Analysis of COVID Patients Employing Approximate Entropy and Deep Learning for Classification and Early Diagnosis

Diego Rodrigo Cornejo[†], Antonio G. Ravelo-García[§‡], María Fernanda Rodríguez[†], Luz Alexandra Díaz[†], 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
[‡] Interactive Technologies Institute (ITI/LARSyS and ARDITI), 9020-105 Funchal, Portugal

Abstract

Due to its rapid propagation and enormous number of infected people, COVID-19 is the greatest pandemic in the past 100 years, with millions of deaths. The need for accessible, quick, and non-invasive diagnostic techniques persists despite a decline in cases recently. Because of this, in the current work we develop a densely connected neural network that uses heart rate data to identify between patients with COVID and healthy individuals. The Stanford University database was used, which underwent a feature extraction and the usage of approximation entropy. With an accuracy of 93% and an AUC of 0.956, the results demonstrated to be more than good at categorization, supporting the usefulness of this approach for the accurate identification of COVID cases.

1. Introduction

The SARS-CoV-2 is the virus that causes coronavirus illness. It first originated in Wuhan, China, in December 2019, when a newly discovered β-coronavirus, later known as coronavirus disease 2019 (COVID-19), caused a wave of pneumonia cases [1]. Since it first emerged, the outbreak has expanded over the globe. On January 30, 2020, it was considered a public health emergency [2], affecting over 185 nations, with over 7,145,800 cases reported and 407,067 deaths as of June 9, 2020 [3, 4].
This disease is characterized by fever, coughing, dyspnea, lethargy, headaches, muscle soreness, and diarrhea [5, 6]. The disease primarily affects the respiratory system, and while everyone has the same chance of developing a severe case, the severity and course of the illness are determined by factors like age and underlying medical conditions or co-morbid conditions like cancer, diabetes, cardiovascular disease, or chronic respiratory diseases [7]. The diagnosis is challenging because, despite the illness's mutation, the symptoms are similar to those of other respiratory conditions. For that reason there is a need of tools that facilitates for specialists to diagnose COVID-19.
Many studies have currently used artificial intelligence to identify cases of infection even before symptoms appear or in asymptomatic individuals using a variety of clinical variables, including blood tests, computed tomography, X-rays, heart rate, and others [9,10,12]; some of these studies use smart devices to collect data non-invasively [11, 12].
The present work uses approximate entropy and feature extraction over heart rate registers to analyze and predict COVID-19 cases using data from portable devices. This constitutes a valuable potential diagnostic tool, given the reality of the disease and the fact that, despite its decreased fatal incidence, there is still a growing interest in research on it for the development of useful tools for its early diagnosis and effective treatment.

2. Method

2.1. Preprocessing

For this work, the Stanford University database [13] was used. This database consists on measurements of the heart rate and number of steps, also the symptoms, date of formal diagnosis and recovery date are reported.
From this database, we considered the heart rate and it is mandatory to standardize the data since it comes from different devices and sampling times. So, first, a subsampling was performed, averaging the data in windows of 1 minute. For each individual patient that was infected the date of onset of their symptoms was located and then we extracted two 5-day windows of data. The first window corresponds to the heart rate of the patients starting two days before the onset of their symptoms, that way we consider the early stage of COVID. The second window starts 7 days before beginning of the first window, this one corresponds to a healthy state. For healthy patients the windows are chosen at random, the two windows corresponding to a healthy

state [14].
In order to extract statistically significant features for model training we used the Time Series Feature Extraction Library (TSFEL) to extract 390 features of the 7 day period of the windows. Also we computed the Approximate Entropy (ApEn), used to measure the irregularity and complexity of a set of temporal data, first defined by Steve Pincus as [15], where N is the number of elements in the time-series:

