Improving the diagnostic performance of the neural network for COVID-19VN via weight assignment to pathological symptoms

Le Minh Ngoc[1], Nguyen Thi Thuy[2*], Dinh Van Quang[3], Dinh Van Dai[3], Bui Van Tan[3]

[1]Department of Science and Technology Management, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam;
[2]Faculty of Electronics and Telecommunications, Electric Power University, 235 Hoang Quoc Viet Street, Bac Tu Liem, Hanoi, Vietnam.
[3]National Hospital For Tropical Diseases, Dong Anh, Hanoi, Vietnam.
*Corresponding author: thuynt@epu.edu.vn.
Received: 01 Feb. 2024; Revised 03 Apr. 2024; Accepted 10 May 2024; Published 20 May 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.95.2024.29-37

ABSTRACT

In this article, we present how to create a database of Covid-19 diseases at National Hospital of Tropical Diseases (called the CovidVN database) and then develop a learning neural network based on this database to diagnose this disease. The CovidVN Database is built based on the processing of real diagnostic test results of Covid-19 patients with a large number of samples and in accordance with the structure of the Israeli Health System Covid-19 disease database (called COVIDIsr Database). Then two MLP Artificial Neural Networks corresponding to these two databases will be developed using the Deep Learning Toolbox of MATLAB software; the results of training these networks and their accuracy are compared with each other to assess the relative quality of the CovidVN database. Other, the paper presents the method of assigning weight corresponding to pathological symptoms for network input parameters. The results showed that the weighting of input attributes corresponding to the pathological symptoms is significant.

Keywords: Artificial neural networks; DATABASE Coronavirus disease (COVID-19); MultiLayer Perceptron (MLP); Assigning weight.

1. INTRODUCTION

The application of Artificial Intelligence technology to Medical Diagnosis and Treatment is booming in the world [1] and is starting in Vietnam [2, 3], of which there are applications [4] using image recognition neural networks to diagnose Covid. However, a prerequisite to apply machine learning methods in medical diagnosis is to have a database of pathological diagnoses, which only a few countries in the world have built and published, like the US, Israel, etc. but for Vietnam, absolutely not!

Therefore, in order to apply machine learning methods (Artificial Intelligence technology) to build a support system for the diagnosis of Covid 19 disease, in this work, we will first design a database used to diagnose Covid-19 disease (referred to as COVIDVN database); then based on this database will build an MLP Neural Net-work to support the diagnosis of Covid disease. The quality performance of this Neu-ron Network will be compared with a Neural Network of the same design but based on the Covid-19 Open Research Dataset (CORD-19). Other, this Neuron Network will be compared with a Neural Network of the same design and COVIDVN database but based on the method of assigning weights to pathological symptoms.

2. MATERIALS AND METHODS

2.1. Materials

Database CORD-19. General description. Attribute structure.

Retrieved from Allen University's Open Research Dataset (CORD-19) for AI in partnership with the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft Research, IBM, and the National Library of Medicine - National Institutes of Health, in coordination with The White House Office of Science and Technology Policy [4]. This database contains characteristic parameters obtained from 578 patients and 6 input attributes such as gender, cough, fever, pharyngitis, dyspnea, headache associated with 3 outputs corresponding to 3 diagnostic results: negative, positive for Influenza Covid-19 and other.

Database COVID19VN. Design. Realization.

Based on the attribute structure of the Open Research Dataset CORD-19 database, a design for the COVIDVN database can be made with 6 attributes (gender, cough, fever, sore throat, shortness of breath, headache), a sample size greater than 500, and an equivalent data collection form. The implementation of this design was carried out at Ha Noi Central Hospital for Tropical Diseases - the frontline hospital for the prevention of infectious and tropical diseases - during the peak of the Covid-19 epidemic in early 2022 with the participation of doctors, nurses on staff and patients who had come to the specialized clinic of the hospital. As a result, the first database of Covid-19 flu (COVIDVN) has been built for Vietnamese patients (COVIDVN) (shown in table 1). It consists of the characteristic parameters obtained from 504 patients and the 6 input attributes mentioned above and is associated with 3 outputs: negative, positive for Influenza Covid-19 and other.

