Advanced in Engineering and Intelligence Systems

Journal Web Page: https://aeis.bilijipub.com

Real-time Classification and Hepatitis B Detection with Evolutionary Data Mining Approach

Asadi Srinivasulu[1] , Goddindla Sreenivasulu[2] , Olutayo Oyeyemi Oyerinde[3,*]

[1] *Data Science Research Lab Blue Crest University, Monrovia, Liberia-1000*
[2] *Sri Venkateswara University, Tirupati Andhra Pradesh, Tirupati District, India – 517501*
[3] *School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, 2050, South Africa*

Highlights

- A new deep learning-based classification approach is developed for the diagnosis of hepatitis B
- The proposed model includes the automatic extraction of features with minimum redundancy and possible dimensions
- A series of evaluation criteria to compare the proposed method with the previous methods are developed
- The accuracy analyses indicate that the proposed approach has a functional superiority compared to other previous methods, with a precision of 97.50%

Article Info

Received: 01 February 2023
Received in revised: 27 March 2023
Accepted: 13 March 2023
Available online: 28 March 2023

Keywords

Hepatitis B,
Classification,
Data Mining,
Deep Learning

Abstract

Hepatitis is a disease that occurs in all ages and levels of the life of people. Hepatitis disease does not only have a deadly effect, but its identification, diagnosis, and early detection can help to treat the disease in the body and care and maintenance. Hepatitis has a variety of types that this type of study deals with hepatitis B. In this research, a new classification approach is developed for the diagnosis of hepatitis B disease using an optimized deep-learning method. This method, which involves the automatic extraction of features with minimum redundancy and minimum possible dimensions, and then modeling data from a low to a high level, can be used as a data mining method in the discovery and extraction of knowledge in computer-aided medical systems to be employed. Also, a series of evaluation criteria, including accuracy, to compare with the previous methods and to ensure the proposed approach is presented.

1. Introduction

One of the most important health issues in the world is hepatitis, which is one of the diseases that can occur in all ages. The most common age people experience is birth and also childhood. Hepatitis disease has only one deadly effect, but the early detection, diagnosis and anticipation of this disease can be useful in preventing other diseases. Hepatitis is one of the most common infectious diseases. It can be estimated that 1.5 million deaths occur every year [1]. Hepatitis is a virus that infects and damages liver cells by six different types of viruses in the liver. These viruses include A, B, C, D, E and G, which are known by their scientific names HAV, HBV, HCV, HDV, HEV and HGV [2]. This study examines the early detection of hepatitis B or HBV disease.

Hepatitis B is a DNA-transmitted DNA virus that spreads through the skin as well as sex, and affects around 300-400 million people every year [3]. The disease can cause a chronic liver disease that can lead to a risk of liver cancer and the loss of life. Identification, diagnosis and

[*] Corresponding Author: Olutayo Oyeyemi Oyerinde
Email: olutayo.Oyerinde@wits.ac.za

early diagnosis of this disease are very important. Because medical results are always uncertain, so eliminating human resources directly and examining outcomes using intelligent methods and human supervision can be an interesting issue in this area. Hence, the provision of intelligent medical systems that can predict and diagnose early illnesses is considered a special necessity in the field of medical science.

Today, the discovery of information from various data, especially in the field of medicine, has led to dramatic developments in this field. One of the sciences that deals with the discovery of information and knowledge of data are data mining, which is the science of identifying and extracting hidden features of data. Intelligent medical systems can use data mining methods to discover new knowledge of information, given the data given to them as inputs. One of the ways in which data mining has been used abundantly in recent years is the neural network approach. One of these neural networks, which has certain complexities, is the deep learning approach, which, given the particular type of training and testing of data, can create a new structure in the methods of predicting hepatitis. Hepatitis B data basically has a number of common features, including the gene, a protein containing HBx itself, containing 154 types of amino acids, a protein core, a protein level, and a polyurethane protein. There are other features, such as nucleotide and pure protein. In analyzing these features, there are three general categories, N, P, and a special mode. In the general category N, there is an explosion N, a WN cluster, and Fasta N. In the general category P, the explosion P, the WP cluster and Festa P, and in the general category of genotype, there is also resistance. It can be used to diagnose and anticipate the early onset of hepatitis B disease, so that it can be used to extract data from a person's experiment based on the characteristics and other features, a specific set of information.

