Evaluation of Supervised machine learning classifiers for detecting the degrees of possibility of Coronavirus disease infection

Abbas Rammal (rammal_abbass@hotmail.com)
Lebanese University
Elie Yammine
Lebanese University
Walid Fahs
Islamic University of Lebanon
Rida Khatoun
Telecom ParisTech

Research Article

Keywords: Machine learning, Supervised learning algorithms, Classification, Coronavirus disease, Model prediction, Support Vector Machine, Naïve Bayesian Classifier, Decision Tree Classifier, Linear Discriminant Analysis, Artificial Neural Network, Extended Gamma, Associative Classifier Naive Associative Classifier

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-958969/v1

Evaluation of Supervised machine learning classifiers for detecting the degrees of possibility of Coronavirus disease infection

Abbas Rammal*[1,2,3], Elie Yammine[2], Walid Fahs[1], Rida Khatoun[4]

[1]Islamic University of Lebanon, Faculty of Engineering, 30014, Wardenieh, Lebanon.

[2]Lebanese University, Faculty of Science, Statistics and Informatics department, Beirut, Lebanon.

[3]Lebanese University, Faculty of Information, Data Science department, Beirut, Lebanon.

[4] Telecom ParisTech, Paris, France

rammal_abbass@hotmail.com, e.yammine@st.ul.edu.lb, walid.fahs@iul.edu.lb,

rida.khatoun@telecom-paris.fr

Corresponding author

Correspondence to Abbas Rammal

First author: Abbas Rammal, Ph.D., Assistant professor, Lebanese University, Faculty of Science, Statistics and computer sciences department, Beirut, Lebanon, abbas.rammal@ul.edu.lb

Second author: Elie Yammine, Master's degree in applied statistics, Lebanese University, Faculty of Science, computer science and statistics department, Beirut, Lebanon, e.yammine@st.ul.edu.lb

Third author: Walid Fahs, Phd, Assistant professor, Islamic University of Lebanon, Faculty of Engineering, transportation, and civil engineering department, Wardenieh, Lebanon, walid.fahs@iul.edu.lb

Fourth author: Rida Khatoun, Ph.D., Assistant professor, Telecom ParisTech, Paris, France, rida.khatoun@telecom-paris.fr

1. Abstract

The COVID-19 virus spread from China throughout the world, causing out of control challenges to the public health community. During the epidemic, it is difficult to act against infectious disease due to its unknown trends, so the spread prediction becomes difficult in light of the scarce data. With the absence of treatment for COVID-19 infection, countries have taken some mitigation steps and policies, such as general lockdown and social distance measures, but there has been variation in the extent of viral spread due to several additional factors. Prediction methods based on Artificial intelligence (AI) and Machine Learning (ML) can help in suggesting new policies or even assessing the effectiveness of the existing ones. Such methods have attracted wide attention from researchers implementing statistical modeling and machine learning methods.

The objective of this study is to examine different supervised classification approaches to detect the degrees of possibility of Coronavirus disease infection in different countries. Naïve Bayesian Classifier (NBC), Decision Tree Classifier (DTC), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Artificial Neural Network classification (ANN) are machine learning algorithms used for the prediction of coronavirus disease cases according to the time of their evolution while considering data collected from official reports and scientific journals. Since we collected mixed data, we suggested to also apply the supervised classifiers available for mixed data, such as Extended Gamma (EG) and Naive Associative Classifier (NAC).

The results showed that ANN and DTC supervised algorithms allow better discrimination between the degrees of possibility of Coronavirus disease infection among advanced methods like NBC, SVM, NAC, EG and LDA.

2. **Keywords**

Machine learning, Supervised learning algorithms, Classification, Coronavirus disease, Model prediction, Support Vector Machine, Naïve Bayesian Classifier, Decision Tree Classifier, Linear Discriminant Analysis, Artificial Neural Network, Extended Gamma, Associative Classifier Naive Associative Classifier.

3. Introduction

In December 2019, the novel coronavirus appeared in Wuhan city in China [1] as the first case was reported to the World Health Organization (W.H.O) on 31 December 2019. The virus created a global threat and was announced as coronavirus disease (COVID-19) by the W.H.O on 11 February 2020. The COVID-19 epidemic caused serious global socio-economic turmoil [2]. As of 29 July 2020, more than 17 million cases were recorded. The virus was detected in 215 countries causing more than 665,000 deaths [3].

The early detection of any disease, whether it is infectious or not, is a critical task for rapid treatment to save more lives [4]. Fast diagnosis and screening processes help prevent the spread of pandemic diseases like SARS-CoV-2 [5]. Moreover, predicting the probability degree of getting infected with covid-19 may help national authorities in taking proactive mitigation measures to contain possible epidemic waves of infections. The biggest challenge is to build a technique to improve the detection of covid-19 disease. Several Data Science and Artificial Intelligence techniques have been used to aid the Coronavirus Response [6, 7]. In the literature review, deep learning methods to predict the structure of proteins and their interactions with chemical compounds to facilitate new antiviral drugs/vaccines or recommend current drugs is widely used [8]. A machine learning-based prognostic model to predict if a patient infected with Covid-19 would survive the infection based on age and other risk factors [9]. Data-Based Analysis, Modelling and Forecasting infection rates and spread/patient prognosis to enable hospitals/health officials to better plan resourcing and response [10].

Machine learning technologies are used to improve the accuracy of prediction for screening both infectious and non-infectious diseases [11]. The machine learning methods were used in modeling former pandemics (e.g., Ebola, Cholera, swine fever, H1N1 influenza [12], dengue

fever [13], Zika, oyster norovirus [14]). Table 1 represents notable Machine Learnings methods used for outbreak prediction.

Table 1. Notable Machine Learning techniques for outbreak prediction

References	**Outbreak infection**	**Machine learning methods**	**Resources used to collect data**
Geospatial Health [15]	Dengue fever	Neural Network	The computerized database for 2016 has been obtained from Taiwan Centers for Disease Control (Taiwan CDC, 2016), which contains all national dengue fever records daily from 1998, including age, gender, township of residence and time of disease onset.
Malaysian Journal of Public Health Medicine [16]	Dengue/Aedes	Bayesian Network	The Aedes outbreak database sourced from the Ministry of Health at pre-determined localities around the Klang Valley, Malaysia.
Global Ecology and Biogeography [12]	H1N1 flu	Neural Network	Data on the prefectural distribution of the H1N1 influenza pandemic in 2009 were provided by the Ministry of the Environment of Japan.
Current Science [17] Infectious Disease Modelling [18]	Dengue	Naïve Bayes Classification and regression tree (CART)	The data has been collected on the monthly dengue cases from the Government of the National Capital Territory (NCT) of Delhi (India).
Environment International [19]	Oyster norovirus	Neural Network	The epidemiological data were obtained from various online data sources. Specifically, historical norovirus outbreaks in Louisiana oyster harvest areas were collected from Louisiana Morbidity Reports released by Louisiana Department of Health and Hospitals

