merit of the solution found. The search process in the ABC is divided into three parts: the worker bees search phase, the watcher bees phase, and the scout bees phase.

for the feature selection, a binary string of length N is considered, where N is the total number of primary features.
If feature i is present, the degree of i in the corresponding binary string is equal to "1", and if the sentence i is omitted in the corresponding solution, the degree of i is equal to "0".

$$\text{Solution}_i = \begin{cases} \text{If feature i exists in the solution} & 1 \\ \text{If feature i is deleted in the solution} & 0 \end{cases} \qquad (4)$$

At first step, the bees in the colony are divided into two categories: Worker bees and watchful bees. Worker bees first start to search without any knowledge of the space around the hive and choose the initial solutions randomly, and keep the location of their food source in their memory.

In each iteration, each worker bee chooses a food source in the neighborhood of its previous solution. After all the worker bees have completed their foraging process, they share their information about food sources with the bees in the hive.

If the amount of nectar or the quality of the current food source is higher than the previous source, the bee remembers that new food source and forgets the previous solution, otherwise, it leaves the same previous solution in its memory (search phase of worker bees). The amount of nectar related to each food source corresponds to the quality of the problem solution expressed by that source, which is calculated from equation (5).

$$Nec_i = 1/cost_i \qquad (5)$$

$$cost = w_1\left(\frac{\sum_{i=1}^{N} Solution_i}{N}\right) + w_2(Error) \qquad (6)$$

In relation (5), Nec_i is the amount of nectar of food source i and cost_i is the value of the objective function of bee i's solution. The first term in the objective function expresses the percentage of selected features, and the second term expresses the classification error for that bee. w_1 and w_2 are two constant coefficients that adjust the influence ratio of the two terms in the overall objective function.

After determining the nectar of the food sources found by the worker bees, the recruitment process is carried out. In each repetition, a number of worker bees who have found the largest amount of nectar are selected and watchful bees are given to them as soldiers.Then these soldiers start searching in the neighborhood of the food sources with the most nectar so that they can obtain solutions with more suitable quality in the neighborhood of the previous sources (searching phase of watchful bees). If it is assumed that the number of selected worker bees is equal to M, the number of soldiers assigned to each of these worker bees is obtained from equation (7):

$$N_k = \frac{Nec_k}{\sum_{i=1}^{M} Nec_j} . N_o \tag{7}$$

where N_k is the number of watcher bees that are placed as soldiers to search in the neighborhood of the solution of the chosen bee k. And Nec_k and Nec_j are the amount of nectar of food source k and i. Also, N_0 is the total number of watchful bees. In each iteration, worker bees whose search has been ineffective (did not improve) in the last few iterations, become scout bees. Then each scout bee chooses a random solution without any knowledge of the surrounding space of the hive (scout bee search phase). It is done in this way to search for the neighborhood for worker bees and guard bees that in order to search for a new solution in the neighborhood of a previous solution, according to the binary structure of the problem, the probability that a change will occur in one dimension of the previous solution is equal to a parameter called P_{change}, And the value of this parameter has been determined as a linear variable from the value of $P_{change\ max}$ to the value of $P_{change\ min}$ during the execution of the program.

$$P_{change} = P_{change_{max}} + (iter\ / iter_{max}).(P_{change_{min}} - P_{change_{max}}) \tag{8}$$

In relation (8), itermax is the total number of iterations of the algorithm and iter is the number of the current iteration.

IV. CLASSIFICATION WITH RADIAL BASIS FUNCTION (RBF) NEURAL NETWORK

The RBF network has a three-layer structure including input layer (X), hidden layer (H) and output layer (Y) [13].The main architecture of the three-layer RBF neural network is shown in Figure (4). The input layer consists of source nodes that connect the network to its environment and then pass the inputs to the hidden layer without applying weights. In the hidden layer, weights are assigned to the data using the nonlinear function of radially symmetric kernel functions. The hidden layer consists of a number of neurons (nodes) and a parameter vector called the center. The nodes in the hidden layer each have an activation function that changes the input signal (data) in terms of weight [14]. The activation function calculates the Euclidean distance between the input vector and the center of the hidden unit in the grid. Finally, the output layer is a summation unit and operates linearly.

