Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Detection and risk assessment of COVID-19 through machine learning

B. Luna-Benoso, J. C. Martínez-Perales *, J. Cortés-Galicia, U. S. Morales-Rodríguez

Escuela Superior de Cómputo, Instituto Politécnico Nacional, Mexico City, Mexico

ARTICLE INFO

Article history:
Received 28 July 2023
Received in revised form
5 January 2024
Accepted 15 January 2024

Keywords:
Machine learning
Artificial neural networks
Decision trees
Random forests
COVID-19

ABSTRACT

COVID-19, also known as coronavirus disease, is caused by the SARS-CoV-2 virus. People infected with COVID-19 may show a range of symptoms from mild to severe, including fever, cough, difficulty breathing, tiredness, and nasal congestion, among others. The goal of this study is to use machine learning to identify if a person has COVID-19 based on their symptoms and to predict how severe their illness might become. This could lead to outcomes like needing a ventilator or being admitted to an Intensive Care Unit. The methods used in this research include Artificial Neural Networks (specifically, Multi-Layer Perceptrons), Classification and Regression Trees, and Random Forests. Data from the National Epidemiological Surveillance System of Mexico City was analyzed. The findings indicate that the Multi-Layer Perceptron model was the most accurate, with an 87.68% success rate. It was best at correctly identifying COVID-19 cases. Random Forests were more effective at predicting severe cases and those requiring Intensive Care Unit admission, while Classification and Regression Trees were more accurate in identifying patients who needed to be put on a ventilator.

1. Introduction

In late 2019, in Hubei province, China, there were reports of many patients in hospitals suffering from pneumonia and respiratory failure due to a newly identified coronavirus named SARS-CoV-2. The World Health Organization later called this disease COVID-19 and declared it a pandemic in March 2020 (Alhadi et al., 2023). Although there have been different variants of COVID-19, such as Alpha, Beta, Delta, Gamma, Epsilon, Zeta, Eta, and omicron, some common symptoms presented by patients include headache, nasal congestion, fever, and fatigue, among others. However, they can also present numerous complications, including respiratory and pulmonary symptoms, acute respiratory distress syndrome (ARDS), and low oxygen saturation (Firouzabadi et al., 2023; Van Kessel et al., 2022; Ballering et al., 2022), which can lead to the patient's death. Just in December 2020, 79.2 million confirmed cases and more than 1.7 million deaths had been reported worldwide (WHO, 2020). By December 2021, the statistics show a total of 273 million reported cases and 5.3 million deaths (WHO, 2021), while by December 2022, there were 649 million confirmed cases and more than 6.6 million deaths worldwide (WHO, 2022). On the other hand, by February 2023, 762 million confirmed cases and 6.8 million deaths had been reported worldwide (WHO, 2023). The statistics show a greater growth of reported cases and deaths during the first years compared to more recent years; however, the number of registered cases continues to increase. Although the rapid antigen test and RT-PCR test are fast ways to diagnose COVID-19 (Ferté et al., 2021), methodologies have been developed in the field of computer science for the timely detection of COVID-19 and its severity.

There are works that analyze the severity of new COVID-19 variants compared to others (Nyber et al., 2023) but do not carry out a prediction of severity. They only carry out a study of reported cases in their databases; other works compare different methods but only for the detection of COVID-19 through symptoms (Rufino et al., 2023). On the other hand, there are works that evaluate the severity of patients with COVID-19 and their possibilities of entering the Intensive Care Unit (ICU) (Boussen et al., 2022), but they do so using respiratory rate and oxygen saturation signals, whereas other works use clinical data but only limit themselves to the prediction of intubated cases (Arvind et al., 2021). Developing a model that can identify COVID-19 infections and predict the severity of the cases, including whether a

* Corresponding Author.
Email Address: jmartinezp@ipn.mx (J. C. Martínez-Perales)
https://doi.org/10.21833/ijaas.2024.01.025
Corresponding author's ORCID profile:
https://orcid.org/0000-0001-9421-5923

patient may require admission to an Intensive Care Unit or mechanical ventilation, based solely on the symptoms reported by the patient, would be highly beneficial. This research aims to apply machine learning techniques to detect COVID-19 and categorize positive cases as either severe or non-severe. We utilize the database from the National Epidemiological Surveillance System (SINAVE) in Mexico City, which tracks potential COVID-19 cases. This database includes demographic details and symptoms of individuals tested for COVID-19, distinguishing between positive and negative results, as well as information on the severity of the cases. The machine learning component involves the use of Artificial Neural Networks, specifically Multi-Layer Perceptrons, as well as Classification and Regression Trees and Random Forests as classification methods. The effectiveness of these methods is evaluated based on accuracy and the performance metrics derived from their confusion matrices. This study offers valuable insights and could significantly contribute to the efforts of researchers working on technological solutions to combat COVID-19.