Hepatitis B disease is one of the deadly infectious diseases whose advanced condition leads to liver cancer. Identifying, detecting and anticipating hepatitis B can prevent many of the dangerous effects of this infectious disease. Hence, the existence of intelligent medical diagnostic and predictive systems that can detect and extract new information and knowledge from the data.

The use of a deep learning approach to teaching hepatitis B data and its testing to predict the presence and diagnosis of the disease is a new method in the field of intelligent diagnostic and medical forecasting systems. This method, which is based on auto-extraction operations with minimum redundancy and minimum dimensions, and then data modeling from low to high levels, can be used as a data mining method for the discovery and extraction of knowledge.

The rest of this paper can be categorized as follows. In section 2, a series of smart methods that identify, detect, and predict hepatitis B virus infection are reviewed. In Section 3, the dataset used in this paper is introduced. Section 4 introduces the proposed approach and Section 5 describes the simulation and its results. The conclusion is also stated in section 6.

2. Literature Review

This section reviews a series of smart methods that identify, detect and predict hepatitis B virus infection. Numerous techniques based on data mining have been developed to diagnose, predict and classify diabetes. In [4], a comprehensive review of the state-of-the-art in the area of diabetes diagnosis and prediction using data mining is provided. The aim of this paper is twofold; Firstly, data mining-based diagnosis and predictiion solutions in the field of glycemic control for diabetes are explored and investigated. Secondly, in the light of this investigation, a comprehensive classification and comparison of techniques that have been frequently used to diagnose and predict diabetes based on important key metrics are provided. For the classification of hepatitis data and the discovery of a series of knowledge based on cavernous data methods, neural networks such as the multilayer perceptron neural network have been used as pattern recognition methods [5]–[9]. Other classical methods, such as Neu Bisin, the K-8's closest neighboring method, and other neural networks [10], have been proposed to diagnose and predict hepatitis disease to this day. In [11], hepatitis disease is predicted using a combination of backup vector carriers and a gradual refrigeration optimization algorithm. The data set used is the same as the data provided by the UCI, and the results indicate that the proposed method is accurate to 96.25%. This research is one of the best examples of intelligent diagnostic and predictive methods for hepatitis.

In [12], proposed a combined method of decision support system based on large data sets and machine learning methods to predict hepatitis disease. The research claims that their results are 100% accurate based on accuracy. This research also uses UCI data. In [13], a medical cost estimate for the treatment of hepatitis at the time of diagnosis and timely prediction of the disease is presented based on neuro-fuzzy network approach. The study uses 110 people to predict the disease, as well as estimate their health care costs.

In [14], the establishment of a neural network expert system for predicting and diagnosing hepatitis B has been addressed. The type of neural network, the generalized regression neural network, is a kind of kernel-based neural network that generates regression and has many similarities to the networks. This kind of neural network has common features with a probabilistic neural network, because it has the probability of formulating data dimensions as a probabilistic classification approach. In [15], they also provided an artificial intelligent support system for deciding interferon behavior in chronic hepatitis B disease. This method uses a decision tree and its type is the decision tree of C5.0 and the boosting method. The predicted results indicate a 100% accuracy in the diagnosis of chronic hepatitis B disease.

