Classification of COVID-19 associated symptomatology using machine learning

Julian Andres Ramirez-Bautista [a], Silvia L. Chaparro-Cárdenas [a], Wilson Gamboa-Contreras [a], William Guerrero-Salazar [a] & Jorge Adalberto Huerta-Ruelas [b]

[a] Departamento de Investigación, Fundación Universitaria de San Gil-Unisangil, San Gil, Colombia. jramirez@unisangil.edu.co, schaparro@unisangil.edu.co, wgamboa@unisangil.edu.co, wguerrero@unisangil.edu.co

[b] Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada-Instituto Politécnico Nacional, Querétaro, México. jhuertar@ipn.mx

Abstract

The health situation caused by the SARS-Cov2 coronavirus, posed major challenges for the scientific community. Advances in artificial intelligence are a very useful resource, but it is important to determine which symptoms presented by positive cases of infection are the best predictors. A machine learning approach was used with data from 5,434 people, with eleven symptoms: breathing problems, dry cough, sore throat, running nose, history of asthma, chronic lung, headache, heart disease, hypertension, diabetes, and fever. Based on public data from Kaggle with WHO standardized symptoms. A model was developed to detect COVID-19 positive cases using a simple machine learning model. The results of 4 loss functions and by SHAP values, were compared. The best loss function was Binary Cross Entropy, with a single hidden layer configuration with 10 neurons, achieving an F1 score of 0.98 and the model was rated with an area under the curve of 0.99 aucROC.

Keywords: computer-aided diagnosis: COVID-19; disease diagnosis; machine learning; artificial neural networks.

Clasificación de la sintomatología asociada a la COVID-19 mediante aprendizaje automático

Resumen

La situación sanitaria provocada por el coronavirus SARS-Cov2 plantea grandes retos a la comunidad científica. Los avances en inteligencia artificial son un recurso muy útil, pero es importante determinar qué síntomas presentados por los casos positivos de infección son los mejores predictores. Se utilizó un enfoque de aprendizaje automático con datos de 5.434 personas, con once síntomas: problemas respiratorios, tos seca, dolor de garganta, secreción nasal, antecedentes de asma, pulmón crónico, dolor de cabeza, enfermedad cardíaca, hipertensión, diabetes y fiebre. Basado en datos públicos de Kaggle con síntomas estandarizados por la OMS. Se desarrolló un modelo para detectar los casos positivos de COVID-19 utilizando un modelo simple de aprendizaje automático. Se compararon los resultados de 4 funciones de pérdida y por valores SHAP. La mejor función de pérdida fue la Entropía Cruzada Binaria, con una configuración de una sola capa oculta con 10 neuronas, logrando una puntuación F1 de 0,98 y el modelo fue calificado con un área bajo la curva de 0,99 aucROC.

Palabras clave: diagnóstico asistido por ordenador; COVID-19; diagnóstico de enfermedades; aprendizaje automático; redes neuronales artificiales.

1 Introduction

Electronic medical data are ubiquitous and available in large quantities with high accuracy due to the increasing availability of such data from a variety of sources, including clinical institutions, individual patients, insurance companies, pharmaceutical industries, and others, highlighting great opportunities for understanding risk factors, and disease spread, continuous health monitoring, among others; enabling targeted prevention approaches [1]. Advances in computer science – especially in machine learning with new and improved methods for data collection and storage – have shown increasing interest related to healthcare data analysis [2,3].

The emergence and spread of the SARS-Cov2 coronavirus, which produces the disease called COVID-19 [4,5] has become a particular challenge for healthcare professionals and the general population [6]. The disease placed a great burden on healthcare systems, and total confinements generated

losses, as many of the people were not infected, but the follow-up of positive cases becomes a complex task [7].

Many studies have focused on identifying infected individuals to isolate them and allow non-infected individuals to work regularly. The use of clinical symptoms is essential to optimize the identification of infected individuals.

In this sense, researchers have developed predictive models that combine several features such as clinical symptoms, laboratory tests, among others. Also, models to detect possible contagion and estimate the risk of infection, and classify the population to help medical personnel and countries' economies [8,9]. References [8,10] predict a SARS-CoV-2 infection by asking 8 basic questions of which 5 refer to symptomatology (fever, cough, sore throat, shortness of breath, and headache) obtaining an accuracy of 90%. Chen et al. explored the distributions of comorbidities and symptoms, in addition to laboratory test results, to correlate between non-severe and severe types of COVID-19, they were able to identify key features between both clinical types using Machine Learning, as an accurate diagnostic decision support tool [11]. Ahamad et al. developed a model using supervised machine learning algorithms to identify features that predict the diagnosis of COVID-19. Using an algorithm called XGBoost, they obtained an accuracy of over 85% in predicting and selecting features that correctly indicate COVID-19 status, indicating that the most frequent and significant predictive symptoms are fever (41.1%), cough (30.3%), lung infection (13.1%) and nasal discharge (8.43%) [12]. Using a dataset with similar features, other models predicting COVID-19 disease with an area under the curve of 0.90 auROC have been reported using a gradient boosting machine built with decision tree base learners as algorithm [8]. Another approach is reported by Khanday et al. where they perform experiments with various algorithms such as random forest, stochastic gradient boosting, decision trees, to classify into four classes, COVID, SARS, ARDS and both (COVID, ARDS), 212 labeled clinical reports, obtaining that logistic regression and Naïve Bayesian multinomial classifier give excellent results by having an accuracy of 94% and 96.2% [13].