2. Literature review

Machine learning is a subfield of artificial intelligence that is concerned with developing techniques that allow computers to learn from a set of data. Several works have been developed using machine learning techniques that allow the detection of COVID-19. Some works include the detection of COVID-19 through the analysis of chest X-ray images (Hu et al., 2022; Bakheet and Al-Hamadi, 2021) or lung radiographs (Panthakkan et al., 2021), while other works include images of both chest and lung radiographs (Duong et al., 2023). There are works that carry out COVID-19 detection using genomic data (Thousif et al., 2022), other works use voice signals (Kanti et al., 2021), and others, in addition to voice signals, do so through coughing and breathing (Pahar et al., 2022; Despotovic et al., 2021). For their part, López-Úbeda et al. (2020) highlighted the importance of textual information processing for classification. In this work, in addition to considering chest radiograph image analysis for COVID-19 detection, they also consider textual reports and conclude that these textual reports contain relevant information to determine the probability that a person presents signs of COVID-19, so they propose a text classification system based on the integration of different information sources applied to the detection of COVID-19 based on chest radiograph reports, for which they used SVM as a classifier. Cisterna-García et al. (2022) also highlighted the importance of diagnostic tests based on data to reduce the mortality rate of COVID-19, and although this work uses data records, it only limits itself to predicting the mortality index and risk of hospitalization from demographic data and comorbidities obtained from clinical histories of patients with COVID-19, in this work they used Random Forest and Logistic Regression for machine learning, obtaining an average accuracy of 71-73% for risk of hospitalization. Works such as the one proposed by Feteira-Santos et al. (2022) use sets of records from the SINAVE for monitoring COVID-19 cases and the National e-Death Certificates Information System (SICO) provided by health information systems. However, they only limit themselves to comparing textual information. Abolfotouh et al. (2022) used records of patient characteristics such as comorbidities, laboratory findings, hospitalization, admission to Intensive Care Units (ICU), and in-hospital and overall mortality, with the aim of describing hospitalization rates, ICU admission, and identifying predictors of in-hospital mortality for COVID-19, that is, they identify causes of in-hospital mortality but do not make predictions of either COVID-19 detection or severe cases. For their part, the work proposed by Perez et al. (2021) performed a multivariate logistic regression analysis to determine the effects of age, sex, previous medical condition, and COVID-19 symptoms to determine the probability of positive cases and hospitalizations during the first wave of the pandemic, March-April 2020, but do not make predictions of severe cases, intubated cases, and cases requiring ICU admission. On the other hand, the work proposed by Huyut (2023) is limited to identifying cases of severe and non-severe COVID-19 patients at the time of admission using routine blood values (RBV) and demographic data. Arvind et al. (2021), for their part, used clinical data but only made predictions of intubated cases. At the same time, Rufino et al. (2023) addressed patient symptoms but only for COVID-19 detection. It is observed that machine learning applied to data records plays an important role in the timely detection of COVID-19 as well as in determining cases of severe and non-severe patients.

3. Methodology

3.1. Artificial neural networks

Artificial Neural Networks (ANNs) are computational models inspired by how biological neural networks function. They are widely used in the field of machine learning for classification tasks. ANNs learn from a training set, adjusting their output data until they produce the desired results. An ANN consists of interconnected nodes organized in layers. The nodes, also known as neurons, are grouped into different layers. Various types of ANNs exist based on their topology, including the Single Perceptron and the Multilayer Perceptron (Liu et al., 2023). The Single Perceptron represents a simple artificial neuron, composed of an input layer $\vec{x} = (x_1, x_2, \dots, x_n)$, a weight vector $\vec{w} = (w_1, w_2, \dots, w_n)$, an activation function φ, and an output layer $\vec{y}$, as shown in Fig. 1.

The activation function φ transmits information to the next layer of interconnected neurons until reaching the output layer. Table 1 presents various activation functions and their corresponding formulas.

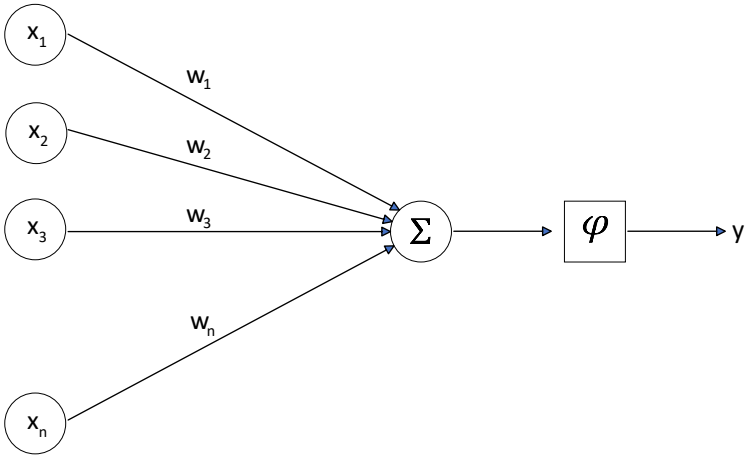

Fig. 1: Simple perceptron with n input neurons and one output neuron

Table 1: Activation functions and formulas

Activation Function	Formula (Equation)
Relu	$\max(0, x)$
Sigmoid	$\frac{1}{1+e^{-x}}$
Softmax	$\frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$, K is the number of classes
Softplus	$\ln(1+e^{x})$
Softsign	$\frac{x}{1+\lvert x\rvert}$
Tanh	$\tanh(x)$
Selu	$\lambda x \; if \; x \geq 0$ $\gamma\alpha(e^{x}-1) if \; x < 0$ with $\alpha = 1.67 \; and \; \lambda = 1.05$
Elu	$x \; if \; x > 0$ $\alpha(e^{x}-1) \; if \leq 0$ with $\alpha = 0.3$

Conversely, a Multilayer Perceptron (MLP) is composed of an input layer, an output layer, and a set of intermediate layers known as hidden layers. Training is carried out using backpropagation.

3.2. Decision trees

A decision tree is a supervised learning model that uses the hierarchical structure of a tree. Each internal node represents a feature or attribute chosen using the maximum gain value, the branches represent decision rules, and the leaf nodes represent the decision outcome (Charbuty and Abdulazeez, 2021). Decisions are made based on a series of questions known as tests, whose sequences of answers transfer the information from the root node to some leaf node. This leaf node contains the information that allows the decision to be made.

3.3. Random forest

It is a supervised learning model in which the dataset is divided into several subsets composed of random samples. Subsequently, an independent decision tree model is used on each subset. Finally, the results of each decision tree are combined using a voting system called majority voting, which consists of giving more weight to those trees with the highest confidence in their result. The result obtained by the random forest model is chosen (Zhang et al., 2023). Fig. 2 shows a typical structure of random forests.