In [16], the use of a multilayered perceptron neural network for the prediction and diagnosis of hepatitis B disease has been used with a sigmoid transfer function approach. The neural network learning method is also based on Levenberg Marquardt. The dataset used is UCI data. The results indicate a precision between 91.9% and 93.8%. In [17], the use of a backup vector machine is combined with another method known as the lethal method for detecting and anticipating hepatitis. The reason for using an overlay method is that it can eliminate data noise before a backup vector machine is classified. The data from this research is based on UCI data. The classification results and then the forecast is 72.73%, which is not a significant result. In [18], a novel method for the diagnosis of hepatitis B virus infection using human blood serum Raman spectroscopy combined with a deep learning model is presented. The sera of 499 people infected with hepatitis B virus and 435 healthy controls were measured in this experiment. The data were subjected to a dimensionality reduction by principal component analysis. Then, the features of multiple scales were preserved and fused by a multiscale fusion convolution operation. The gated recurrent unit network was added to extract time series features and finally output the result of the classification through a softmax layer. A diagnostic model based on a gated recurrent unit and multiscale fusion convolutional neural network was constructed and evaluated by a 10-fold cross-validation method. The combination of Raman spectroscopy and deep learning models is expected to be applied well in the early screening of hepatitis B and is a promising screening method.

3. Dataset

The datasets used will be two links on the Internet, each with separate data with almost identical features, but the size of the sample is different. These two data are available at https://archive.ics.uci.edu/ml/datasets/Hepatitis and https://hbvdb.ibcp.fr/HBVdb. The first data has 155 input data and 19 attributes. The second data, known as HBVdb, contains 78573 data, the latest update on 2/7/2017. This data has 15 features. This research uses the first data with 19 features, of which 13 are binary and 6 are discrete values. In Table (1), you can obtain information about the properties of this data.

In Table 1, the value of the attribute, yes or no, is a Boolean value.

Table 1. Information on the characteristics of the Hepatitis B Dataset

Feature ID	Feature Name	Feature Value
1	Age	10, 20, 30, 40, 50, 60, 70, 80
2	Sex	Male, Female
3	Steroid	Yes, No
4	Antivirals	Yes, No
5	Fatigue	Yes, No
6	Malaise	Yes, No
7	Malaise	Yes, No
8	Big Liver	Yes, No
9	Liver firm	Yes, No
10	Spleen palpable	Yes, No
11	Spiders	Yes, No
12	Ascites	Yes, No
13	Varices	Yes, No
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	Bilirubin	33, 80, 120, 160, 200, 250
16	SGOT	13, 100, 200, 300, 400, 500

17	ALBUMIN	2.1, 3, 0.3, 8, 4.5, 5.0, 6.0
18	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
19	HISTOLOGY	Yes, No

4. Proposed Approach

Once the data is entered as inputs, it is necessary to carry out training on them, which is done using the deep learning of this work. Deep learning is a complex neural network that has learning and testing functions and is considered to be a machine learning method. Deep learning has two main advantages that include the representation of learning and the multi-layered learning of representations. Auto-extraction operations call the representation of learning features with minimal redundancy and the least possible dimension in deep learning. Data modeling from low to high levels in deep learning is called multilayered learning of representations. The type of deep learning provided here is different from the main structure, since the goal is to optimize this method for diagnosing hepatitis B disease. In order to optimize deep learning to find an optimal value close to the main learning parameters of deep learning and its core, a multi-dimensional vector series is used as the relation (1).

$$X = [P_1, P_2, P_3, P_4] \qquad (1)$$

which according to (1), P_1 or σ is the core parameter in the interval [0.0001, 33]. P_2or $C$is equivalent to the complexity and is in the range [0.1, 35000]. P_3 or ε in the range [0.00001, 0.0001]. Also, P_4 or t is equivalent to the error tolerance [0, 0.5]. These selective values are based on common settings in previous articles. As it is clear, the classification stage has two main parts, which include model building and model testing. In the first phase, an educational algorithm runs on data that aims to upgrade a model with an estimate of output. The purpose of this model is to describe the relationship between the class and the predictor. The quality of the model produced in the test phase of the model is evaluated. In principle, the precision criterion is used to evaluate the efficiency of most classification methods, which is related to (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

in (2), TP is a positive item that is categorized as a positive one. TN are negative cases that are negatively classified correctly. FP is a class with negative classes classified as positive and FN cases with positive classes are classified negatively. Therefore, the precision rate in this method is used to measure the quality of the produced solution, which is also called the fit function. Of course, the purpose of fitting the function is to consider such issues in the classification using the criterion of assessment of accuracy as a measurement method. In order to classify with deep learning, all features in the dataset must have a real value. Therefore, the nominal attributes are converted to ordinary data. Then, the normalization of data needs to be done. In order to prevent the magnitude of the values from increasing in the numerical range, as well as to avoid numerical complexity in the calculations, normalization operations are performed. This operation is performed using (3).