Clinically, COVID-19 disease is complex and manifests itself through a limited number of symptoms such as fever, cough, intense headache, among others [8,11]. If these parameters are analyzed with systems based on machine learning algorithms, it is possible to fight this virus and other future viruses by continuously monitoring individuals to improve detection, isolation, and provide disease control recommendations [14].

The study presents the use of machine learning algorithms for COVID-19 detection using the symptoms and physical conditions of 5,434 people with and without the disease. A Keras-Tensor Flow Neural Network was used. The results of four loss functions, based on the performance indicators, F1 score, and area under the curve, were compared to learn the behavior of the neural network as a basis for integrate them into systems that allow preliminary detection of the disease.

The highest F1 classification score using 11 features was 0.98 with an area under the curve of 0.99 aucROC, using the results of the SHAP values, the less relevant features

considered by the model were eliminated, obtaining a feature simplification of 63%, varying the performance by 2 to 3% depending on the metric referred, demonstrating the usefulness of knowing the importance of the features within a classification model, for its simplification without affecting the performance.

2 Materials and methods

2.1 Study data

The dataset used contains the records of 5,434 people, obtained from the Kaggle platform for experimenting with machine learning models. From these data, a model is developed that predicts COVID-19 scores using five binary features: asthma, chronic lung, heart disease, diabetes, and hyper tension; and six initial clinical symptoms: Breathing problem, fever, dry cough, sore throat, runny nose, and headache.

The training-validation set consisted of records from 1,051 individuals without the disease and 4,383 individuals with the disease. The following table describes each of the features of the dataset used by the model where the greatest number of individuals present dry cough and fever, respectively (Table 1).

After an exploratory analysis, no null data or empty cells were found. The figures show the number of positive and negative cases for each characteristic for each class. Fig. 1(a) shows a large number of people presenting cough, fever, sore throat and breathing problem for the COVID-19 positive class, the other features show a balance between presence and non-presence.

On the other hand, the negative COVID-19 class only shows a considerable imbalance in people with breathing problem and sore throat (Fig. 1(b)).

Considering this is an experimental stage of algorithms testing, the UNISANGIL ethics committee determined that the public health dataset used in this study does not require approval for analysis. The development of systems for constant monitoring of physiological parameters supports public health efforts for the monitoring and control of communicable and no communicable diseases.

Table 1.
Features of the data set used by the model in this study

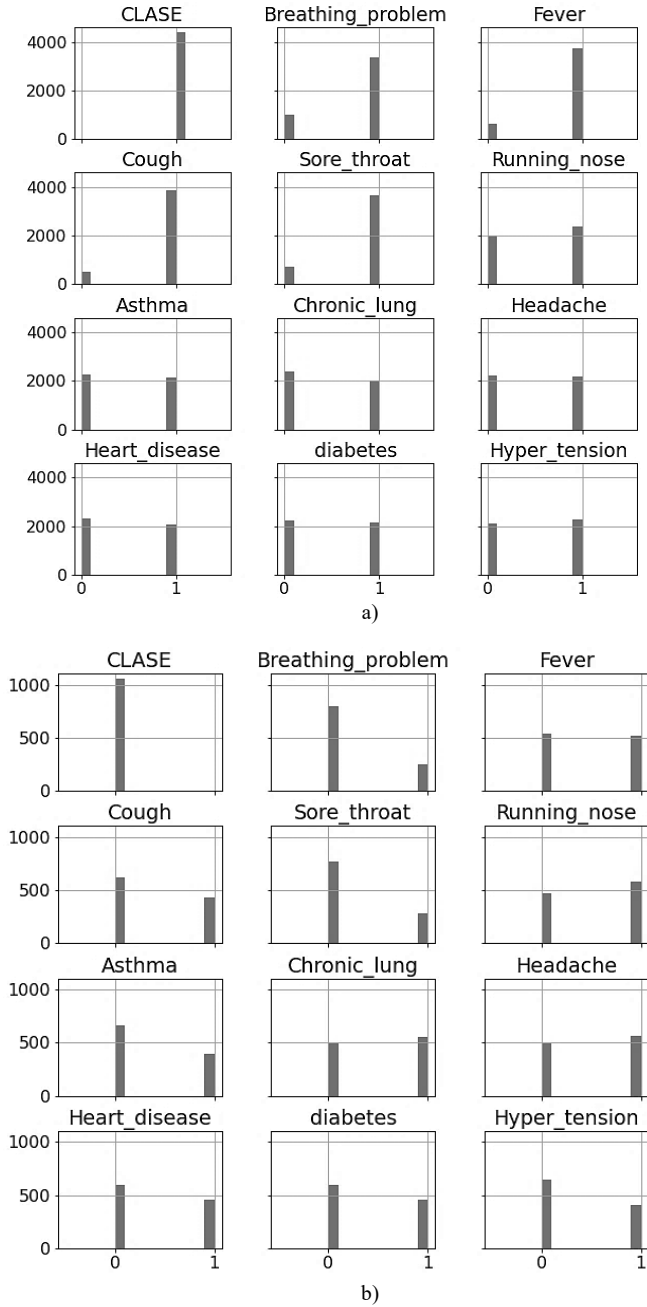| Feature	Total		COVID-19 Negative		COVID-19 Positive	
	N	%	n	%	n	%
	5,434		1,051	19.30	4,383	80.60
Breathing problem	3,620	66.6	251	6.9	3,369	93
Fever	4,273	78.6	516	12	3,757	87.9
Dry Cough	4,307	79.2	429	9.9	3,878	90
Sore throat	3,953	72.7	284	7.1	3,669	92.8
Running nose	2,952	54.3	577	19.5	2,375	80.4
Asthma	2,514	46.2	390	15.5	2,124	84.4
Chronic lung	2,565	47.2	557	21.7	2,008	78.2
Headache	2,736	50.3	559	20.4	2,177	79.5
Heart disease	2,523	46.4	459	18.1	2,064	81.8
Diabetes	2,588	47.6	457	17.6	2,131	82.3
Hyper tension	2,663	49.0	405	15.20	2,258	84.79