Another recent study presents a comparative analysis of machine learning and soft computing models to predict the COVID-19 outbreak as an alternative to Susceptible- Infectious-Removed (SIR) and Susceptible-Exposed-Infectious-Removed (SEIR) models. Among a wide range of machine learning models investigated, two models showed promising results (i.e., multi-layered perceptron, MLP, and adaptive network-based fuzzy inference system, ANFIS) [20]. Also, Machine learning and deep learning can replace humans by giving an accurate diagnosis [21]. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19. X-rays and computed tomography (CT) scans can be used for training the machine learning model. Several initiatives are under-way in this regard. Wang and Wong developed COVID-Net, which is a deep convolutional neural network, which can diagnose COVID-19 from chest radiography images [22]. Moreover, Recent studies [28] show the potential of Artificial Intelligence and Machine Learning tools by suggesting a new model that comes with rapid and valid method SARS-CoV-2 diagnosis using Deep Convolutional Network [23, 24]. In this regard, the remarkable performance suggests the use of the convolutional neural network (Resnet-101) as an adjuvant tool for increase the accuracy of Covid-19 diagnosis. Recent studies used the supervised machine learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS). Logistic regression and Multinomial Naïve Bayes showed better results than other ML algorithms for detecting COVID-19 using clinical text data set [25].

In a comparative analysis framework between various supervised machine learning algorithms in diagnosing COVID-19 infections, Pijush Dutta implemented bagging algorithm, k-nearest neighbor, and random forest for classifying the datasets of COVID-19 [26]. Random Forest gave better results with employing accuracy of 85.71%. Haochen Yao and Nan Zhang built a COVID-19 severeness detection model based on supervised machine learning algorithms [27],

and obtained the best model using the Support Vector Machine algorithm (SVM) with 28 features and overall accuracy of 81.48%. Mahbubunnabi Tamal trained a supervised machine learning algorithms to distinguish between COVID-19 and other diseases, where he found that SVM and Ensemble Bagging Model Trees (EBM) when trained on 71 radiomics features can distinguish between COVID-19 and other diseases with an overall sensitivity of 99.6% and 87.8% and specificity of 85% and 97% respectively [28]. Davide Brinati developed two machine learning classification models using hematochemical values from routine blood exams to discriminate between patients who are either positive or negative to the SARS-CoV-2 [29]: their accuracy ranges between 82% and 86%, and sensitivity between 92% and 95%.

Due to the highly complex nature of the COVID-19 outbreak and variation in its behavior from nation-to-nation, this study suggests applying the machine learning as an effective tool to model the detection and improve the accuracy of prediction for screening coronavirus disease (COVID-19) according to the time of their evolution (in days) in different country. Various supervised machine learning algorithms exist in the literature. The best-known are Linear Discriminant Analysis (LDA) [30], Naïve Bayes Classification (NBC) [31], Decision Tree Classification (DTC) [32, 33], and Support Vector Machine Classification (SVM) [34], Artificial Neural Network (ANN) [35]. Since we collected mixed data, I strongly suggested comparing the classifiers available for mixed data, such as Extended Gamma Associative Classifier (EG) [36], Naive Associative Classifier (NAC) [37]. Irrespective of the method chosen, the use of Cross-validation is a model assessment technique used to evaluate a supervised classification algorithm's performance in making predictions on new COVID-19 datasets that it has not been trained on. This paper aims to investigate the generalization ability of the proposed Machine Learning models and the accuracy of the proposed models for

detecting the degrees of possibility of Coronavirus disease infection from official reports of COIVD-19 data.

4. Mathematical Methods

The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing samples where the values of the predictor feature are known, but the value of the class labels are unknown. The classifier's evaluation is most often based on prediction accuracy (the percentage of correct prediction divided by the total number of predictions) [38]. There are at least three techniques which are used to calculate a classifier's accuracy. One technique is to split the available data by using part of them for training and the other for estimating performance, for example two-thirds for training and the other third for testing. In the second technique, known as cross-validation, the training set is divided into mutually exclusive and equal sized subsets and for each subset the classifier is trained on the union of all the other subsets. The average of the error rate of each subset is therefore an estimate of the error rate of the classifier. The third technique is the leave one out validation, which is a special case of cross validation. All test subsets consist of a single sample. This type of a validation is more expensive computationally, but useful when the most accurate estimate of classifier's error rate is required.

Supervised classification is one of the tasks most frequently carried out by so-called artificial intelligence. Many techniques have been developed based such as Logic-based techniques, Perceptron-based techniques, Bayesian Networks, etc. In this application, we have chosen Decision Trees classifiers as logic learning methods, and Linear Discriminant Analysis, Support Vector Machines and Naïve Bayes Classification as statistical learning algorithms. These techniques are widely used in several research areas, such as biology, medicine fields

and advanced technology [35]. Conversely to artificial neural network, statistical approaches are characterized by having an explicit underlying probability model, which provides a probability that a sample belongs in each class, rather than simply a classification. In addition, the perceptron-based techniques require several parameters which must be studied such that the size of hidden layers and number of neurons. For this, it is difficult to adapt the neural network method as an objective function in the methods of selection of variables such as Genetic Algorithms.

Let $\mathbf{X} = \{\mathbf{x_j}\}_{j=1}^{n}$ be a training set of n samples of observed variables, where each sample is represented by an S-dimensional vector and let c_k denoting the class membership of $\mathbf{x_j}$, where $c_k \in \{c_1, \dots, c_K\}$ with K denoting the number of classes of training set.

4.1.Naive Bayesian classifier

The naïve Bayesian classifier (NBC) is based on Bayes' theorem and predicting independence requirements. A naïve Bayesian model is simple to construct since it does not need costly iterative parameter estimates, resulting in a quick computing time for training [39].

Given $\mathbf{x_j} = [x_{j1}, \dots, x_{js}, \dots, x_{jS}]^T \in \mathbb{R}^S$, selection of observed data retrieved from the training data set X and belonging to the class c_k. The Naïve Bayes classifier can predict that $\mathbf{x_j}$ belongs to the predicted class $\hat{C}$ with the following Maximum a Posteriori Probability (MAP) based on $\mathbf{x_j}$ **:**

$$\text{MAP}(\mathbf{x} = \mathbf{x_j}) = \arg max_{k \in \{1,\dots,K\}} P(C = c_k | \mathbf{x} = \mathbf{x_j}). \quad \text{(eq. 1)}$$