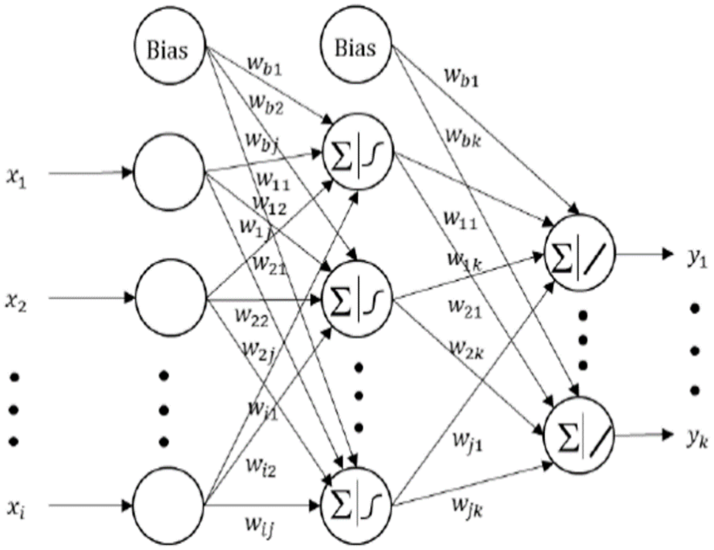

Fig. 2 Proposed three-layer RBF

DOI: 10.17148/IJARCCE.2022.111207

V. INTRODUCING THE DATASET

The dataset used in this research work was extracted from the University of California Irvine 1data repository at https://archive.ics.uci.edu [15]. This dataset includes information collected from 615 applicants for liver disease or hepatitis C test with 12 features. 12 features are considered as input and one feature is output. The output includes five categories: healthy status or blood donor, suspicious status, hepatitis, liver fibrosis and cirrhosis. The full description of the features of the dataset can be seen in Table I.

TABLE I FEATUERS OF DATASET

NO.	Attributes	7	BIL (Bilirubin)
1	Age(in years)	8	CHE (Acetylcholinesterase)
2	Sex(male=1, female=2)	9	CHOL (Cholesterol)
3	ALB (Albumin Blood Test)	10	CREA (Creatinine)
4	ALP (Alkaline phosphatase)	11	GGT (Gamma-Glutamyl Transferase)
5	ALT (Alanine Transaminase)	12	PROT (Proteins)
6	AST (Aspartate Transaminase)	Class	Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis')

To remove the unknown data, the unknown values are replaced with the average value of the feature of that data. The min-max method was also used for normalization. In addition to this, the data set examined in this research had the problem of the disproportion of the size of the categories. There are 533 healthy people, 7 suspected people, 24 people with hepatitis, 21 people with fibrosis and 30 people with cirrhosis. In order to balance the number of categories so that a more accurate result can be obtained from the algorithm, the SMOTE approach has been applied to the data.

The selection of features by the ABC in this research is such that a function is defined whose goal is to maximize accuracy. It is given 12 features as input and the output of the function is an array of length 12 representing the first to twelfth features. A zero of each array house indicates the deletion of that feature number and a one indicates the selection of that feature. The initial population value for this algorithm is 100.

After applying the feature selection algorithm, the second, sixth, eighth and tenth features are removed and the remaining features are selected. In fact, it has kept 8 out of 12 features.

In this study, the K-fold cross validation method [16] is used with K equal to 10. During this process, the dataset is divided into 10 parts and run 5 times to train all the data. The RBF neural network in Figure 4 is used for classification. To investigate the effect of blood indicators on the rate of liver complications, in the form of manual feature selection, blood features including blood gamma glutamy transferase which is a type of blood enzyme, blood protein, blood creatinine and blood cholesterol were selected and worked on.

VI. RESULTS OF THE PROPOSED METHOD SIMULATION

After implementing the proposed research approach, the result in terms of the accuracy of liver fibrosis stage diagnosis was compared with previous works that have worked on the data set used in this research. This comparison can be seen in Fig. 3. It should be considered that the total number of features of this data set is 12, and any model that uses 12 features means working on all the features of the data set. As shown in Figure 3, the application of SMOT increased the accuracy by 4%. Furthermore, the proposed approach using SMOT has higher accuracy than other methods.