Source: The Authors

a)

b)

Figure 1. Features distribution grouped by class. a) COVID-19 positive. b) COVID-19 negative.
Source: The Authors

Table 2.
Sample Dataset

Dataset	Number of individuals	Class
Xtrain	3,423	COVID-19 Positive
Ytrain	924	COVID-19 Negative
Xtest	850	COVID-19 Positive
Ytest	237	COVID-19 Negative

Source: The Authors

The Keras Dense class was used, as a building block of a fully connected layered model, with different loss function described as below:

Binary Cross Entropy also known as log loss, is a loss function used in binary classification tasks, which shows the negative mean of the logarithm of the predicted probabilities of each class, penalizing the probabilities as a function of the distance to the expected value, using the given mathematical formulation [15] (1).

$$Log\ loss = \frac{1}{N}\sum_{i=1}^{N} -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i)) \tag{1}$$

Poisson Loss from the combination of loss frequency and loss severity estimates the loss distribution [16], taking the form of (2).

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N} (\hat{y}_i - y_i * log(\hat{y}_i)) \tag{2}$$

Mean Squared Error is the sum of the squared distances between the true values and the predicted values, greatly penalizing outliers. Due to its sensitivity, regardless of the sign, the values are always positive and 0.0 is the perfect value of the differences [17]. Mathematically it can be expressed as (3).

$$MSE = \frac{\sum_{i=1}^{n}(y_i - y_i^p)^2}{n} \tag{3}$$

Huber Loss presents a lower sensitivity to outliers, considering that when the error is large the absolute error is obtained, which becomes quadratic as the error decreases [18]. Combine the mean square error and the mean absolute error. Its mathematical formulation (4).

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & for\ |y - f(x)| \le \delta, \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{4}$$

2.2 Experiment setup and design

Dataset was classified into two groups: COVID-19 positive and COVID-19 negative. This gives a typical binary classification to know whether people were infected or not. To evaluate the performance of the approach used, the training and test data set were divided into 80% and 20%, respectively. (Table 2).

The process of experimentation with the algorithms was based on Keras Python 3.6 under a win10 operating system. The hardware used for the experiments had an i5-7300HQ CPU, 8G RAM and NVIDIA GeForce GTX 1050 GPU.

The tensor flow was used as the backend with 11 input neurons, a variable number of hidden layers, and one neuron in the output layer. The dataset used had 11 input features and a binary output.

A sequential model with fully connected layers is used, where the dimensions were defined according to the number of inputs, the output, and the variation of the hidden layer, aiming for a small and efficient model. The activation function used was sigmoid. In addition, a comparison with different loss functions was performed. The optimizer used throughout the experimental phase was Adam and the metric used to judge the performance of the neural network was F1 score and auROC parameter from ROC curves. Moreover, considering that the data set is very large, the batch size is used. The characteristics are summarized in Table 3 below.

Table 3.
Features of the Model

Parameter	Value
Optimizer	Adam
Activation Function	Sigmoid
Loss Function	Binary Cross Entropy, Poisson, Mean Squared Error, Huber Loss
Epoch	300
Batch Size	10

Source: The Authors

Table 4.
Comparative results of the loss functions used in the study with a 5-fold cross-validation, in a model with a hidden layer of 10 neurons

Loss functions	1st	2nd	3rd	4th	5th	F1 Score Average	F1 Score Std. Dev
Binary Cross Entropy	0.98	0.98	0.98	0.98	0.98	**0.98**	**+/-0.08%**
Poisson	0.97	0.97	0.97	0.97	0.98	0.97	+/-0.28%
Mean Squared Error	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.13%
Huber Loss	0.97	0.97	0.97	0.97	0.97	0.98	+/- 0.21%
Experiment Features					1 hidden layer and 10 neurons		

Source: The Authors

Table 5.
Comparative results of the loss functions used in the study with a 5-fold cross-validation, in a model with a hidden layer of 5 neurons

Loss functions	1st	2nd	3rd	4th	5th	F1 Score Average	F1 Score Std. Dev
Binary Cross Entropy	0.98	0.98	0.98	0.98	0.98	**0.98**	**+/- 0.34%**
Poisson	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.18%
Mean Squared Error	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.19%
Huber Loss	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.15%
Experiment Features					1 hidden layer and 5 neurons		