According to Bayes' theorem, the probability $P(C = c_k | \mathbf{x} = \mathbf{x_j})$ that we wish to calculate could be defined in terms of $P(C = c_k)$, $P(\mathbf{x} = \mathbf{x_j} | C = c_k)$ and $P(\mathbf{x} = \mathbf{x_j})$ as

$$P(C = c_k | \mathbf{x} = \mathbf{x_j}) = \frac{P(\mathbf{x}=\mathbf{x_j}|C=c_k)P(C=c_k)}{P(\mathbf{x}=\mathbf{x_j})}, \quad \text{(eq. 2)}$$

where $P(\mathbf{x} = \mathbf{x_j})$ in the denominator may be discarded since it does not rely on C, and the

value of C, and the value of $P(\mathbf{x} = \mathbf{x_j})$ is a known constant. $P(\mathbf{x} = \mathbf{x_j}|C = c_k)$ is called the Class-conditional Probability Distribution (CPD). Thus, [40, 41] gives the computation of the Nave Bayes classifier:

$$\mathrm{MAP}(\mathbf{x} = \mathbf{x_j}) = \arg max_{\mathrm{k}\in\{1,\ldots,K\}} P(\mathbf{x} = \mathbf{x_j}|\mathrm{C} = c_k)P(C = c_k). \quad \text{(eq. 3)}$$

The likelihood of a classification error, often known as the classifier's risk, is defined as [31]:

$$R^{NBC}(\mathbf{x_j}) = \sum_{x_j\in\mathbf{X}} \mathbf{1}_{C(\mathbf{x}=\mathbf{x_j})\neq\hat{C}(\mathbf{x}=\mathbf{x_j})} P(\boldsymbol{X} = \boldsymbol{x}_j) = E_{\boldsymbol{X}}\left\{\mathbf{1}_{C(\mathbf{x}=\mathbf{x_j})\neq\hat{C}(\mathbf{x}=\mathbf{x_j})}\right\}, \quad \text{(eq. 4)}$$

where $\mathbf{1}_{\mathrm{C}(\mathbf{x}=\mathbf{x_j})\neq\widehat{\mathrm{C}}(\mathbf{x}=\mathbf{x_j})}$ is the indicator function and $E_{\boldsymbol{X}}$ is the expectancy over $\mathbf{X}$ and $\mathrm{C}(\mathbf{x} = \mathbf{x_j})$ is the real class which the vector $\mathbf{x_j}$ corresponds and $\hat{C}(\mathbf{x} = \mathbf{x_j})$ is the expected class of $\mathbf{x_j}$ provided by the NBC method. The Bayes error is the rate of misclassification (Bayes risk). When addressing classification issues, the Bayes risk is frequently specified as the reference. It is the least error rate when the distribution is given.

4.2.Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a dimensional reducing and classification approach that has been widely used to the analysis of spectral data. Consider the challenge of allocating a sample of observed variables retrieved from the training set $\mathbf{x_j}$ to the class c_k in a classification task. The categorization score is calculated as follows:

$$cf(\mathbf{x_j}) = (\mathbf{x_j} - \mu_k)^T \Sigma_{\mathrm{k}}^{-1}(\mathbf{x_j} - \mu_k) + \ln|\Sigma_{\mathrm{k}}| - 2\ln[P(C = c_k)] \quad \text{(eq. 5)}$$

$P(C = c_k)$ is the prior probability of class c_k, Σ_{k} is the class covariance matrix of class c_k, and μ_k is the mean vector of class c_k, and they are calculated by

$$\widehat{\mu_k} = \frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{x_j} \quad \text{(eq. 6)}$$

$$\widehat{\Sigma_k} = \frac{1}{n_k}\sum_{j=1}^{n_k}(\mathbf{x_j} - \mu_k)(\mathbf{x_j} - \mu_k)^T \quad \text{(eq. 7)}$$

$$P(\widehat{C = c_k}) = \frac{n_k}{n} \quad \text{(eq. 8)}$$

where n_k represents the number of samples in class c_k, and n represents the total number of

samples in the training dataset. The sample $\mathbf{x_j}$ is classified as belonging to the class with the lowest classification score [32].

When the class covariance matrices are believed to be equal, a pooled covariance matrix is formed.

$$\Sigma_{\text{pooled}} = \frac{1}{n}\sum_{k=1}^{K} n_k \Sigma_k \qquad \text{(eq. 9)}$$

and is substituted for the class covariance matrix in eq. (5). Without taking into account constants, the following classification rule for LDA is produced.

$$cf(\mathbf{x_j}) = \left(\mathbf{x_j} - \mu_k\right)^T \Sigma_{\text{pooled}}^{-1} \left(\mathbf{x_j} - \mu_k\right) - 2\ln[P(C = c_k)] \qquad \text{(eq. 10)}$$

Equation (10) relates to the Mahalanobis distance when the prior probability n_i is constant. The LDA classifier's error rate R^{LDA}, which is defined as:

$$R^{LDA}(\mathbf{x_j}) = \frac{1}{n}\sum_{j=1}^{n} \left| sign(\mathrm{c}_k(\mathbf{x_j}) - \hat{c}_k(\mathbf{x_j})) \right|, \qquad \text{(eq. 11)}$$

with $\boldsymbol{x_j}$ belonging to the class c_k as determined by the LDA method classifier in the class $\hat{c}_k$

4.3.Decision Tree Classifier

A decision tree classifier is a non-parametric classifier that does not require any a priori statistical assumptions to be made regarding the distribution of data [43]. The basic structure of the decision tree, however, consists of one root node, a number of internal nodes and finally a set of terminal nodes [44]. A node is a subset of the predictors that is used to determine a split. A non-terminal node or parent node is a node that is further split into two child nodes. Growing a tree consists of selecting the optimal splits to determine a non-terminal node, and the assignment of each terminal node to a class. The data is recursively divided down the decision tree according to the defined classification framework.

Classes are simply assigned to a terminal node by observing which class is mostly commonly observed in that region of the tree. Thus, the challenge is to optimally choose the best variable and split that variable to maximize the purity or similarity among the responses. The impurity

of a parent node τ, denoted $i(\tau)$, is zero when all observations are in the same class. A split s is determined by selecting the best predictor and split value that optimizes the highest reduction in purity [45, 46]

$$\Delta(s,\tau) = i(\tau) - \sum_{b=1}^{B} p(\tau_b/\tau)\, i(\tau_b) \quad \text{(eq. 12)}$$

where τ_b denotes child node b, $p(\tau_b/\tau)$ is the proportion of observations in τ that are assigned to τ_b, and B is the number of branches after splitting. Two common impurity functions are the entropy criterion [33]

$$i(\tau) = -\sum_{k=1}^{K} p_k \log_2(p_k) \quad \text{(eq. 13)}$$

and the Gini index criterion

$$i(\tau) = -\sum_{k=1}^{K} p_k^2 \quad \text{(eq. 14)}$$

where p_k is the proportion of observations in class c_k with $k \in \{1, \ldots, \mathrm{K}\}$. Pruning is based upon successive steps of removing lower branches that lead to improved classification rates. Once the final tree is determined by $\Delta(s,\tau)$, it is natural to evaluate its predictive performance by comparing the observed class to the predicted class from the CT for observation $\mathbf{x_j}$. In a terminal node m, representing a region R_m with n_m observations, let

$$\hat{p}_{mk} = \frac{1}{n_m}\sum_{j=1}^{n_m} \mathbf{1}_{c_k}(\mathbf{x_j}) \quad \text{(eq. 15)}$$

denote the proportion of class c_k observations in terminal node m. We classify the observations in node m to class

$$\widehat{c_k}(\mathbf{x_j}) = \underset{\mathrm{k}}{\operatorname{argmax}}\, \hat{p}_{mk} \quad \text{(eq. 16)}$$