$$\phi^m(r) = \frac{1}{n}\sum_{i=1}^{n} log(C_i^m(r)) \tag{1}$$

$$ApEn(m, r, N)(u) = \phi^m(r) - \phi^{m+1}(r) \tag{2}$$

Here m is the embedding dimension, r is a threshold and A_i and B_i are the measures of proximity between embedding vectors in m and $m+1$ dimensions respectively. We selected $m = 5$ and $r = 0.5$ as the parameters for this work
The Wilcoxon-Mann-Whitney test was then used to compare the characteristics and ApEn values of COVID patients with healthy patients. Only those samples with a p-value less than 0.05 were retained. There were just 33 features left for each subject after this. It is important to stress that following the test, all 10 ApEn values remained. We applied a data augmentation procedure to the heart rate data in order to grow our database. To generate this data augmentation, we first separated the original data and for this we used 45 healthy individuals and 12 Covid patients. The model was tested using the remaining data. For the process of data augmentation we used a Gaussian factor of 0.06 and the COVID group was multiplied by a factor of 6 and the health group by a factor of 4.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

Finally, a standardized normalization was used for the 33 total data (10 corresponding to the ApEn (one value per day) and 23 corresponding to the temporal characteristics). The cases were labeled as COVID and CONTROL and the labels were treated with one hot encoding.

2.2. Neural network architecture

The open source Keras library as an ecosystem of Tensorflow 2 was used to create the artificial neural network. The main architecture consists on densely connected layers piled sequentially. These layers were: an input layer of dimension 33x1 with 20 nodes, two hidden layers, each with ten nodes, and an output layer with 2 nodes consideiring the binary classification of patients with COVID-19 and healthy patients. Each layer used the activation functions "GeLu", "GeLu", and "Tanh" respectively. The output layer was set to use the "Softmax" function. We considered "Categorical Crossentropy" as a loss function and "Adam" as the optimization algorithm, considering as internal metrics "Categorical Accuracy" and the evolution given by the loss function.
To train the model we used 309 examples and then 56 examples (from the original data that was not used to the data augmentation process) were used to test it. Furthermore, the stopping criterion selected was Early Stopping thanks to the callbacks integrated in Keras, in this way evolution of the loss function was monitored and controlled, stopping the training if there was no explicit improvement in the convergence of the model presented.

3. Results

It is possible to observe in Figure 1 the evolution of the loss function, which depicts the error made by the densely connected neural network at the end of each epoch. In the figure we can see a satisfactory evolution because training and validation curves are following the same behavior in a smooth way quickly approaching zero, that means that the parameter that measures the miss-classification made by the neural network gets closer each time more to zero as epochs goes by. Likewise, the graph does not denote overfitting or underfitting behavior of the model.

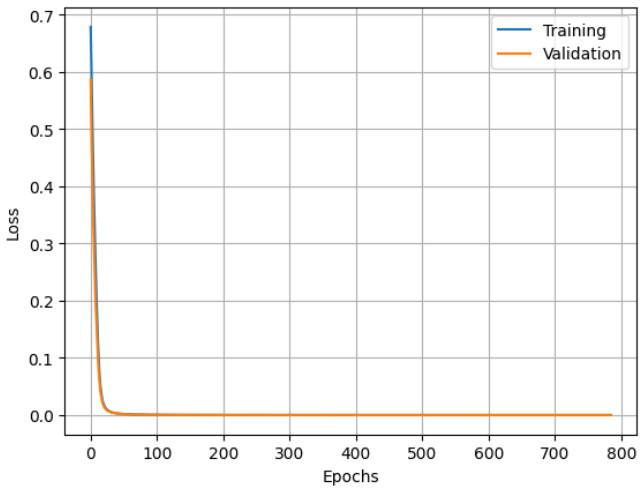

Figure 1. Evolution of the Loss function through the epochs

The figure 2 shows the evolution of categorical accuracy trough the epochs. This represents the number of subjects correctly classified respect to the total number, so that values closer to the number one correspond to a better work of classification made by the network. From the figure it is

possible to see that both curves follow the same trend without a divergence between those. Both also tend rapidly to values closer to one as epochs pass.