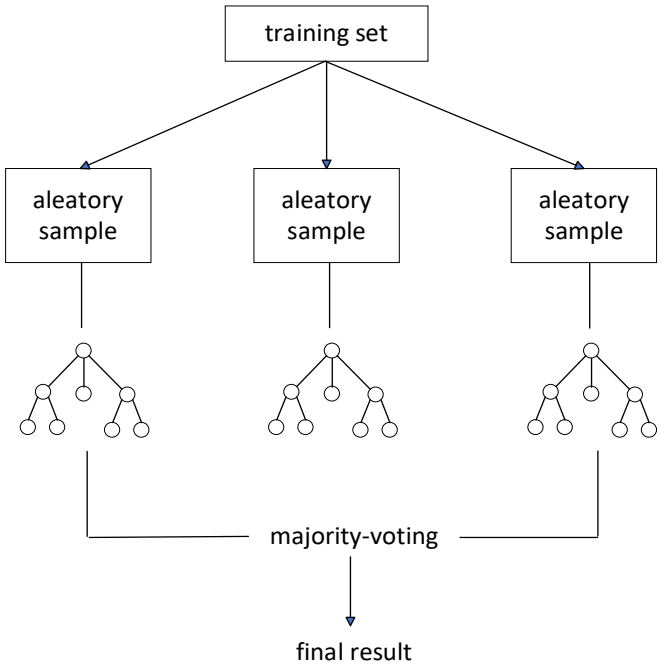

Fig. 2: Random forest model

3.4. Confusion matrix

The purpose of a confusion matrix is to evaluate the performance of a classifier model by describing how real values are distributed with respect to the values produced by the classifier model. The confusion matrix groups in a table the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values that the model produces as results (Fig. 3).

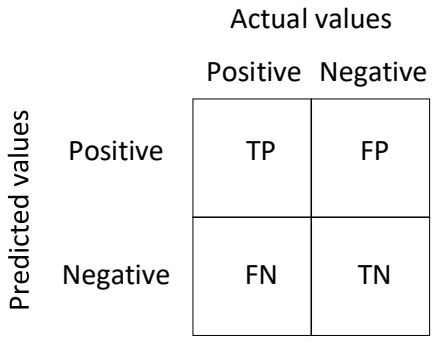

Fig. 3: Confusion matrix

Based on the confusion matrix, metrics such as Sensitivity (SE) and Specificity (SP) can be obtained, which indicate the ability of the classifier model to discriminate between positive and negative cases. The Accuracy (ACC) metric can also be obtained, which indicates the percentage of correct predictions out of the total (Valero-Carreras et al., 2023). The equations that allow to obtain the values of SE, SP, and ACC are given by:

$$SE = \frac{TP}{TP+FN}$$
$$SE = \frac{TP}{TP+FN}$$
$$ACC = \frac{TN+TP}{FN+FP+TN+TP}$$

3.5. COVID-19 data set

The SINAVE of Mexico encompasses a range of epidemiological strategies and activities designed to generate epidemiological information beneficial for public health. This system consolidates information from across the nation and all health system institutions. Specifically, SINAVE maintains a database for tracking potential COVID-19 cases in Mexico City known as SINAVE COVID-19. This database comprises approximately 1 million records with 89 different attributes related to individuals suspected of having COVID-19. It includes demographic details and comorbidities. Attributes cover specific patient information such as gender, age, nationality, occupation, pregnancy status for women and the duration of pregnancy. Additional data include the patient's residential area, symptoms experienced leading to a COVID-19 diagnosis (e.g., diarrhea, headache, muscle pain, vomiting, and difficulty breathing), and conditions like obesity, smoking, and heart disease.

The database also categorizes COVID-19 cases as severe or non-severe based on the patient's condition, including fields like "evolution," "intubated," or "in ICU" for severe cases.

The Multi-Layer Perceptron (MLP) model is noted for its strong learning and classification capabilities, especially with non-linear data sets. However, MLP models can be complex to interpret and require tuning of many hyperparameters to achieve desired results. A significant challenge with MLP is its lengthy training time, exacerbated by the continuous updates and growth of the SINAVE dataset. Decision Trees classify data by recursively dividing the training set into as many homogenous groups as possible, using criteria like the Gini index, Entropy, and Log loss for measuring homogeneity. Despite their simplicity, Decision Trees risk overfitting, which can be mitigated by limiting tree growth during training, such as by restricting the maximum depth. Lastly, Random Forests, which utilize multiple Decision Trees, are less prone to overfitting and can handle incomplete data, making them well-suited to the SINAVE dataset. However, their complexity leads to slower training times.

4. Results

In this section, the experiments and results obtained to identify positive COVID-19 cases and determine whether a severe case is expected are presented using ANNs of the MLP type, Decision Trees, and Random Forests.

The SINAVE dataset consists of around 1 million records with 89 variables, encompassing a wide range of information. These variables include data on patients' residences, personal details like gender, age, and nationality, symptoms of COVID-19 such as fever, cough, difficulty breathing, and chest pain, as well as any chronic diseases patients might have alongside COVID-19 symptoms. Additional details recorded are whether patients received antiviral treatment before being admitted to healthcare facilities and if they were admitted to the ICU or required intubation, which are indicators of the severity of their condition. Another key variable is the result of the COVID-19 antigen test for each patient, although some records lack this test result.

To utilize this dataset effectively, an initial step of data preprocessing is necessary to exclude records missing the antigen test result or other critical information needed for the classification models. Consequently, variables not essential to the objectives of this study, such as patients' residency, nationality, migration status, entry date into the country, or whether the patient speaks an indigenous language, were omitted. After refining the dataset, 50 variables were selected that focus on COVID-19 symptoms, chronic health conditions, and factors indicating the necessity for ICU admission or intubation. Out of these, 46 variables are used as inputs for the classifiers, and four binary variables are used to indicate the outcomes, differentiating between positive and negative COVID-19 cases, predicting severe cases, the need for intubation, and ICU admissions.

During the classification phase, the study first focused on the MLP model. A key initial step was to set specific model parameters, known as hyperparameters, including the number of hidden layers, the number of neurons in each layer, the activation function for each layer, the loss function, the optimization function, and the metrics for evaluating the model's performance. All experiments were conducted with 30 training cycles or epochs. The probabilistic binary cross entropy loss function was selected because the MLP model's outputs are binary (yes/no answers), and accordingly, binary accuracy was used as the evaluation metric. An initial set of hyperparameters was proposed and later adjusted based on the results to fine-tune the MLP model. The initial experiments used a sigmoid activation function, the Adam optimization function, and included 30 epochs of training. Validation was performed using cross-validation with five random splits to assess the model's accuracy.