Source: The Authors

Table 6.
Comparative results of the loss functions used in the study with a 5-fold cross-validation, in a model with two hidden layers of 5 neurons each

Loss functions	1st	2nd	3rd	4th	5th	F1 Score Average	F1 Score Std. Dev
Binary Cross Entropy	0.97	0.97	0.98	0.98	0.98	**0.98**	**+/- 0.38%**
Poisson	0.96	0.96	0.96	0.96	0.96	0.96	+/- 0.15%
Mean Squared Error	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.10%
Huber Loss	0.97	0.97	0.97	0.97	0.97	0.97	+/- 0.16%
Experiment Features					2 hidden layers of 5 neurons		

Source: The Authors

3 Results

The model trained with data from 5,434 individuals, 19.30% negative and 80.63% positive for COVID-19 disease. It was validated with a 5-fold cross-validation to ensure that the results are independent of the partition between training and test data (Tables 4-6).

The F1-Score of the model with a hidden layer and 10 neurons show a similar behavior with the four loss functions with a mean of 0.97 and standard deviation of +/-0.005.

However, the best model was obtained using the Binary Cross Entropy error function achieving a result of 0.98.

The model was scored on the test set using auROC across different thresholds, including false-positive rate, false-negative rate, and overall accuracy. As seen in Tables 7 to 9, the best performance was obtained using the Binary Cross-Entropy loss function with a single hidden layer configuration with 10 neurons.

Although the behavior of the models was similar considering the F1-Score, the aucROC shows differences of 7 percentage points as shown in Tables 7 to 9. The best result obtained, considering this metric, was the one using the Binary Cross Entropy error function with a value of 0.99, regardless of the number of layers and neurons (Fig. 2).

The metrics of all the ROC curves in this study were calculated using the sklearn.metrics module.

The SHapley Additive exPlanations (SHAP) method introduced by Lundberg and Lee [19], whose origins are in game theory, is used to learn the relevant model features. Since it estimates the differences between models with subsets of the feature space, it allows interpreting the predictions of machine learning models using SHAP values, which estimate the contribution of each feature in the model prediction. The SHAP method for interpreting a model uses additive features attribution, where additive features refer to input variables. Thus, it represents the classification result as the sum of the contribution of each feature, as (5).

$$g(z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z_j' \tag{5}$$

Where g is the explanation model, z' is the simplified features vector, M is the maximum simplified features size and $\phi j \in R$ is the feature attribution for a feature j [20,21].

Thus, it is obtained that the most important features considered by the model are those summarized in the SHAP graph in Fig. 3. The presence of breathing problem, cough, fever, and sore throat were key predictors of the disease presence. The features that showed a low impact in almost all cases were runny nose, Asthma, diabetes, chronic lung disease, hypertension, and headache.

Table 7.
Comparative results of the auROC of 5-fold cross-validation, in a model with a hidden layer of 10 neurons

Loss functions	1st	2nd	3rd	4th	5th	auROC Average	auROC Std. Dev
Binary Cross Entropy	0.99	0.99	0.99	0.99	0.99	**0.99**	**+/-0.00%**
Poisson	0.90	0.90	0.89	0.92	0.93	0.91	+/-0.01%
Mean Squared Error	0.92	0.93	0.91	0.92	0.92	0.92	+/-0.01%
Huber Loss	0.92	0.91	0.92	0.92	0.93	0.92	+/-0.01%
Experiment Features	1 hidden layer of 10 neurons						

Source: The Authors

Table 8.
Comparative results of the auROC of 5-fold cross-validation, in a model with a hidden layer of 5 neurons

Loss functions	1st	2nd	3rd	4th	5th	auROC Average	auROC Std. Dev
Binary Cross Entropy	0.99	0.99	0.99	0.99	0.99	**0.99**	**+/-0.00%**
Poisson	0.91	0.90	0.91	0.92	0.90	0.91	+/-0.01%
Mean Squared Error	0.92	0.92	0.93	0.92	0.93	0.92	+/-0.00%
Huber Loss	0.92	0.92	0.92	0.92	0.91	0.92	+/-0.00%
Experiment Features	1 hidden layer of 5 neurons						

Source: The Authors

Table 9.
Comparative results of the auROC of 5-fold cross-validation, in a model with two hidden layers of 5 neurons each