Preprocessing the database. Filter large loss attributes; Standardized.

Before being used to train Neural Networks, the database needs to be preprocessed to filter out the attributes with large losses and then the values of the remaining attributes are normalized so that they are in the range [0,1].

Thus, after preliminary processing, the CORD-19 database used to train machine learning neural networks consists of 2 matrices: the CORD-19Inputs (IP1) matrix with the size of 6 x 578 corresponding to 6 attributes and 578 samples, and the matrix CORD-19Targets (OP1) with the size of 3 x 578 corresponding to 3 diagnosed disease results and 578 samples.

In the same way, the COVIDVN database, after preliminary processing, consists of two matrices: the COVIDVNInputs (IP2) matrix with dimensions of 6 x 504 corresponding to 6 attributes and 504 samples, and the COVIDVNTargets matrix (OP2) with size 3 x 504 corresponding to 3 diagnosed disease results and 504 samples.

Table 1. *Structural Levels of the Body and Diagnostic Methods.*

Sample number	**Attributes - symptoms (Network input)**						**Diagnostic results (Network output)**		
	Gender	**cough**	**fever**	**sore throat**	**dyspnea**	**headache**	**negative**	**positive**	**other**
1	0	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	1	0	0
4	1	0	0	0	0	0	1	0	0

. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
503	1	0	0	0	0	0	0	0	1
504	1	0	0	0	0	0	0	0	1

Note:

- Yes symptoms – 1;
- No symptoms – 0.

2.2. Methods

From the point of view of recognition theory, medical diagnosis belongs to the classification problem which is usually solved by applying Multilayer Perceptron Neural Network (MLP) [1, 7, 8]. Thus, here is the requirement to design 2 direct multilayer networks with the same structure based on 2 databases CORD-19 and COVIDVN.

Design of a Direct Multilayer Network. To ensure the science and convenience of building and surveying the network, the Neural Network Toolbox in Matlab software [8] will be used to design the network. Thus, based on the output/input structure of the aforementioned databases, it is possible to propose a network design with a specific number of layers and neurals as follows:

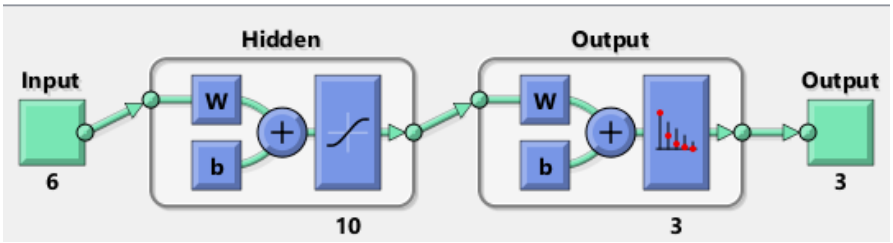

Figure 1. *MLP two-layer diagnostic neural network.*

Specifically, this network has an input with 6 neurals corresponding to 6 attrib-utes, a hidden sigmoid layer with 10 neurals and an output layer with 3 neurals as 3 outputs corresponding to 3 results of the diagnosed disease.

The back-propagation network training algorithm. To optimize/train an MLP network, the most popular method today is to use the backpropagation algorithm. This algorithm converges to a solution that minimizes the mean square error because the weighting and bias correction of the algorithm is done in the opposite direction to the gradient vector of the weighted mean square error function.

The training, testing and validation of the network were performed with sample rates of 70%, 15%, and 15%, respectively.

3. RESULTS AND DISCUSSION

3.1. Results

The results of network training with the CORD-19 database are shown in Fig. 2 and Fig. 3 below:

Based on the parameters and diagnostic rates provided in Fig. 2, we can summarize the results as follows:

* Negative Disease Diagnosis:

- Number of negative patients: 209.
- Correct negative diagnosis rate: 100%.
- Misdiagnosis rate of positive and other: 0%.

