Densely Connected Neural Network and Permutation Entropy in the Early Diagnostic in COVID Patients

Luz Alexandra Díaz[†], Antonio Ravelo-García[§], Esteban Alvarez[‡], María Fernanda Rodríguez[†], Diego Rodrigo Cornejo[†], Victor Cabrera-Caso[†], Dante Condori-Merma[†], Miguel Vizcardo Cornejo[†]

[†]Escuela Profesional de Física, Universidad Nacional de San Agustín de Arequipa, Perú
[§]Instituto for Technological Development and Innovation in Communications, Universidad de Las Palmas de Gran Canaria, Spain
[‡]Escuela de Física, Universidad Central de Venezuela, Venezuela

Abstract

The COVID-19 pandemic has been characterized by the high number of infected cases due to its rapid spread around the world, with more than 6 million of deaths. Given that we are all at risk of acquiring this disease and that vaccines do not completely stop its spread, it is necessary to continue proposing tools that help mitigate it. This is the reason why it is ideal to develop a method for early detection of the disease, for which this work uses the Stanford University database to classify patients with SARS-CoV-2, also commonly called as COVID-19, and healthy ones. In order to do that we used a densely connected neural network on a total of 77 statistical features, including permutation entropy, that were contrasted from two different time windows, extracted from the heart rate of 24 COVID patients and 24 healthy people. The results of the classification process reached an accuracy of 86.67% and 100% of precision with the additional parameters of recall and F1-score being 80% and 88.89% respectively. Finally, from the ROC curve for this classification model it could be calculated an AUC of 0.982.

1. Introduction

The Coronavirus disease 2019, commonly known as COVID-19, is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), belonging to the coronaviridae family. It was first detected in Wuhan, in December 2019, and has been responsible of the greatest pandemic of the last 100 years [1]. Compared to SARS-CoV-1 and Middle East respiratory syndrome (MERS-CoV) viruses, COVID-19 has lower morbidity and mortality, but has spread faster [10] and studies suggest that most of the transmission is respiratory, with viruses suspended in droplets [9]. According to the World Health Organization, until March 2022, globally there have been about 486 millions of confirmed infection cases, and more than 6 millions of deaths [2]. Despite of the virus has muted in several variants [3–5], product of the high number of infections, the clinical manifestation of COVID-19 are similar to many viral illnesses [6] main symptoms are fever, disnea and cough [7, 8]. It mainly affects the respiratory system, and while we are all at risk of developing serious illness, factors such as age and underlying medical or comorbide conditions, such as cardiovascular disease, diabetes, chronic respiratory disease, or cancer, are determinants of disease severity and progression [11]. With the advancement of technology, mainly the field of artificial intelligence, it has been possible to detect infection cases before the onset of symptoms or in asymtomatic persons, using different data of patients such us clinical variables, blood test, computed tomography, X-ray among others [12–14] and deep learning and machine laerning algorithms with a high degree of accuracy, and specially some of them have used variables obtained from smart devices, which are very convenient for a non invasive diagnosis [15]. Taking into account that after the advent of vaccines, the spread of COVID-19 has slightly decreased, but has not stopped, it is very important to develop useful tools that help to decrease the spread and continuity of this pandemic, that is why this work uses an artificial neural network in order to discriminate and predict COVID-19 cases using a database which has variables from wearable devices.

2. Method

2.1. Preprocessed

The database used is from Stanford University [16]. They conducted a study using data of heart rate and quantity of steps using smart watches and from a smartphone app created by the University, people infected with COVID

registered their symptoms, symptoms onset date and diagnosis date.

Of this database, we used only the heart rate data. Taking into account the origin of the data, the standardization of these was a necessary process. First of all, we did an undersampling per minute by averaging the data and rearranging it. Then, for each infected patient, we located the onset date of their symptoms and extracted two 5-day data windows. The first window was taken two days before the onset of symptoms, corresponding to an early state of COVID, and the second window corresponds to 7 days before the last date considered in the first window, this window would correspond to a healthy state. For the case of healthy patients, random symptom dates were used and the same process was carried out, both windows corresponding to a healthy state [17].

At the beginning, we had 28 COVID patients, but we noticed that 4 of them did not have enough data to meet the required 5-day windows, so these 4 were discarded. In order to have the same number of healthy patients, 24 of the total number of healthy patients were randomly selected under the unique condition that they met the necessary data length.