A specific figure, referred to as Fig. 4, illustrates the process of determining the optimal division of the dataset into training and testing portions. The graph plotted the training set size on the horizontal axis against the accuracy percentage on the vertical axis. Based on this analysis, it was decided to allocate 80% of the dataset for training the model and the remaining 20% for testing its performance. This decision was aimed at achieving a balanced approach to model training and validation, ensuring the model is both well-trained and accurately evaluated on unseen data.

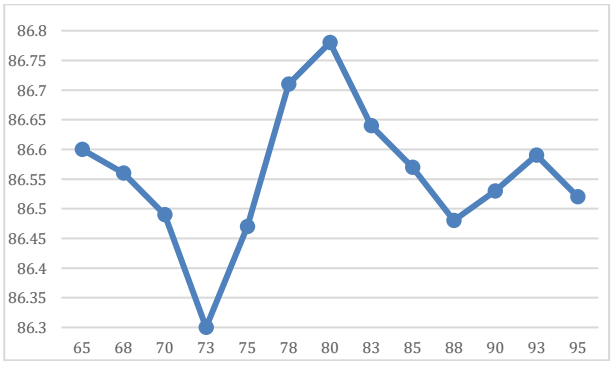

Fig. 4: Percentage of the training set and accuracy of the MLP model

The second stage of the experiments consisted of determining the number of hidden layers that the MLP model must have, as well as the number of neurons per layer. Fig. 5 shows the results of the experiments carried out using one, two, and three hidden layers. The x-axis corresponds to the number of neurons per layer, and the y-axis to the accuracy obtained. In Fig. 5a, the results are shown using one hidden layer, where the best results were obtained, with 79 neurons reaching an accuracy of 87.51%. Using a first hidden layer with 79 neurons, in Fig. 5b, the results are shown to determine with how many neurons the best results are obtained in a second hidden layer. The best results were obtained with 27 neurons, which obtained an accuracy of 87.64%. Using a first hidden layer with 79 neurons and a second hidden layer with 27 neurons, Fig. 5c shows the results obtained when using a third hidden layer. It is observed that the best results were obtained with 30 neurons, with an accuracy of 87.67%.

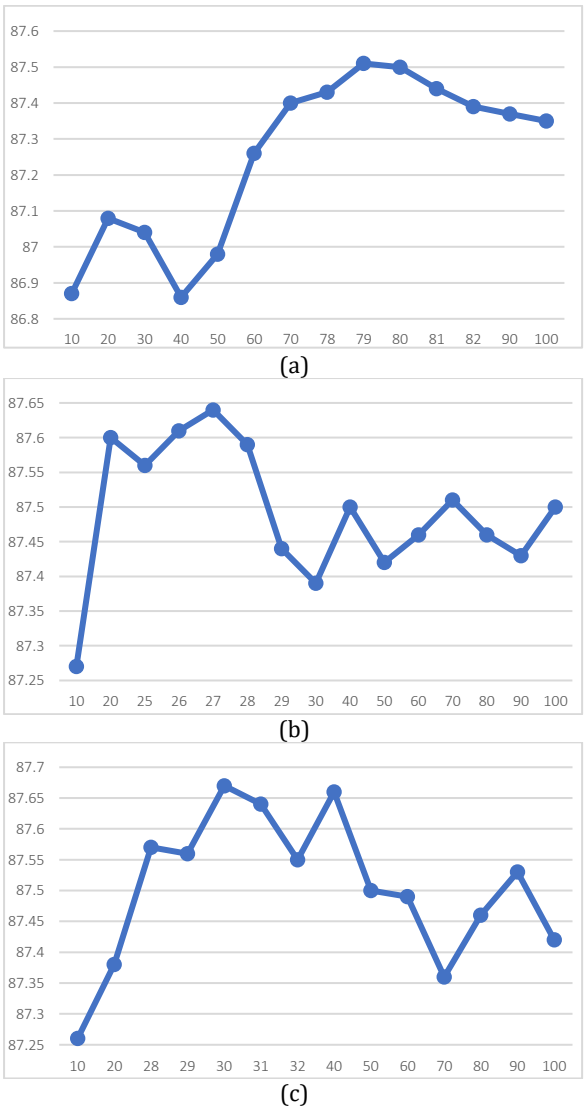

Fig. 5: Results obtained by using a hidden layer in a, a second hidden layer in b, and a third hidden layer in c

The activation function was determined for each hidden layer. Fig. 6 shows the results obtained when applying different activation functions in Fig. 6a to the first hidden layer, in Fig. 6b to the second hidden layer, in Fig. 6c to the third hidden layer, and in Fig. 6d to the output layer. The best results were obtained when applying Softsign, Sigmoid, Softsign, and Sigmoid activation functions to the first, second, and third hidden layer and output layer, respectively.

Fig. 7 shows the result of applying different activation functions to the MLP model, where it was determined to use an Adam activation function.

As a result of the experiments obtained from the different hyperparameters applied to MLP, the best results were obtained by using 80% as the training set and 20% as the test set; configuring the ANN with three hidden layers, the first with 79 neurons, the second with 27 neurons and the third with 30 neurons; the hidden layers use Softsign, Sigmoid and

Softsign activation functions respectively and the output layer uses a Sigmoid activation function; finally, the chosen optimization function was Adam. The accuracy value obtained with these hyperparameters was 87.68%.