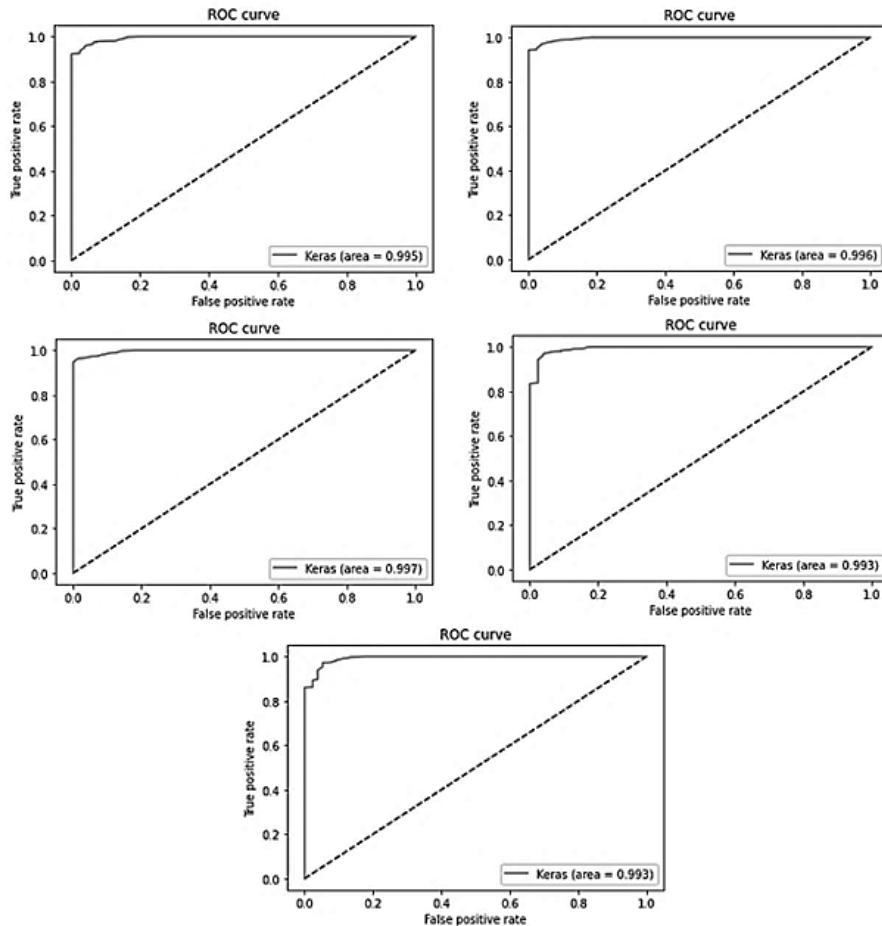| Loss functions	1st	2nd	3rd	4th	5th	auROC Average	auROC Std. Dev
Binary Cross Entropy	0.93	0.93	0.99	0.99	0.99	**0.97**	**0.03%**
Poisson	0.88	0.90	0.89	0.87	0.94	0.90	0.02%
Mean Squared Error	0.92	0.91	0.91	0.92	0.92	0.92	0.00%
Huber Loss	0.91	0.92	0.90	0.92	0.92	0.91	0.01%
Experiment Features	2 hidden layers of 5 neurons						

Source: The Authors

Figure 2. ROC curves showing the performance of the model using the binary cross-entropy loss function, on the test sets in the 5-fold cross-validation.
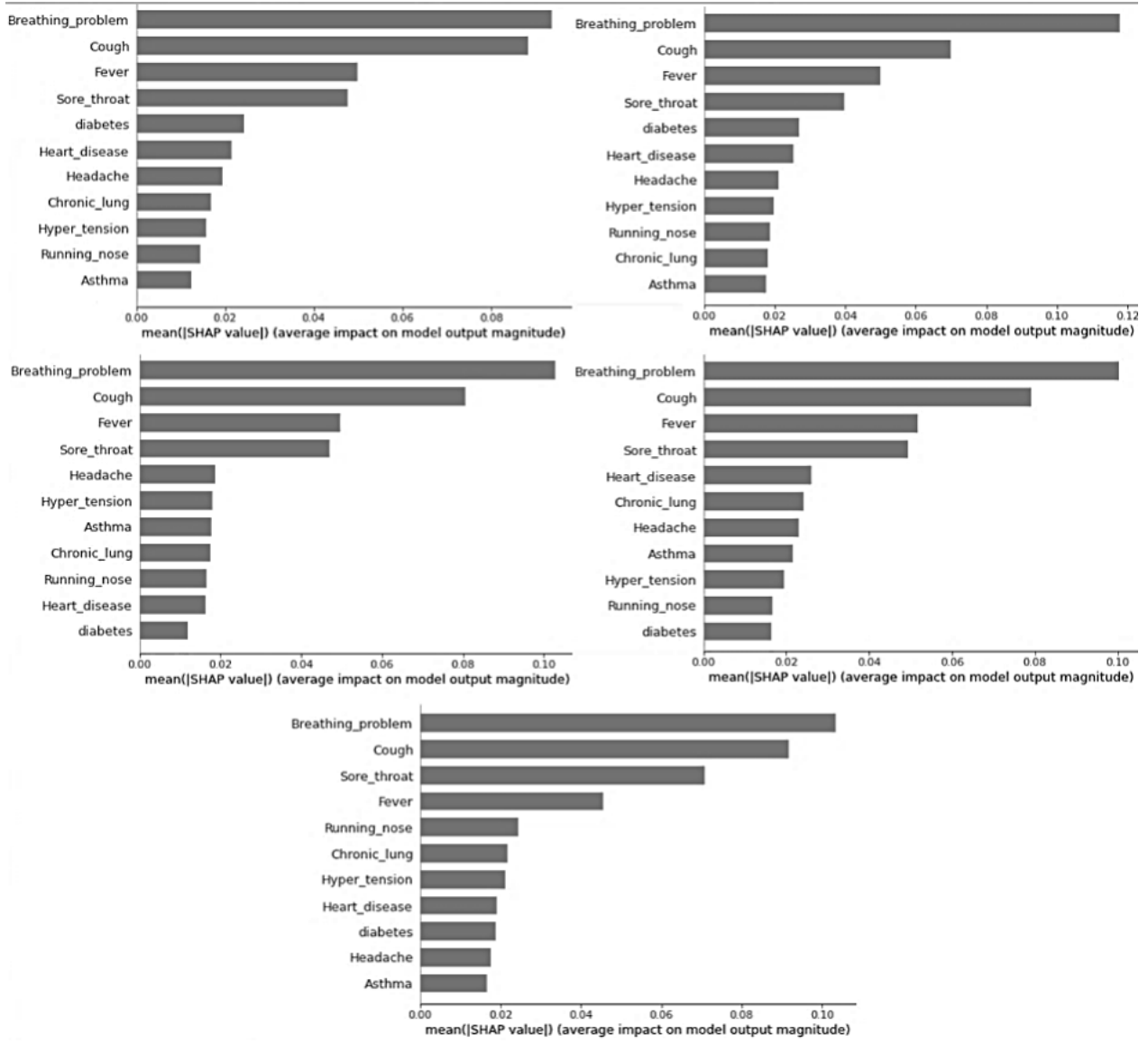Source: The Authors