The misclassification error rate is simply the proportion of observations in the node that are not members of the majority class in that node.

$$R^{DTC}(\mathbf{x_j}) = \frac{1}{n}\sum_{j=1}^{n}\left(1 - \max_{k}(\hat{p}_{mk}(\mathbf{x_j}))\right) \quad \text{(eq. 17)}$$

4.4.Multi-Class Support Vector Machine (SVM)

Support vector machines (SVMs) are a type of learning algorithms that are used for classification and regression. SVM classifiers, like Decision Trees, are non-parametric [47]. The one-against-all technique [48] is now the most fundamental technique for implementing SVM multi-class classification. K binary SVM models are built in this simplest application of the SVM to a K-class issue. Class c k is isolated from the other classes in the kth class SVM problem. To create a final multi-class classifier, all k binary SVM classifiers are concatenated. The term "remaining" refers to the fact that all data points from classes other than c_k are merged to form a single class c_l. Using the usual SVM technique [42], the ideal hyperplane that separates data points from the class c_i and the combined class c_l is obtained. The best separating hyperplane distinguishing the classes c_i and c_k is denoted as:

$$g_k(\mathbf{x_j}) = \mathbf{w_k}.\varphi(\mathbf{x_j}) + \mathbf{b_k} \qquad \mathbf{k} \in \{1, \dots, \mathrm{K}\} \qquad \text{(eq. 18)}$$

where $\mathbf{w_k} \in \mathbb{R}^{\mathrm{S}}$ is the weight vector, b is the bias, and the transfer function φ maps the training data into an appropriate feature space $\mathbb{R}^{\mathrm{S}}$ to allow for complex nonlinear surfaces. The following minimization is used to estimate the parameters of the decision function $g_k(\mathbf{x_j})$:

$$\min J(\mathbf{w_k}, \xi) = \frac{1}{2}\|\mathbf{w_k}\|^2 + C\sum_{j=1}^{n}\xi_j \qquad \text{(eq. 19)}$$

subject to

$$y_j\left(\mathbf{w_k^T}\varphi(\mathbf{x_j}) + \mathbf{b_k}\right) \geq \mathbf{1} - \xi_j \qquad \xi_j \geq \mathbf{0}\ ; j = 1 \dots \mathrm{n}, \qquad \text{(eq. 20)}$$

To loosen the separability restrictions in with scalar $y_j \in \{-1, +1\}$ signifying its class label, $\mathrm{C} \in \mathbb{R}^+$ is a regularization constant, and ξ_j is a slack variable (eq. 19).

The $f_k(\mathbf{x_j})$ decision rule, which allocates the vector $\mathbf{x_j}$ to the class c_k provided by:

$$f_k(\mathbf{x_j}) = sign(g_i(\mathbf{x_j})) \qquad \text{(eq. 21)}$$

The fundamental challenge with this strategy is that the classifier outputs $f_k(\mathbf{x_j})$ are binary numbers. The standard approach to this problem is to disregard the sign-operator in eq. 21. We

say $\mathbf{x_j}$ is in the class with the biggest value of the decision function and is provided by after finding all the optimum hyperplanes given by $g_k(\mathbf{x_j})$ for $\mathbf{k} \in \{1, \ldots, \mathrm{K}\}$.

$$\widehat{c_k}(\mathbf{x_j}) = \underset{k}{\operatorname{argmax}}\, g_k(\mathbf{x_j}) \qquad \text{(eq. 22)}$$

The index of the biggest component of the different classifiers $g_i(\mathbf{x_j})$ for $\mathbf{k} \in \{1, \ldots, \mathrm{K}\}$ is allocated to the data point $\mathbf{x_j}$ in this manner. The SVM classifier's error rate R^{SVM}, which is defined as:

$$R^{SVM}(\mathbf{x_j}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{c_k \neq \hat{c}_k}(\boldsymbol{x_j}), \qquad \text{(eq. 23)}$$

with $\boldsymbol{x_j}$ belonging to the class c_k as assessed by the method classifier in the class $\hat{c}_k$ and $\mathbf{1}_{c_k \neq c_k}(\boldsymbol{x_j})$ being the indicator function defined as:

$$\mathbf{1}_{c_k \neq \hat{c}_k}(\boldsymbol{x_j}) = \begin{cases} 1 \text{ if class } \hat{c}_k \text{ of } \boldsymbol{x_j} \neq \text{ class } c_k, \\ 0 \text{ if class } \hat{c}_k \text{ of } \boldsymbol{x_j} = \text{ class } c_k. \end{cases} \qquad \text{(eq. 24)}$$

4.5.Artificial Neural Network

Artificial Neural network (ANN) method is an algorithm in machine learning and artificial intelligence, which is inspired by the human nervous system to analyze and model complex, nonlinear systems, and parallel computations [49]. Haykin introduces an artificial neural network as a massively parallel learning machine, made up of simple processing units called neurons [50].

A neuron has asset of inputs $x_1, x_2, \ldots, x_R$ and each connection from the inputs to the processing element is affected by different connection strengths known as synaptic weights. A neuron is showed in mathematical and schematic terms by the following equations (25 ,26) and Figure 1.

$$\Sigma = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b \qquad \text{(eq. 25)}$$

$$y = f(\Sigma), \qquad \text{(eq. 26)}$$

Where x_1, x_2, …, x_n are the inputs and w_1, w_2, …, w_n are the synaptic weights of the neuron L. b is the bias, f is the activation function, and y is the output of the neuron.

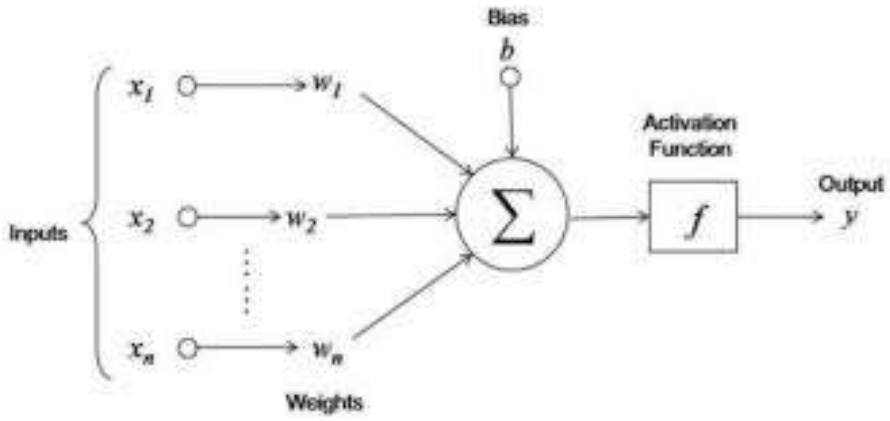


Figure 1. Structure of a neuron

The output of the neuron is defined by applying an activation function to the linear combination of inputs and weights. There are many different activation functions as shown in Figure 2.