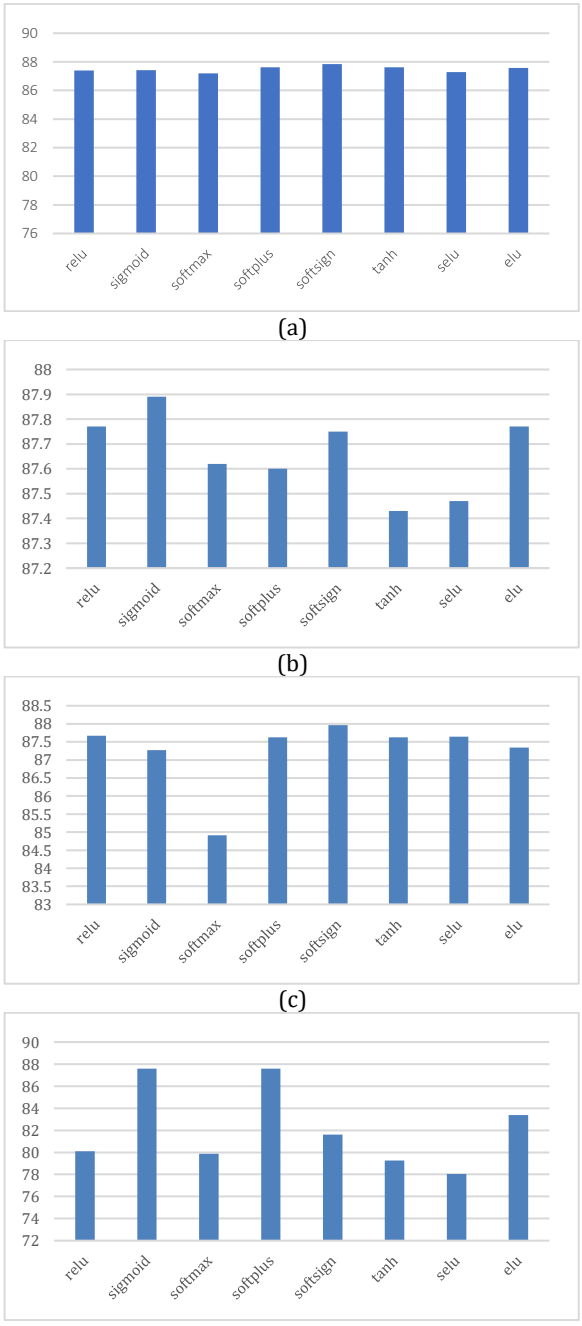

Fig. 6: Different activation functions applied to the first hidden layer in a, the second hidden layer in b, the third hidden layer in c, and the output layer in d

MLP consists of 4 binary outputs. Fig. 8 shows the confusion matrix of each of the output variables and their respective Sensitivity, Specificity, and Accuracy values. Fig. 8a corresponds to positive or negative COVID-19 cases, Fig. 8b corresponds to the prediction of severe cases, Fig. 8c corresponds to those cases that, due to their severity, required intubation, and Fig. 8d to those cases that required admission to the Intensive Care Unit. Nonetheless, a Classification and Regression Tree (CART) was used for the decision tree model. Fig. 9 allows us to determine the percentage of the training set that should obtain the best results with the CART model. The best results are obtained using 85% of the dataset for the training set and 15% for the test set. The splitting criterion is a technique that allows one to decide how a tree should branch. Fig. 10 shows the result of applying the Gini, Entropy, and Log_loss splitting criteria, from which the Entropy criterion shows the best results.

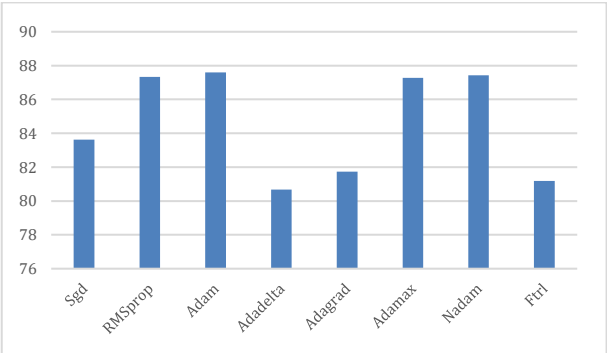

Fig. 7: Different optimization functions applied to MLP

Here are all the hyperparameters that make up the proposed CART model for this work:

- 85% of the dataset corresponds to the training set and 15% to the test set.
- Entropy was used as the division criterion.
- A "Best" division strategy was used.
- The tree has a maximum depth of 6 levels.
- A minimum of 2 examples was considered for separating an internal node.
- A minimum of 5 examples was considered for creating a leaf node.
- 42 characteristics were considered to find the best division in a node.
- The tree comprised a maximum of 180 leaves.
- It does not have a leaf weight fraction, minimum decreasing impurity, class weights, or alpha complexity hyperparameter.

The CART model with the considered hyperparameters showed an accuracy of 86.90% using 5-fold cross-validation.

Fig. 11 shows the confusion matrix of each of the output variables and their respective Sensitivity, Specificity, and Accuracy values using the CART model.

Random forest is a model used in the field of machine learning. The model combines the output of multiple decision trees to obtain a more robust model than the result obtained by a single decision tree. For this work, the tree forests created contain CART trees with the same hyperparameters already presented. Fig. 12 shows tests performed with different numbers of trees and their respective accuracy value. It is observed that the best result was obtained with a forest composed of 80 trees, obtaining an accuracy of 87.40%.