After the extraction of the two mentioned windows, using the Time Series Feature Extraction Library (TSFEL), which is a Python package for feature extraction on time series data, with that, the program done 60 different process to extract 390 features of each window. Subsequently, we subtracted the characteristics of window of the initial state of COVID with the characteristics of the window of the healthy state. We did that with COVID and healthy patients.

To keep the best features we applied the Wilcoxon–Mann–Whitney test between the vector of features of COVID patients and healthy patients, keeping only features with a p-value below 0.05, which means that there are relevant differences between both population groups. After that test we remained 67 features from 390 for each patient.

Finally, we applied permutation entropy (PE) to both windows, where PE is a measure of the complexity of a time series first introduced by Brand and Pompe as [18]:

$$H(n) = -\sum p(\pi) \log p(\pi) \tag{1}$$

It takes into account the occurrence of patterns π (comparisons of one value with its neighbors) within the data and the regularity of their apparitions denoted by $p(\pi)$. For the actual calculation, the base 2 is used for the logarithm.

We did a mean per day in order to obtain 10 data of PE for each patient. These 10 data were added to the list of characteristics of each patient. We did not subtract the permutation entropy values between both windows since the PE characterizes patterns, so when performing a subtraction of these, information and meaning of the parameter would be lost; therefore, the 10 PE values were preserved. In this way, 77 data were obtained for each patient to constitute the input of the densely connected neural network.

As a final process, the MinMaxScaler function included in the Scikit-Learn machine learning library was used to normalize the data with respect to a maximum and minimum value found in the input data series. Thereafter, the tools provided by Keras were used to transform the arrival variables into categorical variables to ensure a correct binary classification.

2.2. Neural network architecture

The neural network was implemented using the open source Keras library as a core ecosystem of Tensorflow 2. The proposed architecture is based on a sequential model, where densely connected layers were stacked. These stacked layers are: An input layer of dimension 77×1 with 20 nodes, 3 hidden layers with 10, 6 and 3 nodes for each one respectively and an output layer of dimension 2 for the binary classification between patients with COVID and healthy patients. All layers except for the output layer has "ReLU" activation function, and the output layer has "Softmax" function. We used "Categorical Crossentropy" as a loss function and "Adam" as the optimization algorithm, considering as internal metrics "Categorical Accuracy", the mean square error and the own given by the loss function.

The 48 subjects (24 with COVID and 24 healthy patients) were randomly divided into three sets: the training, the validation, and the test set. 70% was used in the training of the model, and the other 30% for the final test set. In addition, 20% of the training data were used for the validation of the model at the end of each epoch in the training phase, as well as for the choice of parameters.

Additionally, the stopping criterion that was used was EarlyStopping thanks to the callbacks integrated in Keras, in this way the evolution of the loss function was monitored, stopping the training when there is no explicit improvement in the convergence of the model.

3. Results

70% of all characteristics vectors of each patient were used as input data for the densely connected neural network. We used the proposed architecture with a batch size of 1 and 40 was the epoch limit.

The loss function, which evolution is described in Figure 1, shows the error made by the densely connected neural network at the end of each epoch. We can see in the figure that the evolution of the loss function is satisfactory in view of the fact that the training and validation curves are following the same behaviour, tending both to the value of

zero. Likewise, the graph does not denote overfitting or underfitting behavior of the model.

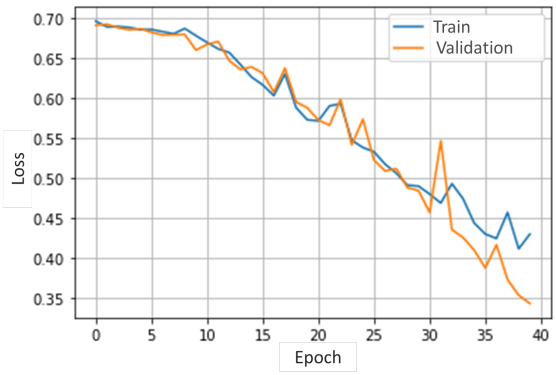

Figure 1. Evolution of the Loss function trough the epochs

Figure 2 shows the evolution of precision through the epochs. Binary precision represents the number of subjects correctly classified in relation to the total number in their current time, so a value closer to one would correspond to a better classification made by the densely connected network. From the figure, we have that both training and validation follow the same trend, that is, there is no divergence, and both tend to a value close to 1.

