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 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12701

Identification of hepatitis disease by combining decision tree algorithm and Harris Hawks Optimization (HHO)

Mohammad Ordouei[1], Mastooreh Moeini[2]

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

Computer Engineering Dep, Islamic Azad University, South Tehran Branch, Tehran, Iran[2]

Abstract: Hepatitis is one of the most common diseases in the world and any early diagnosis can save the lives of many people suffering from this disease. The purpose of this research is to diagnose hepatitis disease using the combined model of the decision tree algorithm and Harris Hawks Optimization.

In this research, the diagnosis of hepatitis disease was made using the decision tree and evolutionary algorithm of Harris Hawks Optimization . HHO algorithm is a population-based and gradient-independent optimization technique. The main idea of the HHO algorithm is the cooperative behavior and chasing style of Harris's falcon in nature, which is known as surprise attack. The effectiveness of the proposed HHO optimizer method, compared to other nature-inspired techniques, was tested on 29 functions and several real-world engineering problems were investigated. The statistical results and comparisons show that the HHO algorithm has very promising and sometimes competitive results compared to other well-known meta-heuristic techniques [6].

Keywords: Decision Tree, Evolutionary Algorithms, Data Mining, Harris Hawks Optimization (HHO)

I. INTRODUCTION

Hepatitis is an inflammatory liver disease. This disease is usually caused by a viral infection, but other possible causes also play a role in the development of hepatitis. These include autoimmune hepatitis and other types of hepatitis that occur as a result of secondary effects of drugs, toxins, and alcohol. Autoimmune hepatitis is a disease that is caused by the production of antibodies by the body against the liver tissue [10]. Another important point about this virus is that it gradually damages the liver. A healthy liver makes the chemicals the body needs and removes toxins from the blood. But when a person gets this disease, the liver becomes inflamed and its normal tissue is destroyed, leaving worn-out tissue instead. [18-10]

The age, weight and gender of the patient are important factors of hepatitis. Timely and early diagnosis of this disease can be effective in its treatment and prevent this disease from turning from acute to chronic Because when this disease becomes chronic, the risk of liver cirrhosis and liver cancer increases And it may involve the patient for years and eventually lead to the death of the patient.

The use of artificial intelligence can be effective in early diagnosis of this disease. With its timely diagnosis, additional costs for the disease can be avoided, as well as the process of treatment and recovery of the patient can be facilitated, and the risks of disease progression and chronicity can be prevented. In this research, by using the decision tree and evolutionary algorithms, it is tried to overcome the subsequent problems caused by the disease by timely diagnosis [13-20].

2-REQUIRED DEFINITIONS- HEPATITIS

Hepatitis means inflammation in the liver and it can be caused by various reasons, some of which are contagious and some are not. Among the factors that cause hepatitis, we can mention excessive alcohol consumption, the effect of some drugs, contamination with bacteria and also viruses. Viral hepatitis leads to liver infection. The cause of hepatitis disease is a virus and at first it can appear like a cold; But chronic hepatitis C disease, unlike common cold, can threaten the patient's life due to liver failure and difficulty in treatment. Most people with hepatitis C and B have no symptoms [15-12].

DOI: 10.17148/IJARCCE.2023.12701

2-1 Evolutionary Algorithms

Evolutionary algorithms are a stochastic, production-based and experimental approach to solving optimization problems. Harris Hawks Optimization is one of the types of evolutionary algorithms that researchers pay attention to. Harris Hawks Optimization or HHO is a new meta-heuristic algorithm that was published in 2019 [14]. The origin and emergence of meta-heuristic algorithms goes back to the genetic algorithm, which was proposed as a mathematical algorithm to solve optimization problems based on Darwin's theory of evolution. Since then, many evolutionary and meta-heuristic algorithms have been introduced by studying and researching various behaviors in nature.

In HHO, a population-based and nature-inspired optimization paradigm, called Harris Hawk Optimizer (HHO), is presented. The main idea of the HHO algorithm is the cooperative behavior and hunting style of the Harris hawk in nature, which is known as surprise attack. First, the falcons scatter on the trunks of trees or tall bushes to find the prey, then they chase the prey to tire it out and finally surround and trap the prey.

2-2 Decision tree

Classification trees are used to classify a set of records and are commonly used in marketing, engineering and medical activities. In the decision tree structure, the prediction obtained from the tree is explained in the form of a series of rules. Each path from the root to a leaf of the decision tree expresses a rule, and finally, the leaf is labeled with the class in which the largest amount of records is assigned.

3. Research background

In 2017 Professor Manugaran et al used a Bayesian hidden Markov model with Gaussian clustering to model genome-wide DNA copy number variation. The proposed Bayesian model with the Gaussian clustering approach has been compared with various existing methods such as the precise time method of pruning, the binary separation method, and the fragment neighborhood method. The experimental results show the effectiveness of the proposed algorithm and the accuracy is 86% [9-21].