$$X_{Normalization} = \frac{X - X(min)}{X(max) - X(min)} \qquad (3)$$

In the following, two methods are used to divide the data into training and test data. The first is the K-Fold mutual validation, one of the most popular strategies for estimating the efficiency of classification methods, and is an appropriate way to prevent optimal localization and too much fitting. In this method, examples of training from the sample tests are independent. In the K-Fold mutual validation, the K value is always standardized to 10. Therefore, the data set is divided into 10 sections, of which 9 sections are applied in the educational process, and the remaining ones, which are 1, are used as a test. The program also runs up to 10 times, enabling each part of the data to reach the test process after training. The accuracy rate for the learning process and the test is calculated by the sum of independent accuracy rates and error rates for each run and divided by 10 times the total implementation. The second method is also a way to keep it open. In this way, the data are divided into two parts, which include training and test data. In this method, there is no specific benchmark for determining the number of training and test data. The main purpose of using these two methods for data segmentation is to evaluate the application of the method more than one perspective. After the preprocessing phase, the first part of the deep learning is represented by a random presentation of the solution, which is done using the upper and lower limit of each parameter, expressed as (4).

$$\begin{aligned} Sol_x = {} & LWB[i] + (UPB[i] \\ & -LWB[i]) \times Random \end{aligned} \qquad (4)$$

which according to (1), $LWB[i]$ is the upper limit and $UPB[i]$is the upper limit. The $Random$value, which is a random value, is in the range (0, 1). Then, the model is trained and tested using all the solutions produced. Then, a reference value using b solution with the best accuracy rate of $b = 5$ is promoted. Then a series of new solutions are generated that are in the form of relations (5), (6) and (7).

$$X_1 = P_1 + (P_2 - P_1) \times r_1 \tag{5}$$

$$X_2 = P_1 + (P_2 + P_1) \times r_2 \tag{6}$$

$$X_3 = P_1 + P_2 \times r_3 \tag{7}$$

in (7), r_1, r_2 and r_3 are random numbers in the interval (0, 1). Using this method, 30 solutions are developed that can be used to teach or test the model.

5. Simulation and Results

Data is entered as input in the system. By dividing data using two methods of K-Fold validation and holding, they are divided into two educational and test sections. 70% are used as training samples and 30% are used as test or test data. First of all, it is necessary to mention the settings for optimal deep learning as presented in Table (1).

In order to create deep learning, the MATLAB toolbox has been used, as well as the NNTRAINTOOL window, which is related to the neural network, and is structured in a deep neural network that has undergone some kind of modification and optimization, and according to the (4) to (7) shows this case. Convolution Neural Network (CNN) is the main technique of deep learning in this method. The deep learning structure after the project implementation can be seen in Fig. 1, which is clear that the values set for this network are in the table (2).

Table 2. Set values for deep learning

Iteration Number	150 cycles
Layers Numbers	2 layers
Input Numbers	2 columns of input data
Output Numbers	2 columns of input data
Training Method	Random weighting function with bias rules
Performance Evaluation Method	MSE