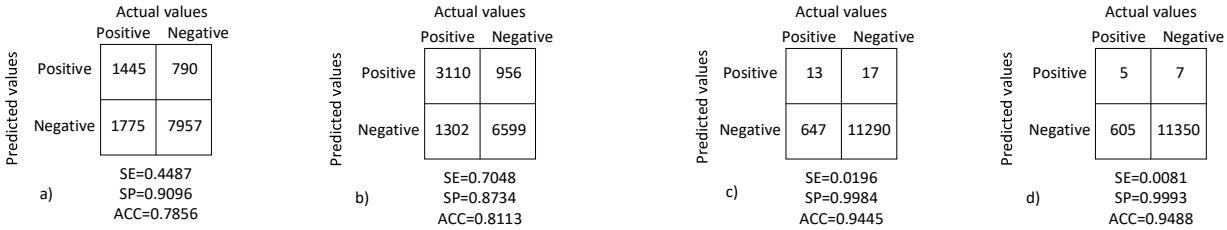

Fig. 8: Confusion matrix of each output variable in the MLP model

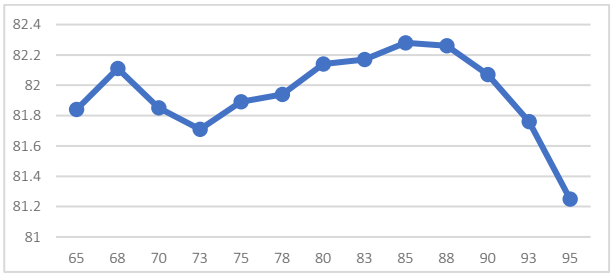

Fig. 9: The percentage of the training set and the accuracy of the CART model

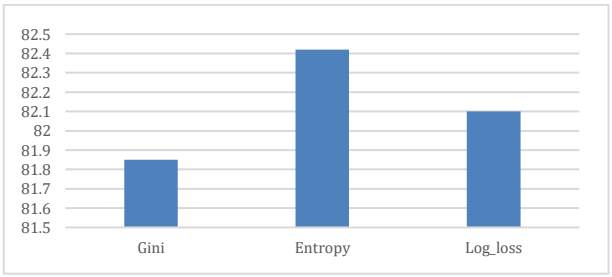

Fig. 10: Division criteria applied to the CART model

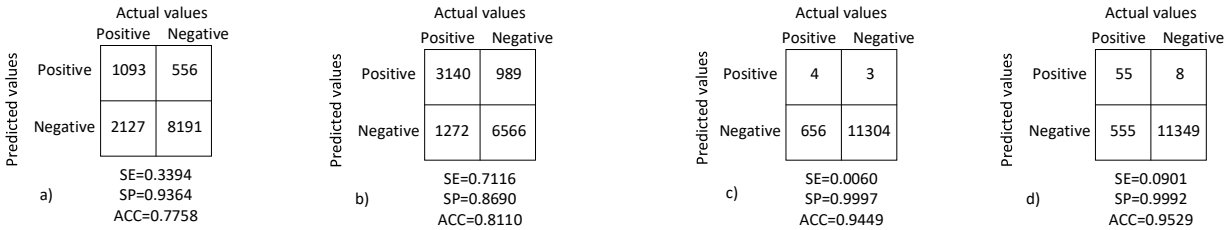

Fig. 11: Confusion matrix of each of the output variables of the CART model

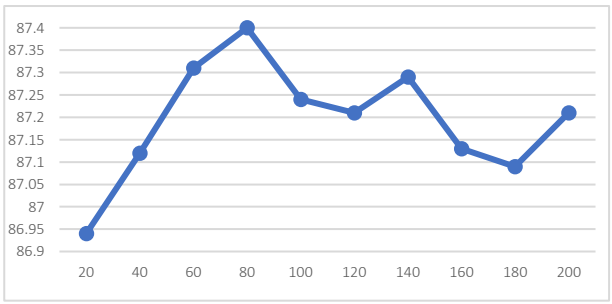

Fig. 12: Random forest with different numbers of trees and their respective accuracy obtained

The Random Forest model used for this work is composed of the following features:

- Each tree in the forest is composed of CART trees with the features exposed in the CART model for solving this problem.
- Random forest is composed of 80 decision trees inside.
- The training data for creating each tree is different.
- It does not estimate a generalization score.
- The decision trees created in the forest start from a tree created previously, to which more estimates are adjusted and added.

The Random Forest model produced an accuracy of 87.45%. Fig. 13 shows the confusion matrix for each output variable and their respective values of Sensitivity, Specificity, and Accuracy using the Random Forest model.

5. Discussion

In this study, three classification models—MLP, CART, and Random Forest—were used to identify COVID-19 from patient symptoms. Additionally, these models assessed whether identified cases were severe, with potential outcomes including the need for intubation or ICU admission. The dataset, provided by Mexico City's SINAVE, was analyzed using 5-fold cross-validation to evaluate the models' accuracy, which measures the correct predictions' proportion against the test dataset.

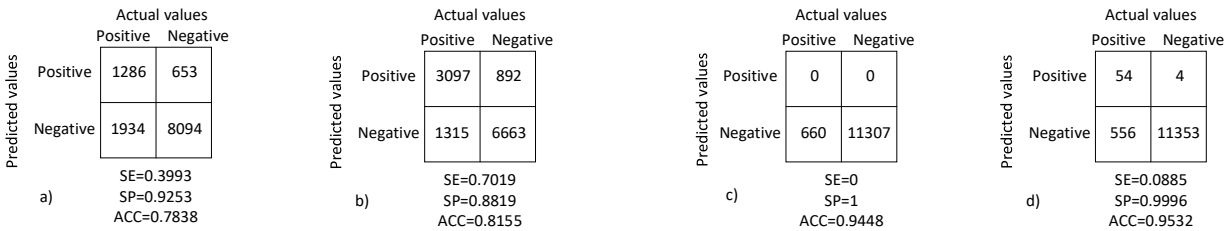

Fig. 13: Confusion matrix of each of the output variables of the Random Forest model

The MLP model demonstrated the highest accuracy at 87.68%, followed closely by Random Forest at 87.45% and CART at 86.90%. The evaluation also considered Sensitivity (the ability to identify positive cases), Specificity (the ability to identify negative cases), and overall Accuracy (the proportion of all correct predictions) from each model's confusion matrix across four scenarios: detecting positive COVID-19 cases, predicting severity, identifying cases requiring intubation, and those necessitating ICU admission.

For identifying positive COVID-19 cases, MLP showed superior accuracy and sensitivity, meaning it was more effective in correctly identifying both positive cases and those actually sick compared to CART and Random Forest. However, CART excelled in specificity, indicating a higher accuracy in identifying patients without COVID-19.

When predicting severe cases, Random Forest led in accuracy and specificity, suggesting it was better at correctly identifying severe cases and more accurately classifying non-severe cases as such. Meanwhile, CART showed the highest sensitivity, indicating a stronger ability to detect severe cases.

In predicting cases requiring intubation, CART had the highest accuracy, MLP showed the greatest sensitivity, and Random Forest achieved perfect specificity. This implies that CART was most accurate in predicting intubated patients, MLP was better at identifying patients who would be intubated, and Random Forest excelled in correctly identifying patients who would not need intubation.