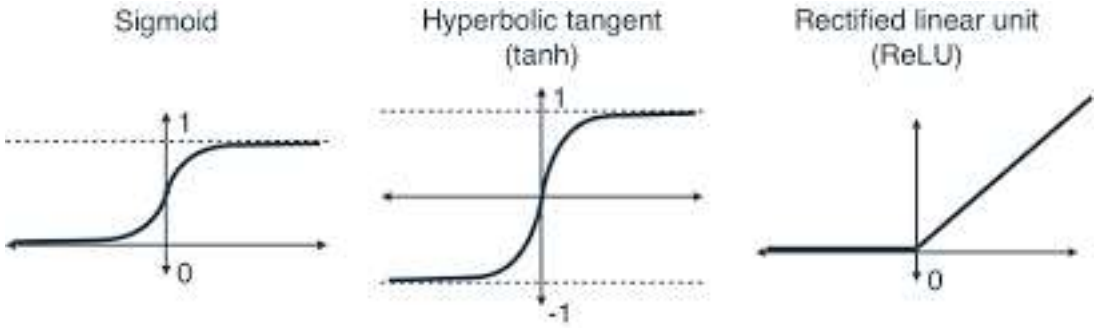


Figure 2. Different activation functions

The procedure in which neurons are arranged in a neural network is known as the network topology or architecture. There are many different types of neural network such as MLP, RBF and PNN. One of the famous ANN architecture which is most widely used in remote sensing application is multilayer perceptron, or MLP that has been used in this study. An MLP is a feed-

forward neural network with one or more layers of neurons between the input and output layer called hidden layers [51].

4.6. Extended Gamma Associative Classifier

The Gamma Associative Classifier (EG) is a supervised learning used as prediction model for classification. This algorithm has able to handle mixed data, or absence of information in the data. The extended gamma association classifier (EG) includes two stages: training and classification. Let X and Y be the training and testing datasets, respectively, where each instance $x_j \in$ X, $y_j \in$ P is described by a set of attributes or factors. [36]

The training stage of the classifier starts with the storage of the training set and incorporates the calculations of various parameters such that the pause and stop parameters. Within the classification stage, EG employments an iterative handle, based on the calculation of the average similarity with each class of the samples to be classified [52]. To analyze the similitude between test sample and training sample, the extended generalized similitude γ_{ext} is utilized. After getting the similitudes, the average of the generalized similarity of said test pattern for each class c_k is calculated. The number of samples belonging to the class k in the training data is given by n, and x_{ij} represents the value of the j-th attribute of the i-th sample of the class c_k, and w_j represents the weight of the jth feature with $w_j \in [0\ 1]$.

$$c_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} w_j \gamma_{ext}(x_j, y_j, \theta)}{n} \quad \text{(eq. 27)}$$

$$\gamma_{ext}(x_j, y_j, \theta) = \begin{cases} \gamma_{num}(x_j, y_j, \theta) & \text{if jth feature is numeric} \\ \gamma_{cat}(x_j, y_j) & \text{if jth feature is categoric} \\ \gamma_{miss}(x_j, y_j) & \text{if } x_j \text{ or } y_j \text{ are missing} \end{cases} \quad \text{(eq. 28)}$$

where

$$\gamma_{num}(x_j, y_j, \theta) = \begin{cases} 1 & \text{if } |x_j - y_j| \le \theta \\ 0 & \text{otherwise} \end{cases} \quad \text{(eq. 29)}$$

$$\gamma_{cat}(x_j, y_j) = \begin{cases} 1 & \text{if } x_j = y_j \\ 0 & \text{otherwise} \end{cases} \quad \text{(eq. 30)}$$

$$\gamma_{miss}(x_j, y_j) = \begin{cases} 1 & \text{if } x_j = y_j = \text{"?"} \\ 0 & \text{otherwise} \end{cases} \quad \text{(eq. 31)}$$

In case a single greatest is found among all the values of c_k , the process ends. If not, the values of the stop and pause parameters, as well as the value of the θ parameter, are taken under consideration in an iterative process [53].

4.7. Naïve Associative Classifier

The Naïve Associative Classifier (NAC) a novel supervised learning model, directly handles mixed and incomplete data [37]. In its training stage, the NAC stores the training set and calculates, for each numerical attribute, the standard deviation. However, the NAC classifier needs to store the entire training set of data. It also uses similarity comparisons with respect to the entire training set to classify a new sample [54].

In the classification stage, the Mixed and Incomplete Data Similarity Operator (MIDSO) and the total similarity operator are used to analyze the similarity between the test samples and the training samples $s^t(x, y)$.

$$s^t(x, y) = \sum_{i=1}^{m} w_i * MIDSO(x, y, A_i) \quad \text{(eq. 32)}$$

$$MIDSO(x, y, A_i) = \begin{cases} s_n(x, y, A_i) & \text{if } A_i \text{ is numeric} \\ s_c(x, y, A_i) & \text{if } A_i \text{ is categoric} \end{cases} \quad \text{(eq. 33)}$$

$$s_c(x, y, A_i) = \begin{cases} 0 & \text{if } \left((x_j \neq y_j) \vee (x_j =' \,?') \vee (y_j = '?')\right) \\ 1 & \text{otherwise} \end{cases} \quad \text{(eq. 34)}$$

$$s_n(x, y, A_i) = \begin{cases} 0 & \text{if } \left((|x_j - y_j| > \sigma_i) \vee (x_j =' \,?') \vee (y_j = '?')\right) \\ 1 & \text{otherwise} \end{cases} \quad \text{(eq. 35)}$$

With w_i represents the weight of the ith attribute. After having obtained the similarities, we calculated the average of the generalized similarity of the test instance for each class kl, noted $s_l(.)$.

$$s_l(.) = \frac{1}{n}\sum_{y \in k_l} s^t(., y) \quad \text{(eq. 36)}$$

If a single maximum is found among all the values of $s_l(.)$, the process ends. If not, any of the classes with maximum similarity is assigned [55].

5. Application and results

In this section, we first present data collection that contains the several indicators that directly or indirectly affects the spread of the coronavirus pandemic. The discrimination analysis of samples of infected patients using the proposed approach of supervised classification algorithms is presented and discussed regarding the size of training and validation dataset. All data treatments were performed using the MATLAB® R2019b environment, and scripts are available upon request.