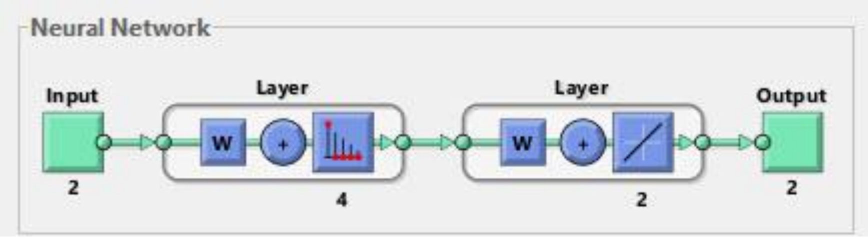
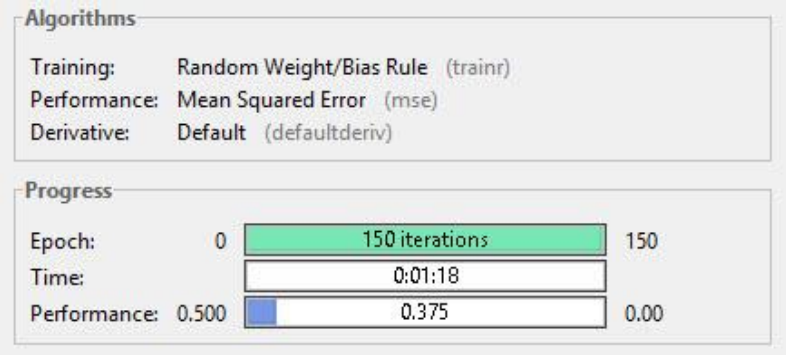


Fig. 1. Deep learning structure (CNN)

The network has two layers, the first layer uses a binary step actuator function with threshold θ. Also, in the second layer, the same actuator function is used. Other settings for algorithms and repetitions can be seen from the table (2) in Fig. 2.



Algorithms

Training: Random Weight/Bias Rule (trainr)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch: 0 150 iterations 150
Time: 0:01:18
Performance: 0.500 0.375 0.00

Fig. 2. Algorithms and other settings used in deep learning (CNN)

Also, after data training, the efficiency, confusion matrix and receiver factor characteristics, known as the ROC curve, are used to display training and test input data. The performance rate based on the mean squared error can be seen in Fig. 3, the confusion matrix in Fig. 4, and the ROC graph in Fig. 5, after training and testing the data.

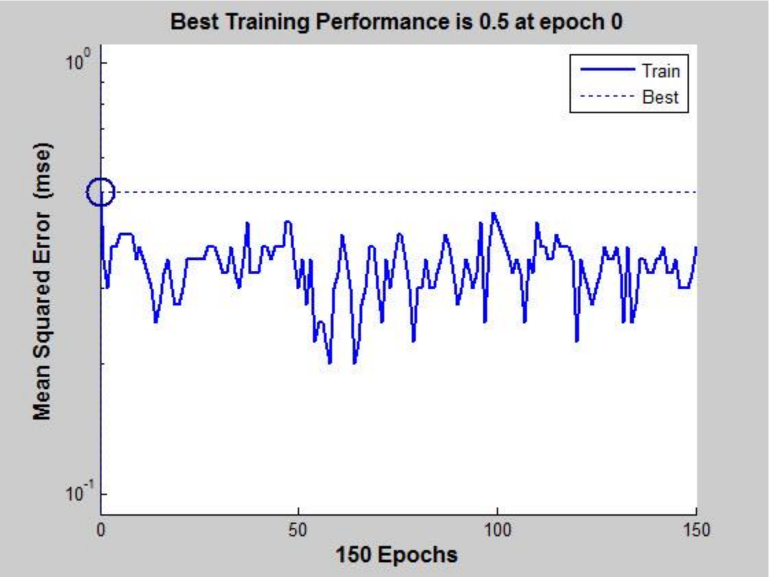
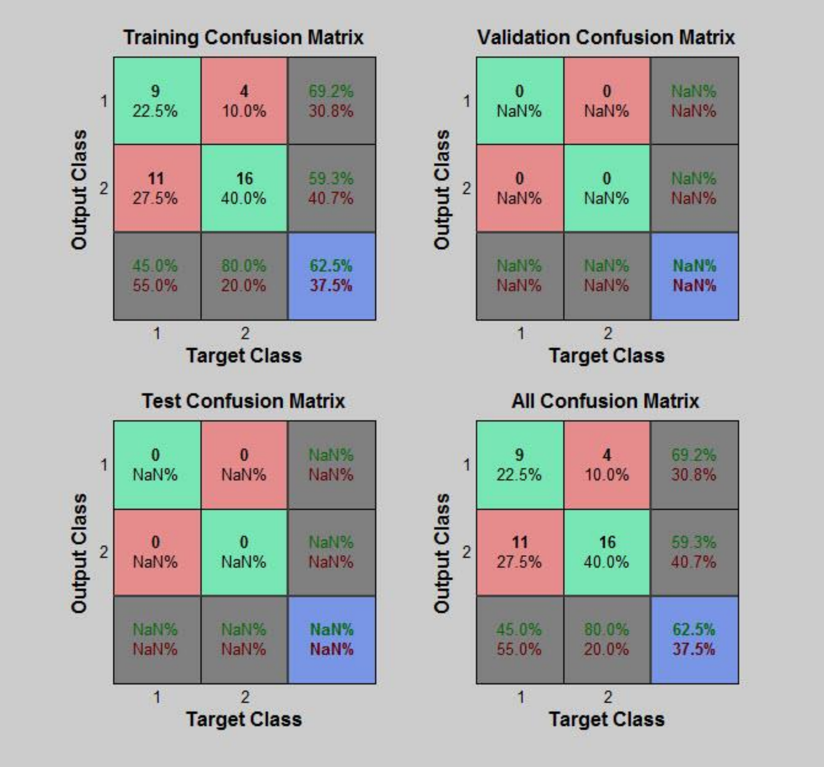