In 2016, Prof. Doi et al investigated the automatic diagnosis of breast cancer based on machine learning algorithm using decision tree. The proposed approach has three stages of a process. In the first step, the data is grouped into a number of clusters using the clustering algorithm. In the second step, outlying distances from breast cancer data are identified using a detection algorithm. In the third step, it is determined whether the cancer is benign or malignant from the beginning of the pre-processed data set using the classification algorithm. In this research, the Wisconsin breast cancer dataset was used and the accuracy was 99.6% [5-19].

In 2015, Professor Lee and colleagues provided a better understanding of the hepatitis C virus life cycle, including general viral properties and proteins. This effort will facilitate the development of sensitive and effective antiviral diagnostic tools. Current treatment is serologic screening testing in high-risk individuals, and nucleic acid testing is recommended to confirm active infections. It is possible that a new gene-free anti-viral anti-hormone treatment will be available within the next few years [4-17].

In 2016, Professor Harris et al investigated the delay in diagnosis among hepatitis C patients in England. The main method of data collection was face-to-face interviews (12 participants) and focus groups (16 participants). The sample of 17 men and 11 women reported an average of 28 years between the risk period of this virus and the first test. These data show that risk awareness does not necessarily lead to action [2-15].

Prof. Al-Amiri et al presented a powerful panel of diagnostic and prognostic bio markers for hepatitis C-related complications as early markers for the diagnosis of hepatitis virus and hepatocellular carcinoma associated with this virus. In addition, this panel can be considered as early markers for tracking the progression of liver fibrosis. 250 patients with this virus, 224 patients with fibrosis and 84 healthy people were used for sampling in this research [5].

In 2018, Panchal and Shah used neural network and expert system to diagnose hepatitis B disease. After providing diagnosis methods for hepatitis disease and determining fuzzy rules for disease diagnosis, he used the generalized regression algorithm and the results obtained showed that this algorithm can have a favorable result for hepatitis diagnosis [7-16].

In 2017, Mendza et al conducted a screening program with the aim of detecting breast cancer. They conducted this research on the information obtained between 1999 and 2007 on 49,501 samples, of which 100 people had cancer. Due

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 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12701

to the large number of healthy samples compared to those with cancer, it was necessary to perform Dimension reduction between healthy samples. For this research, algorithms of self-organizing neural network, support vector machine and spline were used and the obtained results show that these algorithms can effectively help doctors to diagnose breast cancer in women's screening [8-14].

In 2019, Du Gantkin et al. used linear discriminant analysis and fuzzy inference system for the automatic detection system of hepatitis disease. They worked on a database that included 155 samples with 19 features and 2 categories. Using linear discriminant analysis algorithm, they selected features and reduced them and classified them using fuzzy inference system. Finally, according to the results obtained, they achieved 94.16% accuracy for disease diagnosis [10].
In 2018, Dande and Samant studied artificial neural networks in medical diagnosis. In this article, they investigated several diseases and the help of neural networks to prevent wrong diagnosis of diseases. After mentioning some types of diseases and the work that has been done in the past to diagnose them using neural networks, they introduced the basic steps for medical diagnosis using neural networks. They concluded that neural networks are a powerful tool to help doctors diagnose diseases [1-9].

In 2016, Ausi used extreme learning machine to automatically diagnose hepatitis. He conducted this research on a dataset that included 155 samples and 19 features that had 2 categories. In this research, an intelligent automatic hepatitis diagnosis system has been presented, which has reached 91.5% accuracy [3-11].

4- The proposed method

4-1 Classification

After selecting the effective features from among all the features in the data set by Harris Hawks Optimization, the decision tree algorithm is used as a method of training features and classifying them in line with disease diagnosis. In this research, ID3 and C4.5 decision trees are used to classify the trained features. It should be noted that in this section, the data set is divided into two parts: training (70% of samples) and test (30% of samples) randomly using the Rand perm command in MATLAB.

4-2 Evaluation criteria

In this research, in order to evaluate the classification efficiency, we use the accuracy criterion according to the following formula.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad 4\text{-}1$$

TP: The Number of correct features that were correctly detected.
TN: The Number of wrong features that were correctly detected.
FP : The Number of correct features that were incorrectly detected as false
FN: The Number of false features that are falsely correctly recognized.

3-Data collection

The data used in this study was obtained from the authoritative source of UCI. In this database, a total of 29 features affecting hepatitis C are defined, which are used. The characteristics of the dataset used in this research are shown in Table (1-4).