Figure 3. SHapley Additive exPlanations (SHAP) graph showing the important features considered by the ANN model to predict the diagnosis of COVID-19, in the 5-fold cross-validation.
Source: The Authors

Table 10.
Less relevant features considered by the ANN model to predict the diagnosis of COVID-19, in the 5-fold cross-validation

	1st	2nd	3rd	4th	5th
Less relevant features	Chronic lung	Hyper tension	Chronic lung	Asthma	Heart disease
	Hyper tension	Running nose	Running nose	Heart disease	Diabetes
	Running nose	Chronic lung	Heart disease	Running nose	Headache
	Asthma	Asthma	Diabetes	Diabetes	Asthma

Source: The Authors

Considering the results of the SHAP values in the 5-fold cross-validation, the least relevant features in each fold are shown in Table 10. The least relevant features were running nose and asthma, which in 4 of the 5 folds were common.

The model is recalculated using the features of the best configuration, but the number of input neurons is modified, leaving those corresponding to running nose and asthma. As a result, by removing the least relevant features, an F1-Score of 0.97+-0.12 and an aucROC of 0.98 are obtained, with variations of 1%. This is a negligible variation considering the elimination of features from the model.

Finally, the model is obtained again by eliminating the following 3 less relevant features (Diabetes, Chronic Lung and Hypertension), having a total of 6 inputs. The F1-Score result was 0.96 +/-0.34% and the aucROC was 0.96, obtaining a difference of 2 and 3% for each performance index.
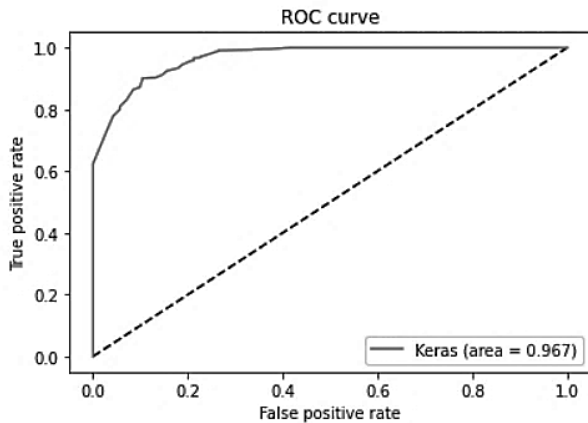
Figure 4. ROC curve showing the performance of the model using six features
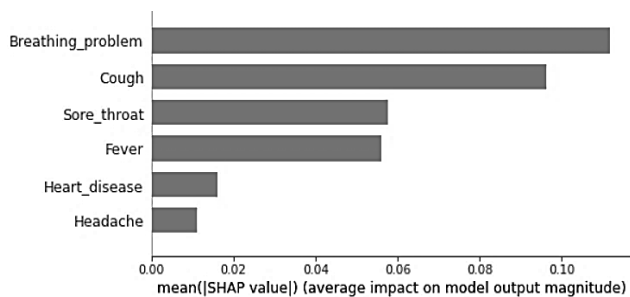Source: The Authors

Figure 5. SHapley Additive exPlanations (SHAP) graph showing the important features considered by the new ANN model with less features
Source: The Authors

Regarding the F1 score obtained in Tables 4 to 6, a decrease of approximately 1% is obtained, but considering the ROC curve, the result obtained with the new model that considers fewer features is still better by approximately 5%, as shown in Tables VII to IX. Fig. 4 shows the ROC curve of the model using fewer features. On the other hand, Fig. 5 shows the SHAP values of the new model, noting that the first two most important features taken by the model do not change in their order or magnitude, although in general the importance of the features was maintained even when some were removed from the initial model.

4   Discussion

Health monitoring using artificial intelligence techniques is a very active field. Currently, specifically in the disease-causing current pandemic situation, a wide range of approaches have been used for monitoring and evaluation of patients with COVID-19. Promising solutions have been proposed in screening using clinical symptoms as a preliminary step. The study shows the use of preliminary symptomatology and clinical condition of the patient to detect possible COVID-19 using a machine learning algorithm and information from 5,434 people with and without the disease.