Fig. 3. Average Efficiency Based on Squared Error

The optimal value for the mean squared error after 150 rounds of repetition of training and data testing is 0.5, which is somewhat optimal in its type.



Training Confusion Matrix

Output Class	1	2	
1	9 22.5%	4 10.0%	69.2% 30.8%
2	11 27.5%	16 40.0%	59.3% 40.7%
	45.0% 55.0%	80.0% 20.0%	62.5% 37.5%

Target Class

Validation Confusion Matrix

Output Class	1	2	
1	0 NaN%	0 NaN%	NaN% NaN%
2	0 NaN%	0 NaN%	NaN% NaN%
	NaN% NaN%	NaN% NaN%	NaN% NaN%

Target Class

Test Confusion Matrix

Output Class	1	2	
1	0 NaN%	0 NaN%	NaN% NaN%
2	0 NaN%	0 NaN%	NaN% NaN%
	NaN% NaN%	NaN% NaN%	NaN% NaN%

Target Class

All Confusion Matrix

Output Class	1	2	
1	9 22.5%	4 10.0%	69.2% 30.8%
2	11 27.5%	16 40.0%	59.3% 40.7%
	45.0% 55.0%	80.0% 20.0%	62.5% 37.5%

Target Class

Fig. 4. Confusion matrix

From the top left, the matrix of clutter of training samples is initialized in terms of output class and target class. Then, on the top right, Validation of the confusion matrix based on the output class and the target class. In the lower left-hand side, the data test step for the confusion matrix is based on the output classes and target classes, and in the bottom right, all values of the confusion matrix based on the output class and the target class are measured, which is the final value It is 62.5% at 37.5%.