* Positive Disease Diagnosis:

- Number of positive patients: 229.
- Correct positive diagnosis rate: 37.5%.
- Rate of misdiagnosis as negative and other: 62.5%.

* Diagnosis of other diseases:

- Number of other patients: 140.
- Other correct diagnosis rate: 0.71%.
- Rate of false positive and negative diagnoses: 99.29%.

* Overview:

- Total number of patients: 578 (209 negative + 229 positive + 140 others).
- Correct diagnosis rate: 51.2%.
- Misdiagnosis rate: 48.8%.

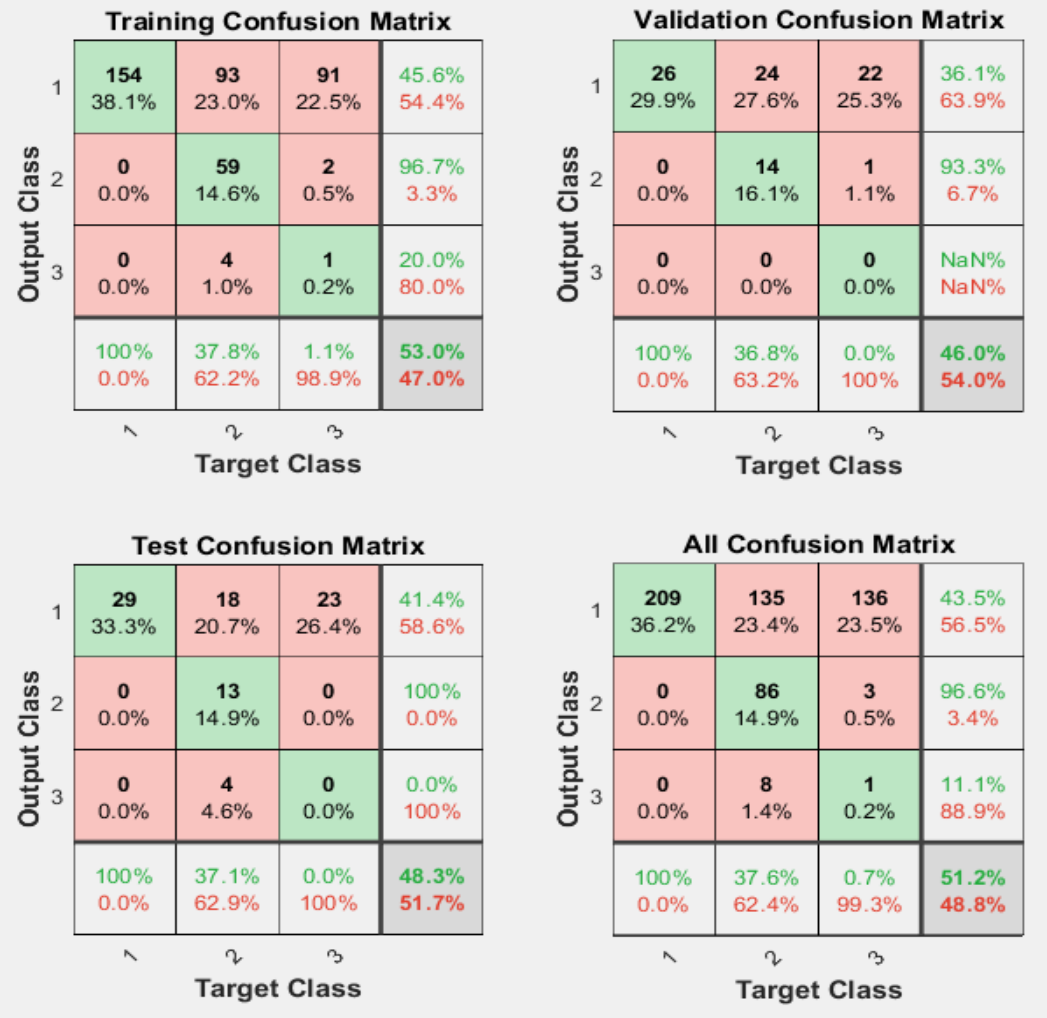

Figure 2. *Results matrix of training, validation, testing and overall error of the network based on database CORD-19.*