5.1.Data set collection

The essential pavement procedure to examine the test hypothesis and to evaluate the predicted outcome is in collecting and processing the data. Thus, to predict the behavior of coronavirus spread, four countries have been chosen. These countries have adopted different methodologies to deal with the coronavirus pandemic and achieved different results related to the methodology used: China, Lebanon, Italy, and Iran. The date range for the data for each of the four countries is as the following:

China: 9 January 2020 – 28 March 2020 (80 days)

Lebanon: 21 February 2020 – 31 March 2020 (40 days)

Italy: 31 January 2020 – 31 March 2020 (61 days)

Iran: 19 February 2020 – 31 March 2020 (42 days)

After studying the situation of the virus in these countries, we noticed that there are several indicators that affect directly or indirectly the spread of the coronavirus. These indicators are presented as follows :

- The governance reactions: the different responses and reactions from governments of the four countries during the outbreak. These indices are used to explore whether the government response affects the rate of infection and identify correlates of intense responses.
- The medical resources: This factor refers to the health system policies such as the COVID-19 testing regime or emergency investments into healthcare (ICU beds...) and the health services quality in these four countries. The sensitivity effects of this factor on the results are proposed to be investigated in this study.
- The quarantine commitment of the people in the country with the government Guidelines. This factor must have theoretically a straight depending on relation with the variation of the spread intensity of the virus between the countries.
- The special events: this factor takes into consideration the existence of simultaneous events that affect the spread of the coronavirus: other disasters, economic problems, war, political problems or disturbance, official holidays, etc.
- The economical level and governmental aids: This factor refer to the economic policies, such as income support to citizens or the provision of foreign aid. Depending on the direct relationship between this factor and the quarantine commitment of the people in the country, we have proposed it to be present in this study.
- Previous experience history of the four governments that determines the existence of experience in such kind of critical disaster management or not.

- The use of technology devoted to control the virus spread in these countries that help in the health and hospitalization services, the lockdown control, and the restrictions of the infected zones.
- The population age average/ The population density which is considered as the number of the people in one km^2/ in the four countries which causes a meaningful variation in the situation of the infected records.

These direct and indirect factors are considered as parameters used by our model to predict the future behavior of the spread of COVID-19. The data is combined into a series of novel indices that aggregate various measures of each factor. Parameters are measured depending on a specific criterion presented in the following table 2:

Table 2. The direct and indirect factors that affecting the spreading level of the coronavirus in different countries.

Effective factors	Indications	Values
Governance reactions (Lockdown)	No lockdown measures at all (value =0).	0, 0.25, 0.5, 0.75, and 1
	Disabling major facilities: Stop flights at airports, public transport, ports, tourist places, universities and school closures and restrictions in movement in the public places (value =0.25)	
	Partially lockdown measures (value=0.5)	
	Fully lockdown measures (value=0.75)	
	Providing virus checks abundantly (value=1)	
Medical resources	0 = no medical resources, 0.25 = low, 0.5 = medium, 0.75 = good, 1 = high	0, 0.25, 0.5, 0.75, and 1
	• Number of nurses for 1000 people	
	• Number of doctors for 1000 people	
	• Number of icu beds for 1000	
Previous experience history	yes / no	1/0

The used technology	0 = no technology, 0.25 = low, 0.5 = medium, 0.75 = good, 1 = high	0, 0.25, 0.5, 0.75, and 1
Special events	yes / no	1/0
Economical resources and governmental aids	0 = no economic resources, 0.25 = low, 0.5 = medium, 0.75 = good, 1 = high	0, 0.25, 0.5, 0.75, and 1
Population density / km2	*	Number
Family number	*	Number
Quarantine commitment (Procedure)	1 = no commitment, 0.75 = low, 0.5 = medium, 0.25 = good, 0 = high	0, 0.25, 0.5, 0.75, and 1
The infected patients by the Covid-19	- new confirmed daily cases	Number

Thus, we have collected real data for the four countries, from different official sources with all parameters and daily records (recovered, death, confirmed, PCR tests).

Many factors have been known to be associated with the initial levels of the Coronavirus outbreak at the country level including geographical factors [56], demographical parameters [57], healthcare services [58], and economic status [59].

As for the technique of data collection, we relied on studies conducted in these four countries to obtain them. In Iran, the demographical data were extracted from the Statistical Centre of Iran [60], healthcare services and economic levels were extracted from [61]. The number of daily confirmed cases was obtained from the Iranian Ministry of Health and Medical Education website.

In China, these factors have been selected using data from the Figshare site, nationwide survey in China and World Health Organization (WHO) [62]. In Italy, the collected data were extracted from the most recently updated databases on the website of the Italian Institute of Statistics

(ISTAT) [63], and the ISTAT's Health for All database [64, 65]. In Lebanon, the data has been collected from the Ministry of Health [66], Interior and Finance to obtain these factors and the number of daily confirmed cases.

In another study, we conducted a Path analysis which is a statistical technique that uses both bivariate and multiple linear regression techniques to test the causal relations among the variables specialized in the model [67]. We evaluated the parameter estimates of the previously stated regressors on covid-19 daily cases to assess the most influential factors that contribute to the prediction of day-by-day covid-19 infections. The results showed that quarantine commitment ($\beta = -0.823$) and full lockdown measures ($\beta = -0.775$), has the largest direct effect on covid-19 daily infections. The results also show that more experience ($\beta = -0.35$), density in society ($\beta = -0.288$), medical resources ($\beta = 0.136$), and economy resources ($\beta = 0.142$) have indirect effects on covid-19 daily infections.