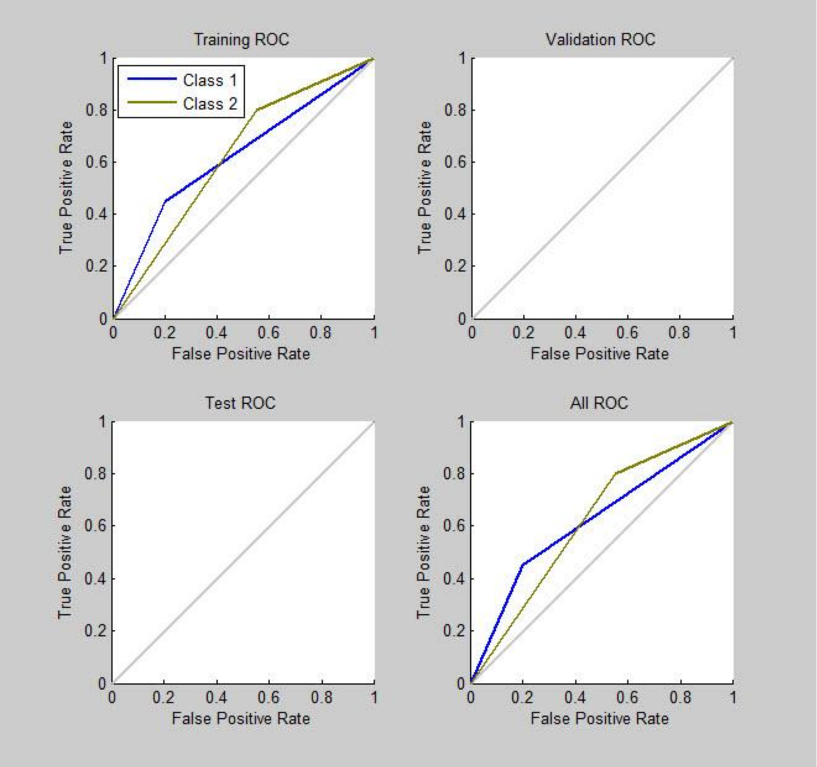


Fig. 5. ROC chart

Fig. 5. shows the ROC curve from the top left. The upper right shows the ROC curve validation. The bottom left, the ROC curve for data testing and the lower right shows all the ROC charts. At the end, a graph for the deep learning depth optimization aiming at detecting hepatitis B is presented, which is in Fig. 6.

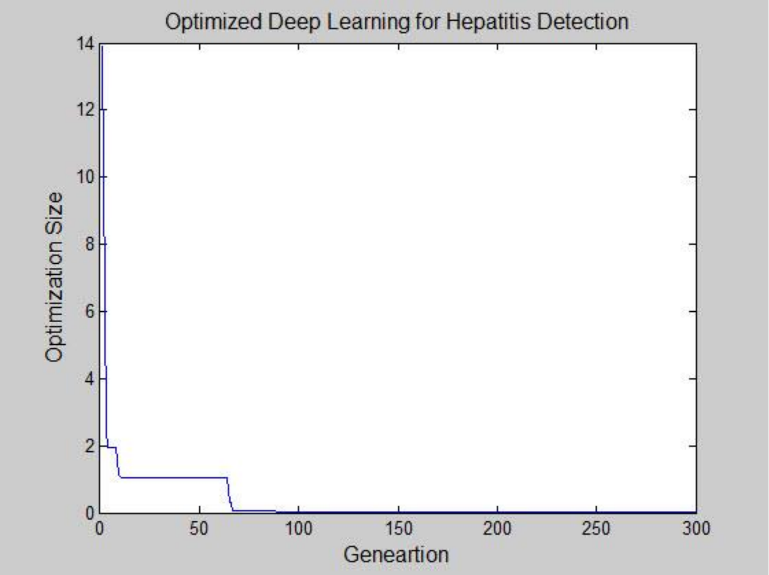


Fig. 6. Deep learning optimization rates for diagnosis of hepatitis B

It is worth noting that in various performances, this optimization rate will be different due to the production of different numbers, which is due to the fact that there is a random case in improving deep learning that has a random property in optimization. The accuracy of the proposed method for diagnosis of hepatitis B is also 97.50%.

Considering that most of the methods used to diagnose hepatitis as an evaluation criterion and to ensure their proposed approach has been used accurately, this study has also used this approach for comparison. The results of the precision assessment method for measuring the rate of detection of hepatitis B disease in the proposed method are presented in the table below (Table 3).

Table 3. compares the criteria for the accuracy of the proposed method compared to similar previous methods

Methods	**Accuracy (%)**
SVM and SA [11]	96.25 %
MLP [16]	91.9 % - 93.8 %
SVM [17]	72.73 %
Proposed Method	97.50 %

It is clear that the proposed method yields better accuracy than its predecessor.