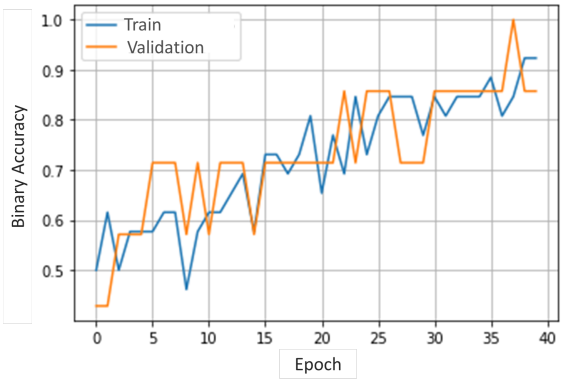

Figure 2. Evolution of the Accuracy trough the epochs

We can handle relevant information about the classifier performance of the densely connected neural network. From it, we can see the total number of true positive, true negative, false positive and false negative classifications and its respective rates, with that we can calculate parameters to describe the performance of the model: precision, accuracy, recall and the F-score. From the Figure 3, we have that the accuracy is 86.67%, the precision is 100%, the recall is 80% and the F-score is 88.89%.

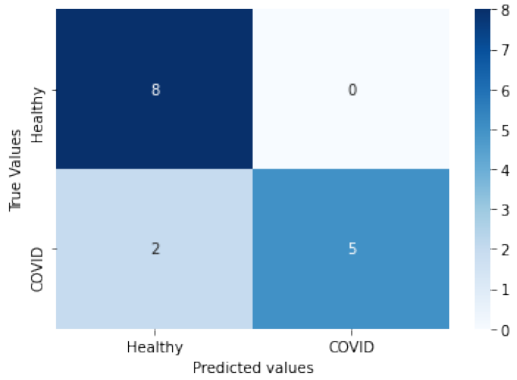

Figure 3. Confusion matrix

To visualize the success rate of the classification made by the densely connected neural network we plotted the corresponding ROC curve where the true positive rate is on the y axis (also called as sensitivity) and the false positive rate is on the x axis (also called 1- specificity). Taking into account that we have the top left corner of the plot as the "ideal" point, which is not very realistic; but, larger area under curve or AUC means better performance of classification.

The ROC curve for the classification is shown in Figure 4. From this we can see that the value of the AUC is really close to 1, where the exact value is 0.982. It means that the classification work done by the densely connected neural network was quite good.

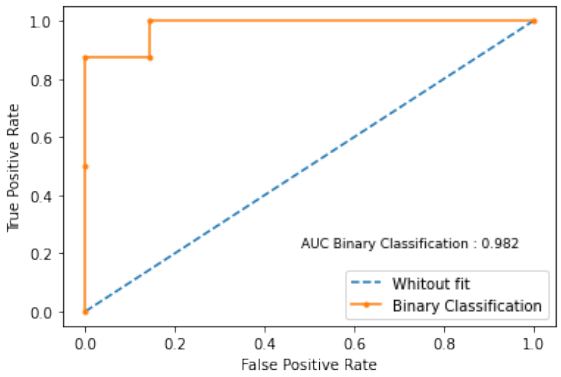

Figure 4. ROC curve

4. Discussion and conclusions

This work used a densely connected neural network, also commonly called as "deep learning", to classify patients in two groups, patients with COVID and healthy patients. For that aim, we used different statistical characteristics of the heart rate and permutation entropy to characterize COVID patients and healthy patients.

The presented architecture has shown great results, talking about accuracy and computing time, although the limited number of patients. We had to use only 48 patients (24 COVID patients and 24 healthy patients) because of the limitation and irregularity of the database used.

Despite the inconveniences presented with the database, the model obtained had a satisfactory evolution in the loss and binary accuracy as we can corroborated trough the graphics obtained. All of them convergent and following a good tendency. Also, from the all performance of classification we have a 86.67% of accuracy and 100% of precision. Both results are good and indicates that the model had a good classification work like we can see in ROC presented, with an AUC really close to 1, that means, with a very good values of specificity and sensitivity.

Regarding related works, we have the work carried out by Skibinska et al., which uses different machine learning algorithms, obtaining different accuracy results, with a minimum of 50% and a maximum of 78%. Our work, by adding the permutation entropy feature, a useful tool for analyzing time series such as heart rate, and implementing a densely connected neural network, achieves an accuracy of 86.7%, signifying a significant percentage improvement.