Finally, for predictions concerning ICU admission, Random Forest had the highest accuracy and specificity, whereas CART had the highest sensitivity. This suggests that Random Forest was most accurate in predicting ICU admissions and distinguishing patients who did not require ICU care, while CART was better at identifying patients who were admitted to the ICU. In comparison with other works that do something similar to the work exposed, Cisterna-García et al. (2022) showed a hospitalization risk accuracy result of 0.75, while in this proposed work an accuracy of 0.8155 (Random Forest) was obtained regarding the prediction of patient severity, which is an indicator of a patient requiring hospitalization, in addition to the best accuracy values for intubated cases and ICU admission cases being 0.9449 (CART) and 0.9532 (Random Forest), data that directly correspond to patients requiring hospitalization. On the other hand, Arvind et al. (2021) carried out the prediction of intubated COVID-19 cases and report an accuracy of 0.84 with the model proposed by the authors, however, in this work an accuracy of 0.9449 was obtained with the CART model.

On the other hand, when using KNN for detecting severe cases, Huyut (2023) reported an accuracy of 0.8, like the value obtained in this work of 0.8155 (Random Forest). Although better results are achieved with the work exposed than with other works, however, one of the limitations is that 3 different models are being used separately to achieve it, it would be interesting to generate a model that could obtain the best results that these three models give separately.

6. Conclusion and future work

This work proposed to carry out the detection of COVID-19, as well as the prediction of severity, prediction of being intubated, and prediction of being admitted to the Intensive Care Unit of a COVID-19 patient through symptoms presented, for this Artificial Neural Networks of Multilayer Perceptron, Decision Trees and Random Forests were used. The data used were provided by the National System of Epidemiological Surveillance (SINAVE) of Mexico City. The results were compared via the accuracy that the models yielded using 5-fold cross-validation and the sensitivity, specificity, and accuracy metrics obtained through the confusion matrix. Some works shown in state-of-the-art detect COVID-19 in patients. However, they do so through analysis of chest X-ray images (Bakheet and Al-Hamadi, 2021; Duong et al., 2023; Hu et al., 2022; Panthakkan et al., 2021), on the other hand, other works, like the one proposed, use data records, however, they only limit themselves to one of the proposed problems such as prediction of intubated cases (Arvind et al., 2021), identification of severe cases (Huyut, 2023), or show the risk of hospitalization of a COVID-19 patient (Cisterna-García et al., 2022), which is associated with the severity of the patient. The proposed work, for its part, contemplates the four mentioned problems using records of symptoms presented by patients, and three different classifiers were used that together show better accuracy results compared to the works with which it was compared (Cisterna-García et al., 2022; Arvind et al., 2021; Huyut, 2023). With MLP, the best accuracy of 87.68% was obtained. However, in terms of the accuracy yielded by the confusion matrix regarding the prediction of severe cases, Random Forest obtained the best result with 0.8155. For the prediction of intubated cases, CART obtained the best result with 0.9449. Finally, for those cases requiring admission to the Intensive Care Unit, Random Forest obtained the best result with 0.9532. It is observed that to have a work that considers detecting positive COVID-19 cases and their severity requires more than one classifier to obtain favorable results. Consequently, researchers could perform experiments with different databases of symptom-related COVID-19 records that allow detection and consider a greater number of models to determine which ones yield the best results in each of the predictions made.

Acknowledgment

The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, EDD, EDI, SIP, and ESCOM) and CONAHCYT for their financial support in developing this work.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abolfotouh MA, Musattat A, Alanazi M, Alghnam S, and Bosaeed M (2022). Clinical characteristics and outcome of COVID-19 illness and predictors of in-hospital mortality in Saudi Arabia. BMC Infectious Diseases, 22: 950.
https://doi.org/10.1186/s12879-022-07945-8
PMid:36526994 PMCid:PMC9758036

Alhadi B, Khder MM, Rashid S, Taha K, and Manzour AF (2023). Health care workers' perceptions of their hospitals' preparedness during the COVID-19 virus pandemic in three different world regions. Clinical Epidemiology and Global Health, 21: 101278.
https://doi.org/10.1016/j.cegh.2023.101278
PMid:37033720 PMCid:PMC10066860

Arvind V, Kim JS, Cho BH, Geng E, and Cho SK (2021). Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. Journal of Critical Care, 62: 25-30.
https://doi.org/10.1016/j.jcrc.2020.10.033
PMid:33238219 PMCid:PMC7669246

Bakheet S and Al-Hamadi A (2021). Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification. Computers in Biology and Medicine, 137: 104781.
https://doi.org/10.1016/j.compbiomed.2021.104781
PMid:34455303 PMCid:PMC8382592

Ballering AV, van Zon SKR, Olde TC, and Rosmalen JGM (2022). Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study. Lancet, 400(10350): 452-461.
https://doi.org/10.1016/S0140-6736(22)01214-4
PMid:35934007

Boussen S, Cordier PY, Malet A, Simeone P, Cataldi S, Vaisse C, Roche X, Castelli A, Assal M, Pepin G, Cot K, Denis JB, Morales T, Velly L, and Bruder N (2022). Triage and monitoring of COVID-19 patients in intensive care using unsupervised machine learning. Computers in Biology and Medicine, 142: 105192.
https://doi.org/10.1016/j.compbiomed.2021.105192
PMid:34998220 PMCid:PMC8719000

Charbuty B and Abdulazeez A (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01): 20-28.
https://doi.org/10.38094/jastt20165

Cisterna-García A, Guillén-Teruel A, Caracena M, Pérez E, Jiménez F, Francisco-Verdú FJ, Reina G, González-Billalabeitia E, Palma J, Sánchez-Ferrer A, and Botía JA (2022). A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study. Scientific Reports, 12: 18126.
https://doi.org/10.1038/s41598-022-22547-9
PMid:36307436 PMCid:PMC9614188

Despotovic V, Ismael M, Cornil M, Mc Call R, and Fagherazzi G (2021). Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. Computers in Biology and Medicine, 138: 104944.
https://doi.org/10.1016/j.compbiomed.2021.104944
PMid:34656870 PMCid:PMC8513517

Duong LT, Nguyen PT, Iovino L, and Flammini M (2023). Automatic detection of COVID-19 from chest X-ray and lung computed tomography images using deep neural networks and transfer learning. Applied Soft Computing, 132: 109851.
https://doi.org/10.1016/j.asoc.2022.109851
PMid:36447954 PMCid:PMC9686054