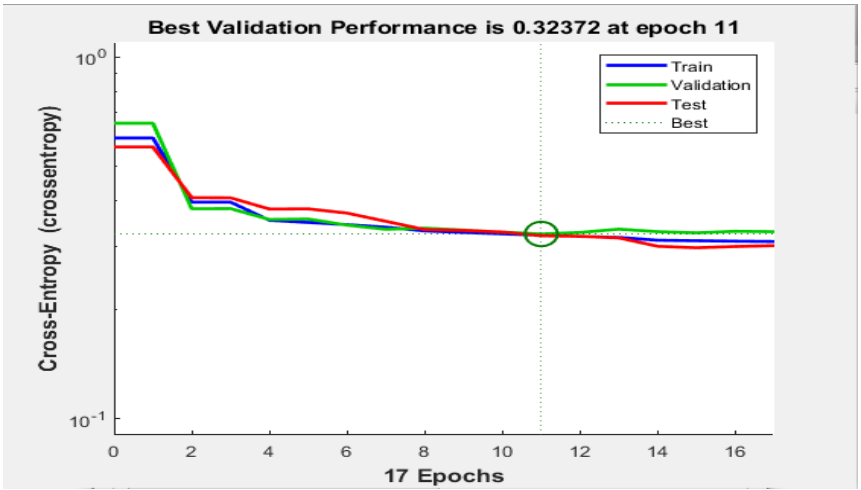

Figure 3. *Best Validation Perfomance of the network based on database CORD-19.*

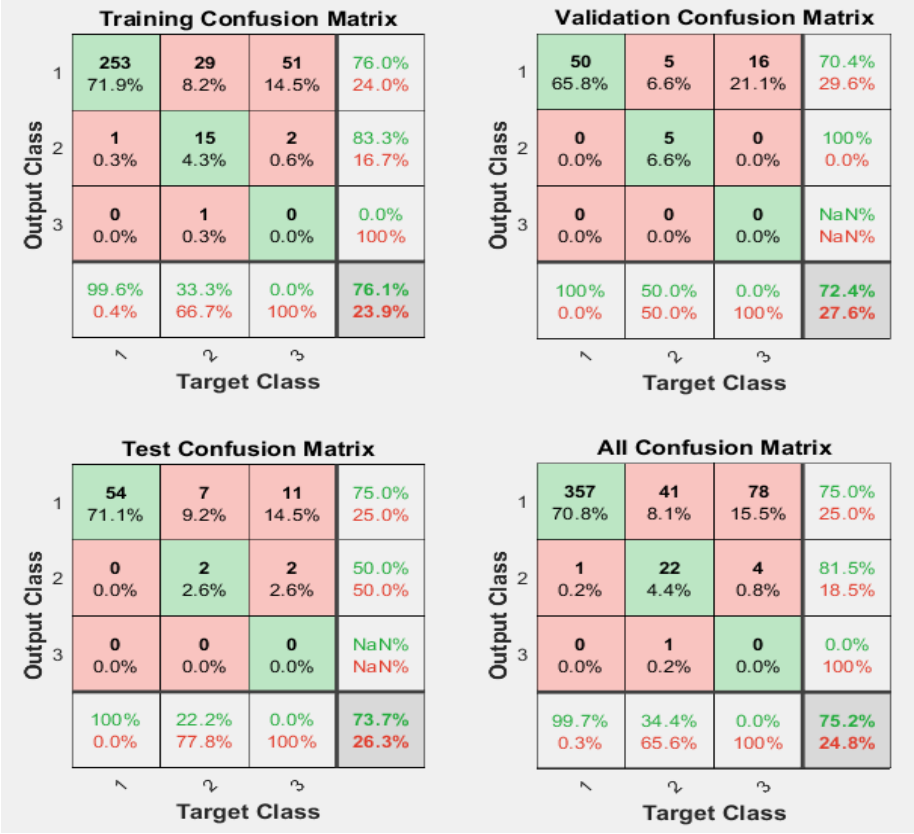

Figure 4. *Results matrix of training, validation, testing and overall error of the network based on database COVIDVN.*

The results of network training with the COVIDVN database are shown in Fig. 4 and Fig. 5 below:

Based on the parameters and diagnostic rates provided in Fig. 4, we can summarize the results as follows:

* Negative Disease Diagnosis:

- Number of negative patients: 358.
- Correct negative diagnosis rate: 99.7%.
- Misdiagnosis rate of positive and other: 0.3%.

* Positive Disease Diagnosis:

- Number of positive patients: 64.
- Correct positive diagnosis rate: 34.4%.
- Rate of misdiagnosis as negative and other: 65.6%.

* Diagnosis of other diseases:

- Number of other patients: 82.
- Other correct diagnosis rate: 0%.
- Rate of false positive and negative diagnoses: 100%.