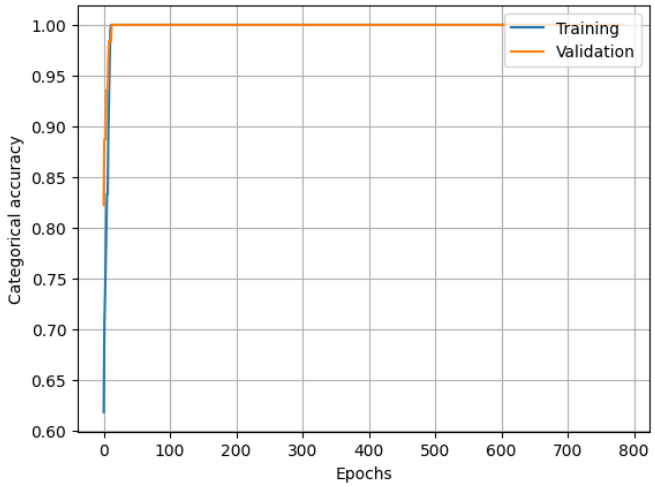

Figure 2. Evolution of the Categorical Accuracy through the epochs

In figure 3 we show the confusion matrix of the model where we can extract important information about the classification performance of the densely connected neural network. In this matrix the total number of true positives, true negatives, false positives and false negatives are represented together with their respective rates. We can then use the information of this matrix to compute parameters that best describe the performance of the model (accuracy, precision, recovery and the F-score). From the observed values, there is a total accuracy of 93%, precision of 100% and 75%, recall of 91% and 100%, and an f-1 score of 95% and 86% in the classification of control patients and with COVID, respectively.

We can illustrate the success rate of the classification performed by the densely connected neural network with the help of a ROC curve. In this curve, the y axis corresponds to the true positive rate called "sensitivity" and the x axis to the false positive rate called "1-specificity". An "ideal" point on this plot would be the one with values close to one in specificity and sensitivity, so the closer the points are to the top left corner, the better the classification job was done, that is, larger area under curve (called AUC) means better classification performance. Figure 4 is the plot of the ROC curve for the classification. Here it is possible to see that the value of the AUC is really close to 1 with an exact value of 0.956. That is a proof of the good classification work done by the network.