6. Conclusion

One of the most commonly diagnosed diseases that is seen in third-world societies is hepatitis, which has many different types. One of this hepatitis is hepatitis B. Given the fact that medical sciences require high costs, it is necessary to provide intelligent methods that can handle patients' information and achieve results. Therefore, in this research, we present an intelligent approach to the diagnosis of hepatitis B disease that is optimized based on deep learning. The results of the evaluation in terms of accuracy indicate that the proposed approach has a functional superiority compared to other previous methods with a precision of 97.50%.

For future work, we need to pre-process the data and use hybrid techniques that incorporate different models in parallel instead of using an individual model for accurate disease diagnosis, classification and prediction. For preprocessing, we need to use dimensionality reduction, denoising, feature selection, and feature extraction techniques in combination with classification and prediction schemes for optimal performance and results.

REFERENCES

[1] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, "Detection and prediction of diabetes using data mining: a comprehensive review," *IEEE Access*, vol. 9, pp. 43711–43735, 2021.

[2] W. M. Lee, "Hepatitis B virus infection," *New England journal of medicine*, vol. 337, no. 24, pp. 1733–1745, 1997.

[3] J. Cohen, "The scientific challenge of hepatitis C." American Association for the Advancement of Science, 1999.

[4] I. Maida *et al.*, "Severe liver disease associated with prolonged exposure to antiretroviral drugs," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 42, no. 2, pp. 177–182, 2006.

[5] S.-H. Chiu, C.-C. Chen, and T.-H. Lin, "Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer," *Artif Intell Med*, vol. 44, no. 3, pp. 221–231, 2008.

[6] K. Kayaer and T. Yildirim, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," in *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, 2003, p. 184.

[7] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif Intell Med*, vol. 34, no. 2, pp. 113–127, 2005.

[8] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Syst Appl*, vol. 36, no. 1, pp. 944–949, 2009.

[9] B. S. Blumberg, "Hepatitis B virus, the vaccine, and the control of primary cancer of the liver," *Proceedings of the National Academy of Sciences*, vol. 94, no. 14, pp. 7121–7125, 1997.

[10] M. A. Feitelson, "Hepatocellular injury in hepatitis B and C virus infections," *Clin Lab Med*, vol. 16, no. 2, pp. 307–324, 1996.

[11] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput Methods Programs Biomed*, vol. 108, no. 2, pp. 570–579, 2012.

[12] Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," *Appl Soft Comput*, vol. 13, no. 8, pp. 3429–3438, 2013.

[13] R. J. Kuo, W. C. Cheng, W. C. Lien, and T. J. Yang, "A medical cost estimation with fuzzy neural network of acute hepatitis patients in emergency room," *Comput Methods Programs Biomed*, vol.

122, no. 1, pp. 40–46, 2015.

[14] D. Panchal and S. Shah, “Artificial intelligence based expert system for hepatitis B diagnosis,” *International journal of modeling and optimization*, vol. 1, no. 4, p. 362, 2011.

[15] A. G. Floares, “Artificial intelligence support for interferon treatment decision in chronic hepatitis B,” *International Journal of Medical and Health Sciences*, vol. 2, no. 8, pp. 255–260, 2008.

[16] O. Çetin, F. Temurtaş, and Ş. Gülgönül, “An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function,” *Dicle Tıp Dergisi*, vol. 42, no. 2, pp. 150–157, 2015.

[17] A. H. Roslina and A. Noraziah, “Prediction of hepatitis prognosis using support vector machines and wrapper method,” in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, 2010, pp. 2209–2211.

[18] Z. Guo, X. Lv, L. Yu, Z. Zhang, and S. Tian, “Identification of hepatitis B using Raman spectroscopy combined with gated recurrent unit and multiscale fusion convolutional neural network,” *Spectroscopy Letters*, vol. 53, no. 4, pp. 277–288, 2020.