* Overview:

- Total number of patients: 504 (358 negative + 64 positive + 82 others).
- Correct diagnosis rate: 75.2%.
- Misdiagnosis rate: 24.8%.

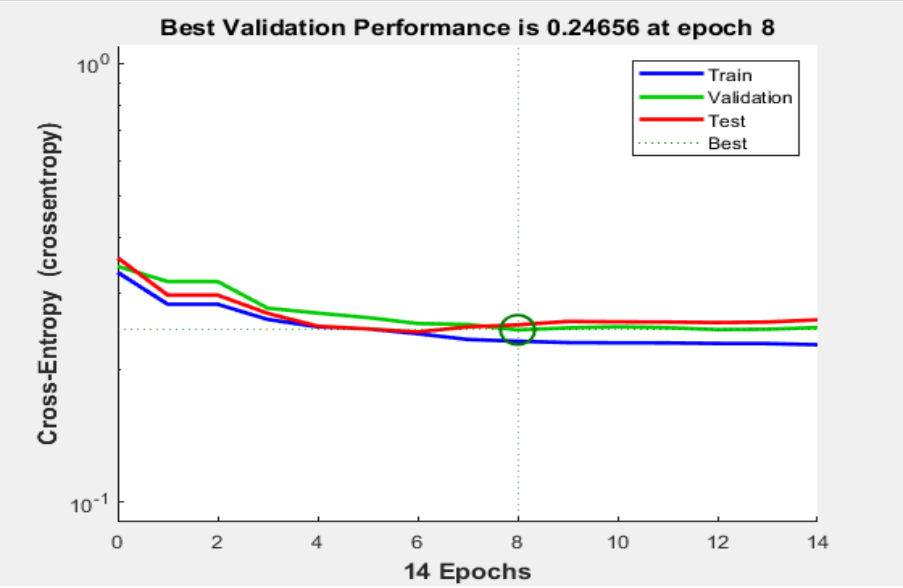

Figure 5. *Best Validation Perfomance of the network based on database COVIDVN.*

Thus, both to test the validity and effectiveness of the above method and to improve the quality of the Covid-19 Flu diagnosis network, in this study the method of assigning weights to attributes will be applied. Input appropriate to the body's structural levels for the artificial neural network to support the diagnosis of Covid-19 Flu.

With the main purpose of determining the importance of weight assignment by comparing the network training results between two cases without and with weights, we maintain the structure of the network constant (number of layers, number of cells). neurals for each layer, bias), meaning that the initial network structure remains the same.