Table 1-4: Introducing the characteristics of the data set

Feature symbol	Feature definition	Feature type	value
Age	Age	Continuous	years (32-61)
Gender	Gender	Discrete	male = 1; Female = 0
BMI1	Body Mass Index	Discrete	[25 ; ‘2 = [18:5 ; 25] ‘ 1 = [0:18:5] 5 = [35 ; 40] ‘4 = [30 ; 35] ‘3 =30]
Fever	Fever	Discrete	Yes = 1; No = 0
Nausea/Vomiting	Nausea/Vomiting	Discrete	Yes = 1; No = 0
Headache	Headache	Discrete	Yes = 1; No = 0
Diarrhea	Diarrhea	Discrete	Yes = 1; No = 0
Fatigue	Fatigue	Discrete	Yes = 1; No = 0
Bone ache	Bone ache	Discrete	Yes = 1; No = 0

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 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12701

Jaundice	Jaundice	Discrete	Yes = 1; No = 0
Epigastria pain	Epigastria pain	Discrete	Yes = 1; No = 0
WBC	WBC	Discrete	‘2 = [4000-11000] ‘1 = [0-4000] 3 – [11000-12101]
RBC	RBC	Discrete	[3000000- ‘1 – [0-3000000] – [5000000-5018451] ‘2 – 3
HGB	HGB	Discrete	For men: [17.5 ‘Y2=[14-17.5] ‘1= [2-14] 3–20]
			for women: ‘2=[12.3-15.3] ‘1=[2-12.3] 3=[15.3-20]
Plat	Plat	Discrete	[100000 – ‘1 = [93013-100000] 3 = [255000-226465] ‘2 =255000]
AST1	(first week)AST1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT1	(first week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT4	(4th week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT12	(12th week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT26	(26th week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT36	(36th week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
ALT48	(48th week)ALT1	Discrete	3 = [40-128] ‘2 = [20-40] ‘1 = [0-20]
RNA Base	RNA Base	Discrete	2= [5-1201086] ‘1 = [0-5]
RNA 4	RNA 4	Discrete	2= [5-1201086] ‘1 = [0-5]
RNA 12	RNA 12	Discrete	2= [5-1201086] ‘1 = [0-5]
RNA EOT	RNA EOT	Discrete	2= [5-1201086] ‘1 = [0-5]
RNA EF	RNA EF	Discrete	2= [5-1201086] ‘1 = [0-5]
Baseline Histological Grading	Baseline Histological Grading	Discrete	[1-16]
Baseline Histological	Baseline Histological	Continuous	1= No Fibrosis 2=Portal Fibrosis 3=Staging Few Septa 4=Stagig Many Septal 5=Cirrhosis

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 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12701

The results of the simulation

In order to compare the results, genetic algorithms and PSO are also used to combine with ID3 and C4.5 decision trees. Also, instead of a decision tree, we use a neural network. The results of the accuracy of combined algorithms (HHO-ID3), (HHO-C4.5), (GA-ID3), (GA-C4.5), (PSO-ID3) and (PSO-C4.5), (HHO-MLP),(GA-MLP) and (PSO-MLP) are shown in table (2-4).

Table 2-4: Accuracy results of applying combined algorithms on the data set used in the present research in the field of hepatitis.

Algorithm	HHO-ID3	HHO-C4.5	GA-ID3	GA-C4.5	PSO-ID3	PSO-C4.5	HHO-MLP	GA-MLP	PSO-MLP
Accuracy	95.2631	95.9112	92.2317	92.8741	93.5163	94.1789	95.4186	92.4018	94.001

As shown in Table 2-4, the combination of Harris Hawks Optimization and C4.5 was able to predict hepatitis C disease with an accuracy of 95.9112 and by reducing the number of features in the database. HHO-C4.5 integrated algorithm has selected 10 effective features on hepatitis C diagnosis among 29 features.

Table 3-4 shows the features selected by HHO algorithm.

Table 3-4: Features selected by Harris Hawks Optimization

Row	Feature Selected by HHO
1	Jaundice
2	Fever
3	Bone ache
4	WBC
5	HGB
6	RBC
7	Plat
8	Fatigue
9	Nausea/Vomiting
10	Diarrhea

II. CONCLUSION

Considering the importance of this type of disease all over the world, especially in Iran, in this research, an attempt was made to provide an automatic system for the detection of hepatitis C disease by using artificial intelligence techniques. Based on this, using the Harris Hawks Optimization in combination with the decision tree classification algorithm, we presented an automatic system on the dataset obtained from the HCV dataset from the UCI reference, which is able to diagnose this disease with 95.9112%. Since the proposed dataset has 29 features, the combined algorithm first reduces the feature dimensions and then diagnoses the disease. In order to compare the proposed algorithm and evaluate its performance, genetic and PSO algorithms were used instead of Harris Hawks Optimization. Also, the MLP algorithm was combined with Harris Hawks Optimization, Genetic and PSO algorithms. The results show the efficiency of the combination algorithm of decision tree and Harris Hawks Optimization. In this research, two decision trees ID3 and C4.5 were used, and the accuracy of the HHO-C4.5 algorithm was higher compared to the HHO-ID3 algorithm.