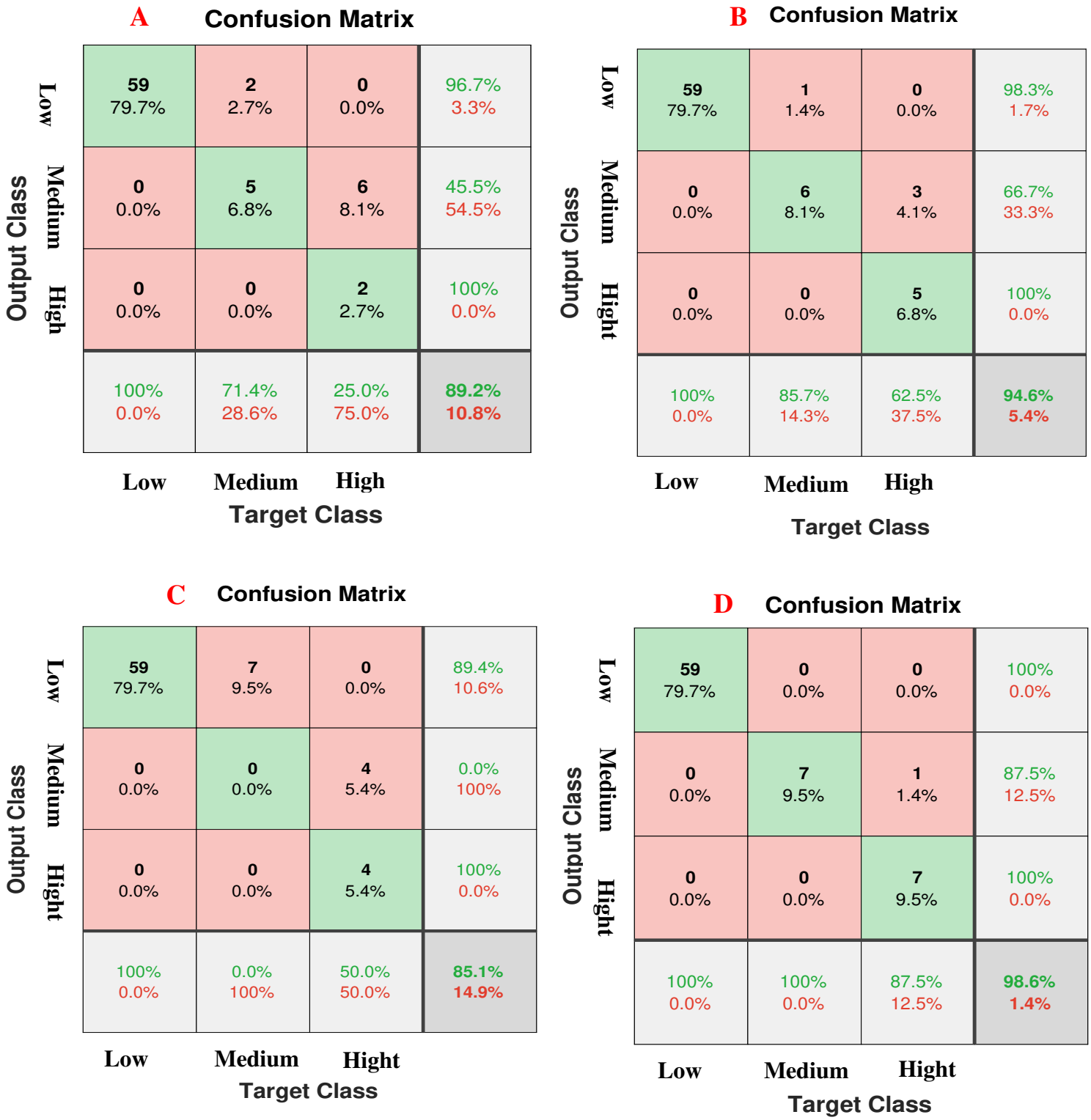
These direct and indirect factors allow to measure the score of each sample which represents the times in days in four countries: China, Iran, Italy, and Lebanon. The daily records are basically the cumulative records for a dependent day-by-day scale. In other words, the record of the next day is the sum of the records of the current day and the new records obtained in the same day. The samples were classified into 3 categories which are the degrees of possibility of Coronavirus disease infection according to their evolution (in days): Low (if the standardized cumulative cases are between 0 and 0.09), medium (if the standardized cumulative cases are between 0.09 and 0.3) and High (if the standardized cumulative cases are above 0.3). The data collected is an imbalanced data meaning that a certain degree of possibility of infection is dominant over other degrees, where in our case the low-risk degree dominates the other two degrees. This is due to the data collected at the beginning of the covid-19 pandemic where most countries are still at low risk of infection and the virus is not fully spread. Regarding the distribution of the degree of possibility of getting infected by covid-19 among the countries

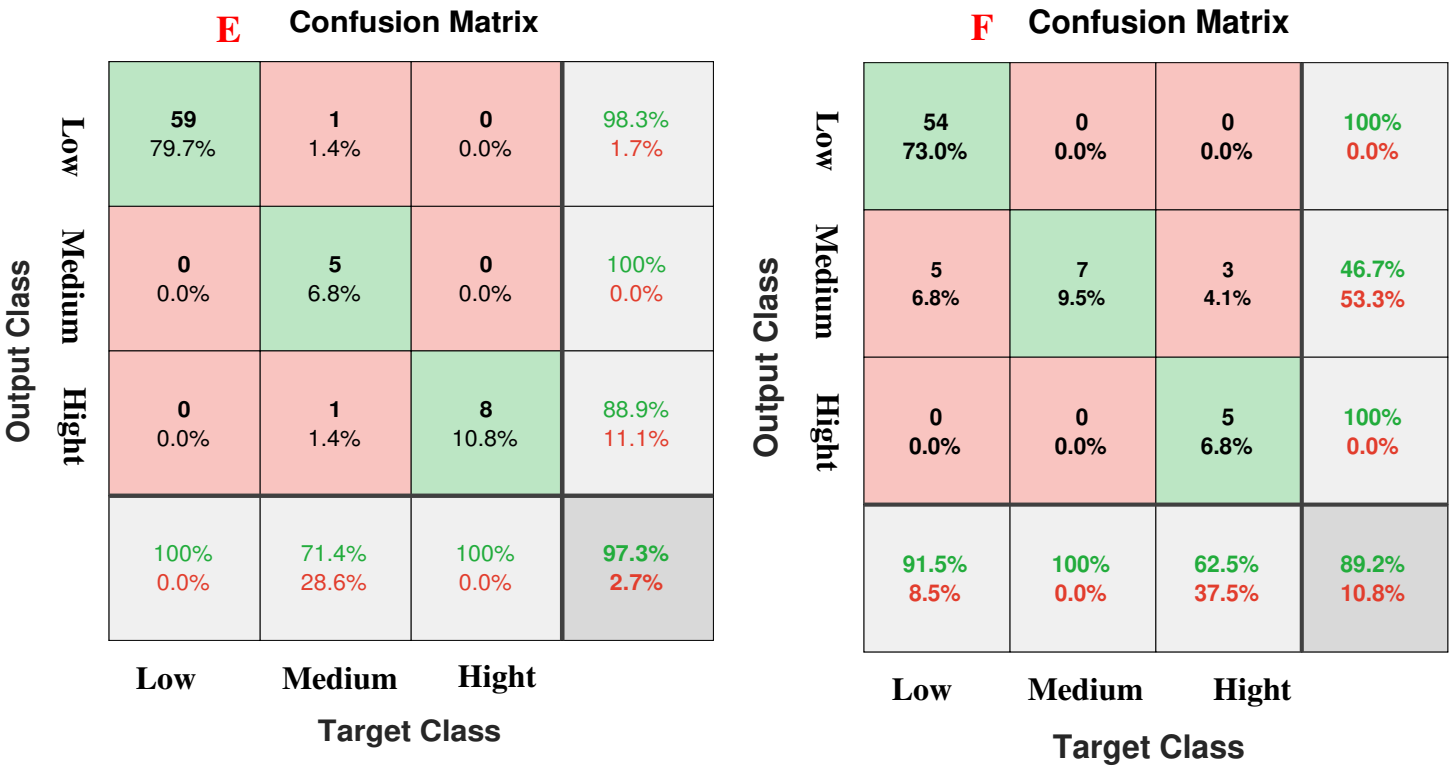
studied, China (80 days) and Lebanon (40 days) remained at low risk for the whole durations while Iran maintained a low risk of infection in the first 20 days, medium risk of infection for the next 15 days and high risk of infection for the last 7 days of the 42 total days studied in this country. As for Italy, it remained at a low risk for the first 36 days, medium risk for the next 7 days, and high risk for the rest 18 days of the 61 total days studied in Italy.