To evaluate the importance of the input attributes, we will assign a weight of 0 to the estimated attribute and a weight of 1 to the remaining attributes. Specifically:

* Case 1: Assign a weight of 0 to the cough symptom attribute, the remaining parameters are assigned a weight of 1.

* Case 2: Assign a weight of 0 to the fever symptom attribute, the remaining parameters are assigned a weight of 1.

* Case 3: Assign a weight of 0 to the sore throat symptom attribute, the remaining parameters are assigned a weight of 1.

* Case 4: Assign a weight of 0 to the shortness of breath symptom attribute, the remaining parameters are assigned a weight of 1.

* Case 5: Assign a weight of 0 to the headache symptom attribute, the remaining parameters are assigned a weight of 1.

Table 2. *Neural network assessment test key results for weight assignment options-COVIDIsr Database (CORD-19).*

Case	Based on	Successful diagnosis rate (%)	Failure rate of diagnosis (%)	Best Validation Perfomance	Convergence Epoches
1	CORD-19	51.2	48.8	0.32372	11
2	COVIDVN	75.2	24.8	0.24656	8

The results of testing the effectiveness of the Covid 19 diagnostic neural network for the case of assigning attribute weights are given in table 2.

Table 3. *Neural network assessment test key results for weight assignment options COVIDVN.*

Case	Successful diagnosis rate (%)	Failure rate of diagnosis (%)	Best Validation Perfomance	Convergence Epoches
0	75.2	24.8	0.24656	8
1	72.4	27.6	0.22906	4
2	75.2	24.8	0.21714	21
3	75.8	24.2	0.21175	34
4	74.4	25.6	0.22945	17
5	74.4	25.6	0.22474	21

3.2. Discussion

1 - Due to the difference in input data between the CORD-19 Database and the COVIDVN Database, the following results were obtained:

• The success rate for COVIDVN database is up to 72.4% and for COVIDIsr Data-base (CORD-19) it is only 51.2%;

• The convergence period for the COVIDVN database is 8 Epoches while for the CORD-19 database it is 11 Epoches;

• The best validation result for COVIDVN database is 0.24656 while for Database CORD-19 it is 0.32372.

2 - These results show that the performance of the network based on the COVIDVN database is higher than that of the network based on the CORD-19 database. This difference in diagnostic efficiency of the two Networks may be due to the difference in accuracy in the way/methods of testing the diagnostic results.

3 - The low diagnostic success rate of both Networks (< 75%) is probably due to the inappropriate use of non-high weighted attributes corresponding to macroscopic body structure levels.

4- Assigning weights to pathological symptoms on COVID VN database the following results were obtained:

• The best success rate is for case assign a weight of 0 to the sore throat symptom attribute, the remaining parameters are assigned a weight of 1 is up to 75.8% and lowest is for case assign a weight of 0 to the cough symptom attribute, the remaining parameters are assigned a weight of 1 it is only 72.4%;

• The convergence period for the case assign a weight of 0 to the cough symptom attribute, the remaining parameters are assigned a weight of 1 is 4 Epoches while for the case assign a weight of 0 to the sore throat symptom attribute, the remaining parameters are assigned a weight of 1 it is 34 Epoches;

• The best validation result for case assign a weight of 0 to the sore throat symptom attribute, the remaining parameters are assigned a weight of 1 is 0.21175 while for case assign a weight of 0 to the shortness of breath symptom attribute, the remaining parameters are assigned a weight of 1 it is 0.22945;

•The results of assigning input attributes to Covid-19 Flu data based on success rate, convergence period, and best validation of attribute importance are as follows: Cough attribute > difficulty breathing > headache > fever > sore throat;

•These results show that the performance of the network based on the method of assigning weights to pathological symptoms is higher than that of the network based on the no method of assigning weights to pathological symptoms. This difference in diagnostic efficiency of the two Networks may be due to the difference in accuracy in the way/methods of testing the diagnostic results.

4. CONCLUSIONS

This study uses database of COVID 19 influenza (COVIDVN) built at the Central Hospital for Tropical Diseases.