So, the quality of the obtained results show that the densely connected neural network work optimally. Taking into account that, a good result could be obtained with more data, this type of work could be help to implement a fast and effective automatic early diagnosis of COVID - 19.

Acknowledgements

Virrectorado de Investigación de la Universidad Nacional San Agustin de Arequipa, contrato de subvención IBA-IB-02-2021-UNSA.

References

[1] Voskarides, K. (2022). SARS-CoV-2: tracing the origin, tracking the evolution. *BMC Medical Genomics, 15* , 62.

[2] World Health Organization (2022). Coronavirus (COVID-19) dashboard.

[3] Banoun, H. (2021). Evolution of SARS-CoV-2: Review of mutations, role of the host immune system. *Nephron, 145* (4), 392-403.

[4] Kannan, S. & et al. (2021). Evolutionary analysis of the delta and delta plus variants of the SARS-CoV-2 viruses. *J Autoimmun, 124*, 102715.

[5] Hirabara, S. & et al. (2022). SARS-COV-2 Variants: differences and potential of immune evasion. *Front Cell Infect Microbiol, 11*, 781429.

[6] Salian, V.& et al. (2021). COVID-19 transmission, current treatment, and future therapeutic strategies. *Molecular Pharmaceutics, 18* (3), 754-771.

[7] De Vito, A., Geremia, N., Fiore, V., Princic, E., Babudieri, S. & Madeddu, G. (2020). Clinical features, laboratory findings and predictors of death in hospitalized patients with COVID-19 in Sardinia, Italy. *Eur Rev Med Pharmacol Sci., 24* (14), 7861-7868.

[8] Hoang, V. T., Colson, P., Levasseur, A., Delerce, J., Lagier, J. C., Parola, P., Million, M., Fournier, P. E., Raoult, D., & Gautret, P. (2021). Clinical outcomes in patients infected with different SARS-CoV-2 variants at one hospital during three phases of the COVID-19 epidemic in Marseille, France. *Infect Genet Evol., 95*, 105092.

[9] Meyerowitz, E. A., Richterman, A., Gandhi, R. T., & Sax, P. E. (2021). Transmission of SARS-CoV-2: a review of viral, host, and environmental factors. *Annals of internal medicine, 174* (1), 69–79.

[10] Peeri, N., Shrestha, N., Rahman, M., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W. & Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *Int J Epidemiol, 49* (3), 717-726.

[11] Chen, Y., Klein, S., Garibaldi, B., Li, H., Wu, C., Osevala, N., Li, T., Margolick, J., Pawelec, G. & Leng, S. Aging in COVID-19: vulnerability, immunity and intervention. *Ageing Res Rev., 65*, 101205.

[12] Hasan, A., Al-Jawad, M., Jalab, H., Shaiba, H., Ibrahim, R. & Al-Shamasneh, A. (2021). Classification of Covid-19 coronavirus, pneumonia and healthy lungs in CT scans using q-deformed entropy and deep learning features. *Entropy (Basel), 22*, (5), 517.

[13] Pathan, S., Siddalingaswamy, P. & Ali, T. (2021). Automated detection of Covid-19 from chest x-ray scans using an optimized CNN architecture. *Appl Soft Comput., 104,* 107238.

[14] Li, W., Ma, J., Shende, N. et al. (2020). Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med Inform Decis Mak, 20*, 247.

[15] Gadaleta, M., Radin, J.M., Baca-Motes, K. et al. (2021). Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. *npj Digit. Med., 4*, 166.

[16] Mishra, T., Wang, M., Metwally, A.A. et al. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. Nat. Biomed. Eng. 4, 1208–1220.

[17] J. Skibinska, R. Burget, A. Channa, N. Popescu and Y. Koucheryavy, (2021). COVID-19 diagnosis at early stage based on smartwatches and machine learning techniques. IEEE Access, vol. 9, pp. 119476-119491.

[18] Bandt, Christoph & Pompe, Bernd. (2002). Permutation entropy: a natural complexity measure for time series. *Physical review letters, 88*(17), 174102.

Correspondence

Miguel Vizcardo Cornejo, mvizcardoc@unsa.edu.pe Calle Santa Catalina Nº 117 CP 04000. Arequipa Perú.