It is shown that feature reduction using techniques such as SHAP values can produce simpler models that use only relevant feature sets to solve a problem. In the case presented, comparison of the model with 11 features, with an F1 score of 0.98 +/-0.08% and an aucROC of 0.99, versus the model with 6 features, with an F1 score of 0.96 +/-0.34% and an aucROC of 0.96, yields a difference of 2 and 3% for each performance index. Obtaining a very small performance variance with a feature reduction of 63%.

The authors, based on the source of the data, consider that the study is not free of errors and biases, since the clinical condition understood as asthma, chronic lung disease, heart disease, diabetes and hypotension; and six initial clinical symptoms: Respiratory problem, fever, dry cough, sore throat, runny nose and headache of the 5,434 people were taken from public data from the Kaggle platform for experimentation with machine learning models. But it is useful as a basis to evaluate the performance of the machine learning model using different configuration parameters and to learn about the most relevant features considered by the model, showing a path for future studies using proprietary databases acquired for research purposes.

5   Conclusion and future research

In this study, we used data from public sources, as an experimental stage. We evaluate different loss functions and configuration parameters of an ANN, to obtain an optimal model that can detect the disease and know which are the most relevant features influencing the detection.

From the data used, a model has been developed to predict the diagnosis of COVID-19 with an F1 score of 0.98 and 0.99 aucROC, using eleven basic features. As a final model using the SHAP values a model using only 6 features is obtained achieving an F1 score performance of 0.96 and 0.96 aucROC, observing a very small percentage difference and retaining the relevant features for the model.

The model is intended to benefit the response of health systems to this disease and other respiratory viruses, although the need for more robust data to complement the study and avoid possible biases is emphasized, before algorithm is employed.

References

[1]   Peña-Reyes, C. A. and Sipper, M., Evolutionary Computation in medicine: an overview, Artif. Intell. Med., 19(1), pp. 1-23, 2000, DOI: https://doi.org/10.1016/S0933-3657(99)00047-0.

[2]   Tan, K.C., Yu, Q.C.. Heng, M., and Lee, T.H., Evolutionary computing for knowledge discovery in medical diagnosis, Artif. Intell. Med., 27(2), pp. 129-154, 2003, DOI: https://doi.org/10.1016/S0933-3657(03)00002-2.

[3]   Li, Z., Chen, W., Wang, J. and Liu, J., An automatic recognition system for patients with movement disorders based on wearable sensors, in:

Proc. 9th IEEE Conf. Ind. Electron. Appl. ICIEA 2014, pp. 1948-1953, 2014. DOI: https://doi.org/10.1109/ICIEA.2014.6931487.

[4] Andrikopoulou, M. et al., Symptoms and critical illness among obstetric patients with coronavirus disease 2019 (COVID-19) infection, Obstet. Gynecol., 136(2), pp. 291-299, 2020. DOI: https://doi.org/10.1097/AOG.0000000000003996.

[5] Amenta, E.M., Spallone, A., Rodriguez-Barradas, M.C., El--Sahly, H.M., Atmar, R.L., and Kulkarni, P.A., Postacute COVID-19: an overview and approach to classification, Open Forum Infect. Dis., 7(12), pp. 1-7, 2020. DOI: https://doi.org/10.1093/ofid/ofaa509.

[6] Maghdid, H.S., Ghafoor, K.Z., Sadiq, A.S., Curran, K., Rawat, D.B., and Rabie, K., A novel AI-enabled framework to diagnose coronavirus COVID-19 using smartphone embedded sensors: design study, arXiv, pp. 1-7, 2020, DOI: https://doi.org/10.48550/arXiv.2003.07434

[7] Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B., and Cheng, X., Artificial intelligence and machine learning to fight Covid-19, Physiol. Genomics, 52(4), pp. 200-202, 2020. DOI: https://doi.org/10.1152/physiolgenomics.00029.2020.

[8] Zoabi, Y., and Shomron, N., COVID-19 diagnosis prediction by symptoms of tested individuals : a machine learning approach, npj Digital Medicine, May, art. 93948, 2020. DOI: https://doi.org/10.1101/2020.05.07.20093948.

[9] Alafif, T. and Bajaba, S., Machine and deep learning towards COVID-19 diagnosis and treatment: survey, Challenges, November, art. 47848, 2020, DOI: https://doi.org/10.13140/RG.2.2.20805.47848/1.

[10] Zoabi, Y., Deri-Rozov, S. and Shomron, N., Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4(1), 2021. DOI: https://doi.org/10.1038/s41746-020-00372-6.

[11] Chen, Y. et al., An interpretable machine learning framework for accurate severe vs non-severe COVID-19 clinical type classification, medRxiv, 2020. DOI: https://doi.org/10.1101/2020.05.18.20105841.

[12] Ahamad, M.M. et al., A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients, Expert Syst. Appl., 160, art. 113661, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.113661.

[13] Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R., Rouf, N., and Mohi Ud Din, M., Machine learning based approaches for detecting COVID-19 using clinical text data, Int. J. Inf. Technol., 12(3), pp. 731-739, 2020. DOI: https://doi.org/10.1007/s41870-020-00495-9.