The higher efficiency of the method of assigning weights to pathological symptoms Building Network compared to the no method of assigning weights to pathological symptoms Network allows to confirm the quality of the COVIDVN database as well as the Machine Learning Neural Networks based on it. Thus, it is certainly possible to performance improvement and use the neural network designed here in supporting the diagnosis of Covid-19 disease in VIETNAM conditions.

This study also points to the next development direction in the application of Artificial Intelligence in Health in VIETNAM, which is to create a database of disease diagnosis in general and COVID-19 flu in particular with structure attributes more reasonably corresponding to the microstructure levels of the body.

REFERENCES

[1]. S. Kaur et al., *"Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives,"* in IEEE Access, vol. 8, pp. 228049-228069, (2020), doi: 10.1109/ACCESS.2020.3042273.

[2]. Huynh Luong Nghia- Cong Doan Truong- Dinh Van Quang- Xuan Thu Do- Anh Ngoc Le. An *"Application of a Method of Weighting Assigning for Attributes Based on the Level of Different Body Structures to Improve the Artificial Neural Network for Diagnosing Hepatitis"*. Lecture Notes in Networks and Systems - ICISN, (ISSN: 2367-3370), (2021).

[3]. H L Nghia- N T Thuy and Đ V Quang. *"Pathological diagnosis Neural Network with inputs corresponding with structure levels of the body"*. Journal of military Science and Technology, No 57A, (2018).

[4]. G. Jain, D. Mittal, D. Thakur, and M. K. Mittal, *"A deep learning approach to detect covid-19 coronavirus with x-ray images,"* Biocybernetics and Biomedical Engineering, vol. 40, no. 4, pp. 1391–1405, (2020).

[5]. Shanthi M, Pekka P, Norrving B. *"Global Atlas on Cardiovascular Disease Prevention and Control"* (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3. (2011).

[6]. https://cset.georgetown.edu/publication/Covid-19-open-research-dataset/.

[7]. https://www.mathworks.com/products/matlab-online.html, last visit: 15/05/22

[8]. https://www.cdc.gov/coronavirus/2019-ncov/index.html.

TÓM TẮT

Cải thiện hiệu quả của mạng nơ-ron trong chẩn đoán COVID-19VN bằng phương pháp gán trọng số cho các triệu chứng bệnh lý

Bài báo trình bày cách tạo cơ sở dữ liệu về bệnh Covid-19 tại Bệnh viện Bệnh nhiệt đới Trung ương (gọi là cơ sở dữ liệu CovidVN) và sau đó xây dựng mạng neural dựa trên cơ sở dữ liệu này để hỗ trợ chẩn đoán bệnh. Cơ sở dữ liệu CovidVN được xây dựng trên cơ sở xử lý và tổng hợp kết quả xét nghiệm chẩn đoán thực tế của bệnh nhân mắc Covid-19 với số lượng nhiều mẫu và phù hợp với cấu trúc cơ sở dữ liệu bệnh Covid-19 của Hệ thống Y tế Israel (gọi tắt là Cơ sở dữ liệu COVIDIsr). Sau đó, mạng nơron nhân tạo MLP tương ứng với hai cơ sở dữ liệu này sẽ được phát triển bằng hộp công cụ học sâu của phần mềm MATLAB; Kết quả luyện các mạng này và độ chính xác của chúng được so sánh với nhau để đánh giá tương đối chất lượng của CSDL CovidVN. Mặt khác, bài báo trình bày phương pháp gán trọng số tương ứng với các triệu chứng bệnh lý cho các tham số đầu vào của mạng. Kết quả cho thấy, trọng số của các thuộc tính đầu vào tương ứng với các triệu chứng bệnh lý là có ý nghĩa.

Từ khoá: Mạng nơ-ron nhân tạo; Cơ sở dữ liệu virus corona (COVID-19); Perceptron đa lớp (MLP); Gán trọng số.