DOI: 10.17148/IJARCCE.2023.12701

REFERENCES

[1]. B. McMahan et al., "Communication-Efficient Learning of Deep Networks From Decentralized Data", *Artificial Intelligence and Statistics Proc. PMLR*, vol. 10, no. 1, pp. 1273-82, 2017.

[2]. Mazloomy-Mahmoodabad SS, Khodayarian M, Morowatisharifabad MA, Lamyian M, Tavangar H. Iranian Women's Breast Health–Seeking Behaviors: Husband's Role. Cancer Nursing. 2018 Sep 1;41(5):409-17.

[3]. C. En Guo, S.-C. Zhu and Y. N. Wu, "Primal Sketch: Integrating Structure and Texture", *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 5-19, 2007.

[4]. M. Ordouei and T. BaniRostam, Integrating data mining and knowledge management to improve customer relationship management in banking industry (Case study of Caspian Credit Institution), Int. J. Comput. Sci. 3 (2018), 208–214.

[5]. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, no. 2, pp. 569–571,1999.

[6]. Hogade, N., Pasricha, S. and Siegel, H.J, "Energy and Network Aware Workload Management for Geographically Distributed Data Centers". IEEE Transactions on Sustainable Computing, vol.7, no. 2, pp.400–413. 2021

[7]. A. Wierman, Z. Liu, I. Liu and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response", Proc. Int. Green Comput. Conf., vol.7, no. 6, pp.1-10, 2014.

[8]. M. Ordouei, A. Broumandnia, T. Banirostam and A. Gilani, Optimization of energy consumption in mart city using reinforcement learning algorithm, Int. J. Nonlinear Anal. Appl. In Press, (2022) 1–15.

[9]. J. D. Jenkins et al., "The benefits of nuclear flexibility in power system operations with renewable energy", Appl. Energy, vol. 22 no. 2, pp. 872-884, 2018.

[10]. M. Ordouei and T. Banirostam, Diagnosis of liver fibrosis using RBF neural network and artificial bee colony algorithm, Int. J. Adv. Res. Comput. Commun. Engin. 11 (2022), no. 12, 45–50.

[11]. Haoying Dai, Yanne Kouomou Chembo, "RF Fingerprinting Based on Reservoir Computing Using Narrowband Optoelectronic Oscillators", Journal of Lightwave Technology, vol.40, no.21, pp.7060-7071, 2022.

[12]. M. Ordouei and M. Moeini, Identification of female infertility in people with thalassemia using neural network, Int. J. Mechatron. Electeric. Comput. Technol. 13 (2023), no. 48, 5371–5374.

[13]. Floris Van den Abeele, Jeroen Hoebeke, Girum Ketema Teklemariam, Ingrid Moerman, Piet Demeester, "Sensor Function Virtualization to Support Distributed Intelligence in the Internet of Things", Wireless Personal Communications, vol.81, no.4, pp.14-18, 2015.

[14]. M Moeini, SH Alizadeh, Proposing a new model for determining the customer value using RFM model and its developments (case study on the Alborz insurance company)Journal of Engineering and Applied Sciences, 2016.

[15]. A Moradi, M Ordouei, SMR Hashemi, Multi-period generation-transmission expansion planning with an allocation of phase shifter transformers, Int. J. Nonlinear Anal. Appl. In Press, (2023) 1–12.

[16]. J. Hwang, J. Kim and H. Choi, "A review of magnetic actuation systems and magnetically actuated guidewire- and catheter-based microrobots for vascular interventions", Intell. Serv. Robot., vol. 13, no. 1, pp. 1-14, 2020.

[17]. D. G. Feitelson, D. Tsafrir and D. Krakov, "Experience with using the parallel workloads archive", J. Parallel Distrib. Comput., vol. 74, no.3, pp. 2967-2982, 2014.

[16]. M. Ordouei, I. Namdar." Web Robot Detection Based On Fuzzy System and PSO Algorithm", IJCSN International Journal of Computer Science and Network, Volume 7, Issue 4, August 2018.

[18]. B. Accou, J. Vanthornhout, H. V. Hamme and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network", Scientific Reports, vol. 13, no. 1, pp. 812, 2023.

[19]. Serim Lee, Nahyun Kim, Junhyoung Kwon, Gunhee Jang, "Identification of the Position of a Tethered Delivery Catheter to Retrieve an Untethered Magnetic Robot in a Vascular Environment", Micromachines, vol.14, no.4, pp.724, 2023.

[20]. M Ordouei, A Broumandnia, T Banirostam, A Gilani, Providing A Novel Distributed Method For Energy Management In Wireless Sensor Networks Based On The Node Importance Criteria, Journal of Namibian Studies: History Politics Culture, 2023.