Ferté T, Ramel V, Cazanave C, Lafon ME, Bébéar C, Malvy D, Georges-Walryck A, and Dehail P (2021). Accuracy of COVID-19 rapid antigenic tests compared to RT-PCR in a student population: The StudyCov study. Journal of Clinical Virology, 141: 104878.
https://doi.org/10.1016/j.jcv.2021.104878
PMid:34134035 PMCid:PMC8178956

Feteira-Santos R, Camarinha C, Nobre MA, Elias C, Bacelar-Nicolau L, Silva A, Furtado C, and Nogueira PJ (2022). Improving morbidity information in Portugal: Evidence from data linkage of COVID-19 cases surveillance and mortality systems. International Journal of Medical Informatics, 163: 104763.
https://doi.org/10.1016/j.ijmedinf.2022.104763
PMid:35461149 PMCid:PMC9012514

Firouzabadi N, Ghasemiyeh P, Moradishooli F, and Mohammadi S (2023). Update on the effectiveness of COVID-19 vaccines on different variants of SARS-CoV-2. International Immunopharmacology, 117: 109968.
https://doi.org/10.1016/j.intimp.2023.109968
PMid:37012880 PMCid:PMC9977625

Hu Q, Nauber F, Costa R, Zhang L, Yin L, Magaia N, and Albuquerque VHC (2022). Explainable artificial intelligence-based Edge fuzzy images for COVID-19 detection and identification. Applied Soft Computing, 123: 108966.
https://doi.org/10.1016/j.asoc.2022.108966
PMid:35582662 PMCid:PMC9102011

Huyut MT (2023). Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models. Innovation and Research in BioMedical Engineering, 44(1): 100725.
https://doi.org/10.1016/j.irbm.2022.05.006
PMid:35673548 PMCid:PMC9158375

Kanti T, Mishra S, Panda G, and Chandra S (2021). Detection of COVID-19 from speech signal using bio-inspired based cepstral features. Pattern Recognition, 117: 107999.
https://doi.org/10.1016/j.patcog.2021.107999
PMid:33967346 PMCid:PMC8086594

Liu W, Zhang L, Xie L, Hu T, Li G, Bai S, and Yi Z (2023). Multilayer perceptron neural network with regression and ranking loss for patient-specific quality assurance. Knowledge-Based Systems, 271: 110549.
https://doi.org/10.1016/j.knosys.2023.110549

López-Úbeda P, Díaz-Galiano MC, Martín-Noguerol T, Luna A, Ureña-López LA, and Martín-Valdivia MT (2020). COVID-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine, 127: 104066.
https://doi.org/10.1016/j.compbiomed.2020.104066
PMid:33130435 PMCid:PMC7577869

Nyber T, Bager P, Svalgaard IB, Bejko D, Bundle N, Evans J, Krause TG, McMenamin J, Mosson J, Mutch H, Omokanye A, Peralta-Santos A, Pinto-Leite P, Starrfelt J, Thelwall S, Veneti L, Whittaker R, Wood J, Peboy R, and Presanis AM (2023). A standardised protocol for relative SARS-CoV-2 variant severity assessment, applied to Omicron BA.1 and Delta in six European countries, October 2021 to February 2022. Eurosurveillance, 28(36): 2300048.
https://doi.org/10.2807/1560-7917.ES.2023.28.36.2300048
PMid:37676146 PMCid:PMC10486193

Pahar M, Klopper M, Warren R, and Niesler T (2022). COVID-19 detection in cough, breath and speech using deep transfer learnig and bottleneck features. Computers in Biology and Medicine, 141: 105153.
https://doi.org/10.1016/j.compbiomed.2021.105153
PMid:34954610 PMCid:PMC8679499

Panthakkan A, Anzar SM, Mansoori SA, and Ahmad HA (2021). A novel DeepNet model for the efficient detection of COVID-19 for symptomatic patients. Biomedical Signal Processing and Control, 68: 102812.

https://doi.org/10.1016/j.bspc.2021.102812
PMid:34075316 PMCid:PMC8156912

Perez M, Saad NJ, Lucaccioni H, Costa C, McMahon G, Machado F, Balasegaram S, and Machado RS (2021). Clinical and hospitalisation predictors of COVID-19 in the first month of the pandemic, Portugal. PLOS ONE, 16(11): e0260249. https://doi.org/10.1371/journal.pone.0260249
PMid:34797879 PMCid:PMC8604361

Rufino J, Ramírez JM, Aguilar J, Baquero C, Champati J, Frey D, Lillo RE, and Fernández-Anta A (2023). Consistent comparison of symptom-based methods for COVID-19 infection detection. International Journal of Medical Informatics, 177: 105133. https://doi.org/10.1016/j.ijmedinf.2023.105133
PMid:37393765

Thousif M, Abdul M, and Vankdothu R (2022). COVID-19 detection and classification for machine learning methods using human genomic data. Measurement: Sensors, 24: 100537. https://doi.org/10.1016/j.measen.2022.100537
PMid:36466096 PMCid:PMC9595328

Valero-Carreras D, Alcaraz J, and Landete M (2023). Comparing two SVM models through different metrics base on the confusion matrix. Computers and Operations Research, 152: 106131. https://doi.org/10.1016/j.cor.2022.106131

Van Kessel SAM, Olde HTC, Lucassen PLBJ, and van Jaarsveld CHM (2022). Post-acute and long-COVID-19 symptoms in patients with mild diseases: A systematic review. Family Practice, 39(1): 159-167. https://doi.org/10.1093/fampra/cmab076
PMid:34268556 PMCid:PMC8414057

WHO (2020). Weekly epidemiological update - 29 December 2020. World Health Organization. Geneva, Switzerland.

WHO (2021). Weekly epidemiological update on COVID-19 - 21 December 2021. World Health Organization. Geneva, Switzerland.

WHO (2022). Weekly epidemiological update on COVID-19 - 21 December 2022. World Health Organization. Geneva, Switzerland.

WHO (2023). Weekly epidemiological update on COVID-19 - 13 April 2023. World Health Organization. Geneva, Switzerland.

Zhang H, Quost B, and Masson MH (2023). Cautious weighted random forest. Expert Systems with Applications, 213: 118883. https://doi.org/10.1016/j.eswa.2022.118883