[14] Smarr, B.L. et al., Feasibility of continuous fever monitoring using wearable devices, Sci. Rep., 10(1), art. 21640, 2020. DOI: https://doi.org/10.1038/s41598-020-78355-6.

[15] Usha-Ruby, A., Theerthagiri, P., Jeena-Jacob, I., and Vamsidhar, Y., Binary cross entropy with deep learning technique for image classification, Int. J. Adv. Trends Comput. Sci. Eng., 9(4), pp. 5393-5397, 2020. DOI: https://doi.org/10.30534/ijatcse/2020/175942020.

[16] Valencia, A.M., Construcción de la distribución de pérdidas y el problema de agregación de riesgo operativo bajo modelos LDA: una revisión, Revista Ingenierías Universidad de Medellín, 12(23), pp. 71-82, 2013.

[17] Wang, Z. and Bovik, A.C., Mean squared error: Love it or leave it?. A new look at signal fidelity measures, IEEE Signal Process. Mag., 6(1), pp. 98-117, 2009, DOI: https://doi.org/10.1109/MSP.2008.930649.

[18] Meyer, G.P., An alternative probabilistic interpretation of the huber loss, arXiv:1911.02088v3, Section 2, pp. 5261-5269, 2019, DOI: https://doi.org/10.48550/arXiv.1911.02088

[19] Lundberg, S. and Lee, S.-I., A Unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 2017, pp. 4766-4775, 2017.

[20] Mangalathu, S., Hwang, S.H. and Jeo, J.S., Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach, Eng. Struct., 219, art. 110927, 2020. DOI: https://doi.org/10.1016/j.engstruct.2020.110927.

[21] Štrumbelj, E. and Kononenko, I., Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst., 41(3), pp. 647-665, 2014. DOI: https://doi.org/10.1007/S10115-013-0679-X.

J.A. Ramirez-Bautista, is BSc. Eng. in Electronic Engineer from the Fundación Universitaria de San Gil (UNISANGIL), San Gil, Colombia, in 2013. The MSc. in advanced technology, and the PhD. in advanced technology, with a specialty in mechatronics, from the Research Center for Applied Science and Advanced Technology (CICATA), Instituto Politécnico Nacional, Queretaro, Mexico, in 2016 and 2020. He is working on both the design and programming of clinical decision support systems using deep neural networks and fuzzy Systems. His research interests include fuzzy systems, hybrid systems, interface development, neural networks, and clinical decision support systems. He is currently a full-time professor at the Faculty of Natural Sciences and Engineering of UNISANGIL, Colombia.
ORCID: 0000-0002-6472-5751

S.L. Chaparro-Cárdenas, is BSc. Eng. in Electronic Engineer from the Fundación Universitaria de San Gil (UNISANGIL), San Gil, Colombia, in 2013. The MSc. in advanced technology, and the PhD. in advanced technology, with a specialty in mechatronics, from the Research Center for Applied Science and Advanced Technology (CICATA), Instituto Politécnico Nacional, Queretaro, Mexico, in 2016 and 2021. She was recognized nationally by the Colombian Association of Engineers (ACIEM), node Santander, with the best graduation project 2013–2014. She is currently a professor and researcher in UNISANGIL, Colombia. Her research interests include fuzzy systems, hybrid systems, robotic rehabilitation devices, neural networks, intelligent control and electrophysiology.
ORCID: 0000-0002-2589-259X

W. Gamboa-Contreras, is BSc. Eng. in Electronic Engineer from the Universidad Industrial de Santander (2002), Sp. in Senior Management from the Universidad Industrial de Santander-UIS, Colombia, (2008), MSc., Technology and Innovation Management from the Universidad de Santander-UDES, Colombia, (2020). 18 years of experience as a university teacher and researcher in agro-industrial science and technology and bioengineering. Inventor of 5 patents granted and one under examination, 2 software registrations, two registered trademarks and two pilot plants. National and international recognitions: National INNOVATE 2020 award from ECOPETROL and UNIRED National Engineering Award (ACOFI 2010), National Innovation Award (Seguros la Equidad).
ORCID: 0000-0001-5526-3156

W. Guerrero-Salazar, is BSc. Eng. in Agricultural Engineer and Business Administrator from the University Foundation of San Gil – UNISANGIL, Colombia (1998-2003), Sp. in Environmental Chemistry from the Industrial University of Santander – UIS, Colombia (2015). He has 16 years of experience as a teacher and university researcher in science and technology, as well as experience in entrepreneurship. Inventor of 1 patent. National awards: ACOFI National Award 2009-2011.
ORCID: 0000-0002-2441-5441

J.A. Huerta-Ruelas, received the MSc. in solid state physics and the PhD. in electrical engineering from the Autonomous University of San Luis Potosi, San Luis Potosi, Mexico, in 1995 and 2000, respectively. He held a Postdoctoral Fellowship in the Department of Science and Food Technology, Oregon State University, Corvallis, OR, USA, in 2004. He is a professor with the Advanced Technology Graduate Program, teaching: optical characterization techniques, interaction of radiation with matter, and the writing and publishing of technical and scientific documents. From 2010-2013, he was the director of Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional, Querétaro México. He is currently a member of the National System of Researchers. His current research focuses on the development of optical measuring systems for use in research and industrial process control.
ORCID: 0000-0001-5632-3368