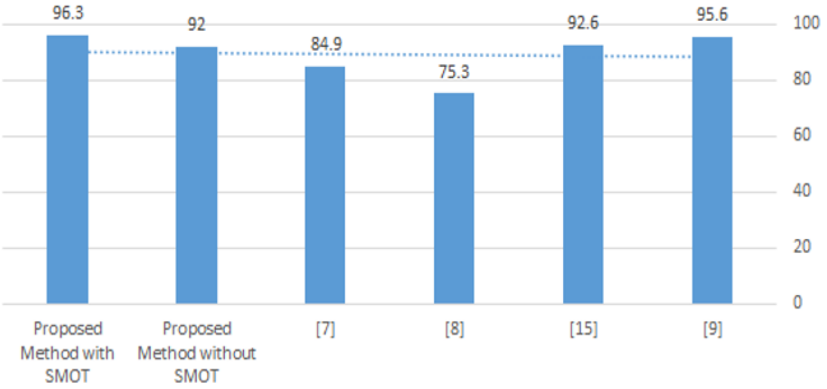

Fig. 3 Comparison of the simulation results of the proposed approach with other methods

VII. Conclusion

In this research, a method based on the RBF neural network and the bee colony algorithm was presented to detect the stage of liver fibrosis without invasive methods and by using the most available patient information. To diagnose the disease, the collected information of 615 patients was used, and the bee colony algorithm was used to select the features of the pre-processing stage. During this feature selection, the collected features of patients were reduced from 12 features to 8 features, which significantly reduced the cost and complexity of the model. At the end, a comparison was made with some previous works. According to the obtained results, the combined method used in this research was more accurate than the compared methods.

References

[1]. Ömer Kayaaltı et al., “Liver Fibrosis Staging Using CT Image Texture Analysis and Soft Computing”, Applied Soft Computing Journal Vol. 25, 2014, Pages 399-413.

[2]. Dawei Yang et al., “Systematic review: The diagnostic efficacy of gadoxetic acid-enhanced MRI for liver fibrosis staging”, European Journal of Radiology, Vol.125, □□□□□□□□□□□□.

[3]. Banihashem, E., and Banirostam, T., 2017, “Diagnosis of Breast Cancer by Combining the Techniques of Data Mining and Artificial Immune System”, International Journal of Computer Science and Network, Volume 6, Issue 5, Pages 539-546.

[4]. Mohammed Eslam et al., “FibroGENE: A gene-based model for staging liver fibrosis”, Journal of Hepatology, Vol. 64, 2016 Pages 390-398.

[5]. Chieh-ChenWu et al., ”Prediction of fatty liver disease using machine learning algorithms”, Computer Methods and Programs in Biomedicine , Vol. □□□□□□□□□□ Pages □□□□□.

[6]. Zheng Fang et al., ” X-ray absorption spectroscopy combined with machine learning for diagnosis of schistosomiasis cirrhosis”, Biomedical Signal Processing and Control, Vol. 60,2020,101944.

[7]. Oladosu Oyebisi Oladimeji et al., “Machine learning models for diagnostic classification of hepatitis C tests”, Frontiers in Health Informatics, Vol. □□□□□□□□□.

[8]. Samar Gawrieh et al., ”Automated quantification and architectural pattern detection of hepatic fibrosis in NAFLD”, Annals of Diagnostic Pathology, Vol. □□□□□□□□□□□□□□□□.

[9]. Somaya Hashem, et al., “Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease” , Computer Methods and Programs in Biomedicine, Vol.□□□, November □□□□□□□□□□□□□

[10]. Syafa’ah, Lailis, et al. "Comparison of machine learning classification methods in hepatitis C virus." Jurnal Online Informatika 6.1 (2021): 73-78. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[11]. Sivachandran M, Krishnakumar T, “An Analysis of blood donors and Hepatitis C patients by using big data techniques”. Proceedings of the First International Conference on Computing, □□□□.

[12]. Piervincenzo Ventrella et al., “Supervised machine learning for the assessment of Chronic Kidney Disease advancement”, Computer Methods and Programs in Biomedicine, Vol.□□□□□□□□□□□□□□□□□□

[13]. Abedini, M., Bijari, A. and BaniRostam, T., 2020, "Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network", International Journal of Advanced Research in Computer and Communication Engineering, Volume 9, Issue 7, Pages 1-4.

[14]. Yousefian, F., Banirostam, T. and Azarkeivan, A., 2017, "Prediction Thalassemia Based on Artificial Intelligence Techniques: A Survey", Int. J. of Advanced Research in Computer and Communication Engineering, Volume 6, Issue 8, Pages, 281-287.

[15]. Nasr, Mahmoud, et al. "A novel model based on non invasive methods for prediction of liver fibrosis." 2017 13th International Computer Engineering Conference (ICENCO). IEEE, 2017.

[16]. Sebastian Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning”, University of Wisconsin–Madison, □□□□□