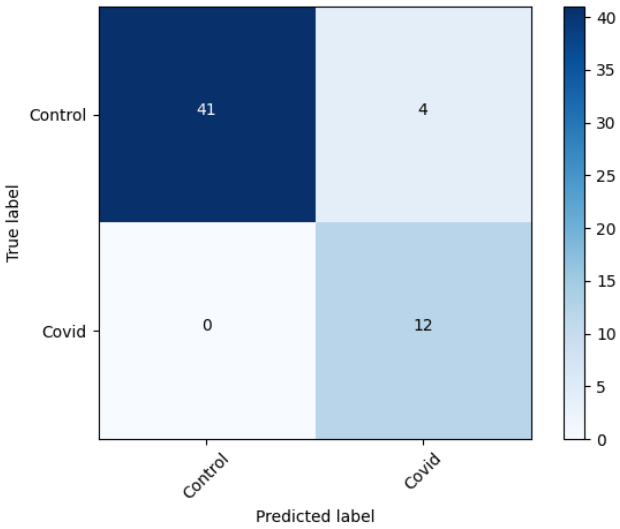

Figure 3. Confusion matrix

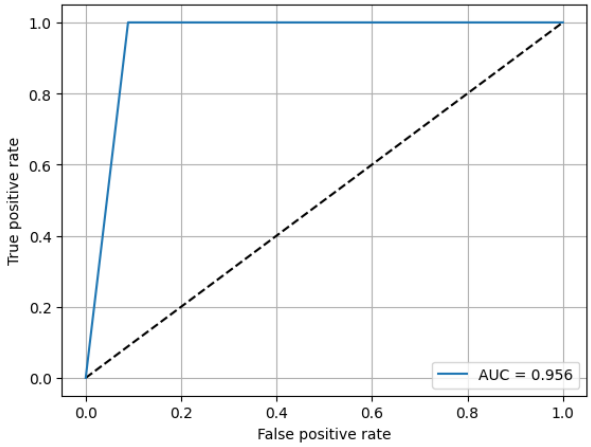

Figure 4. ROC curve

4. Discussion and conclusions

The present work made use of a densely connected neural network or commonly referred to as "deep learning" as a tool to distinguish between COVID patients and healthy patients. We performed preprocessing utilizing several statistical features of the heart rate and used approximate entropy (ApEn) on it to arrive at a good categorization result. The proposed architecture shows great results despite the limited number of patients, taking into account the relationship between precision and computation time. The work only used 48 patients, where 24 corresponded to patients with COVID and 89 healthy patients. Despite this data drawback, the model had a good evolution in loss and categorical precision as we could see in the graphs of the work. All of them converge and follow the same trend,

the loss converging to values close to 0 and the accuracy to 1. Also, from the total work of classification the model reach an accuracy of 93%, which is a very good result that denotes a satisfactory work of classification that could be corroborated in the ROC curve presented, with an AUC value of 0.956, very close to 1, with adequate values of specificity and sensibility.

Regarding related works, it is important to mention the work of Skibinska et al., which tests various machine learning algorithms with different accuracy results, reaching a maximum of 78% [16]. Another work that is relevant to mention is the one presented by Díaz et al., where they make use of deep learning and permutation entropy, obtaining an accuracy of 86.67% in the classification [17]. This work also makes use of a densely connected network and information entropy due to its spectacular performance when working with time series; however, the entropy used is the approximate entropy (ApEn) and the discarding of features is more rigorous. This, together with a good implementation of data augmentation, gives us an accuracy of 93%, thus consisting of a significant percentage improvement compared to previous works.

In conclusion, the excellent quality of the results demonstrates the efficacy of using approximate entropy, feature extraction, and the densely connected neural network as a good classification tool.

Acknowledgements

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

References

[1] Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., ... & Yan, Y. (2020). The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak–an Update on the Status. *Military medical research, 7*, 1-10.

[2] Sohrabi, C., Alsafi, Z., O'neill, N., Khan, M., Kerwan, A., Al-Jabir, A., ... & Agha, R. (2020). World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19). *International journal of surgery, 76*, 71-76.

[3] WHO: World Health Organization, 2020. Coronavirus disease (COVID-2019) situation Reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation- reports/ (accessed 09 June 2020)

[4] JHU: John Hopkins University, 2020. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://www.coronavirus.jhu.edu/map.html (accessed 09 June 2020).

[5] Weng, L. M., Su, X., & Wang, X. Q. (2021). Pain Symptoms in Patients with Coronavirus Disease (COVID-19): A Literature Review. Journal of Pain Research, 147-159.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Díaz, L. A., Ravelo-García, A., Alvarez, E., Rodríguez, M. F., Cornejo, D. R., Cabrera-Caso, V., ... & Cornejo, M. V. (2022, September). Densely Connected Neural Network and Permutation Entropy in the Early Diagnostic in COVID Patients. In *2022 Computing in Cardiology (CinC)* (Vol. 498, pp. 1-4). IEEE.

[13] 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.

[14] 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.

[15] Pincus, S. M. (1991). Approximate Entropy as a Measure of System Complexity. *Proceedings of the National Academy of Sciences, 88*(6), 2297-2301.

[16] Skibinska, J., Burget, R., Channa, A., Popescu, N., & Koucheryavy, Y. (2021). Covid-19 Diagnosis at Early Stage Based on Smartwatches and Machine Learning Techniques. IEEE Access, 9, 119476-119491.

[17] Díaz, L. A., Ravelo-García, A., Alvarez, E., Rodríguez, M. F., Cornejo, D. R., Cabrera-Caso, V., ... & Cornejo, M. V. (2022, September). Densely Connected Neural Network and Permutation Entropy in the Early Diagnostic in COVID Patients. In *2022 Computing in Cardiology (CinC)* (Vol. 498, pp. 1-4). IEEE.

Correspondence:
Miguel Vizcardo Cornejo, Av. Independencia s/n Ciudad Universitaria, Edificio de Fisica, Laboratorio Nro. 305, Arequipa 04001, Perú. Email; *mvizcardoc@unsa.edu.pe*