In addition, the data collected do not include missing values. The application of supervised learning methods may result in better accuracy, unless a missing value is expected to have a very high variance. All Covid-19 data were preprocessed with a Standard Normal Variate (SNV) preprocessing method [68].

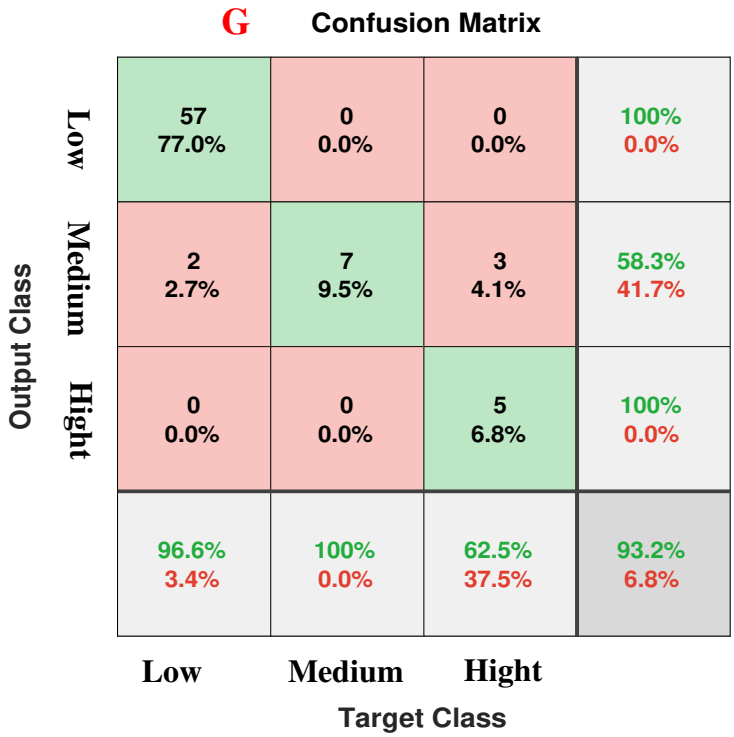
5.2.Analysis of discrimination of samples of infected patients by the Covid-19

To deepen the obtained results, we analyzed the confusion matrices and the ROC curves. A confusion matrix is used to describe the performance of a classification model on a set of test data for which the true values are known. It is a summary of prediction results on a classification problem and allows the visualization of the performance of an algorithm. Hence, the elements in the diagonal are the elements correctly classified, while the elements out of the diagonal are misclassified. It can easily be seen that the overall accuracy can be computed by dividing the sum of the diagonal elements (number of correctly classified samples) by the total sum of the elements of the matrix (total number of samples in the dataset).


A
Confusion Matrix
Output Class
Low
Medium
High
59
79.7%
2
2.7%
0
0.0%
96.7%
3.3%
0
0.0%
5
6.8%
6
8.1%
45.5%
54.5%
0
0.0%
0
0.0%
2
2.7%
100%
0.0%
100%
0.0%
71.4%
28.6%
25.0%
75.0%
89.2%
10.8%
Low
Medium
High
Target Class
B
Confusion Matrix
Output Class
Low
Medium
Hight
59
79.7%
1
1.4%
0
0.0%
98.3%
1.7%
0
0.0%
6
8.1%
3
4.1%
66.7%
33.3%
0
0.0%
0
0.0%
5
6.8%
100%
0.0%
100%
0.0%
85.7%
14.3%
62.5%
37.5%
94.6%
5.4%
Low
Medium
High
Target Class
C
Confusion Matrix
Output Class
Low
Medium
Hight
59
79.7%
7
9.5%
0
0.0%
89.4%
10.6%
0
0.0%
0
0.0%
4
5.4%
0.0%
100%
0
0.0%
0
0.0%
4
5.4%
100%
0.0%
100%
0.0%
0.0%
100%
50.0%
50.0%
85.1%
14.9%
Low
Medium
Hight
Target Class
D
Confusion Matrix
Output Class
Low
Medium
Hight
59
79.7%
0
0.0%
0
0.0%
100%
0.0%
0
0.0%
7
9.5%
1
1.4%
87.5%
12.5%
0
0.0%
0
0.0%
7
9.5%
100%
0.0%
100%
0.0%
100%
0.0%
87.5%
12.5%
98.6%
1.4%
Low
Medium
Hight
Target Class



544



545

546 Table 2. Results of average confusion matrix (of 100 runs) for the training (2/3 of samples) and

547 testing (1/3 of samples) dataset using the supervised algorithms (A) LDA (B) NBC (C) SVM

548 (D) DTC (E) ANN (F) EG (G) NAC methods for the detection of infected patients by the Covid-
549 19 according to the time of evolution (in days) of coronavirus disease in different country
550
551 Cross-validation is a model assessment technique used to evaluate a supervised classification
552 algorithm's performance in making predictions on new datasets that it has not been trained on.
553 This is done by partitioning a dataset and using a subset to train the algorithm and the remaining
554 data for testing. Several methods of cross validation exist such as k-fold, Holdout, Leave-out
555 and Re-substitution. We use k-fold method that partitions samples of infected patients by the
556 Covid-19 into k = 2 until 10 randomly chosen subsets of roughly equal size. With this approach
557 there is not a possibility of high bias if we have limited data, because we would not miss some
558 information about the data which we have not used for training. the discussion of choosing the
559 value of K in K-fold cross validation, is shown in section 5.3. but here we will display the
560 results of K = 3-fold. One subset is used to validate the model (1/3 of data = 74 samples) trained
561 using the remaining subsets (2/3 of data = 149 samples). This process is repeated T=100 times
562 to get the best stable model and to ensure that most of our samples of infected patients by the
563 Covid-19 have been used for validation.
564 The Linear Discriminant Analysis (LDA), Naïve Bayes Classifier (NBC), Support Vector
565 Machine (SVM), Decision Tree Classifier (DTC), Artificial Neural Network (ANN), Extended
566 Gamma Associative Classifier (EG) and Naive Associative Classifier (NAC) proposed
567 supervised algorithms has been applied on the coronavirus data. Tables 2 (A)(B)(C)(D)(E)
568 presents the performance of the LDA, NBC, SVM, DTC and ANN classifications as the mean
569 over the T=100 samplings of the confusion matrix information for the samples of infected
570 patients by the Covid-19 according to the time of evolution (in days) of coronavirus disease in
571 different country. Each table show the confusion matrix for the true label's targets and predicted
572 labels outputs. The rows correspond to the predicted class (Output Class) extract by
573 classification methods and the columns correspond to the infected samples degree true class

(Target Class). We have three degrees of possibility according to their evolution of coronavirus that represents the classes of our COIVID-19 data: Low, Medium and Hight. The diagonal cells correspond to spectra that are correctly classified. The off-diagonal cells correspond to incorrectly classified of degree of infected samples (time of evolution) by the Covid-19.

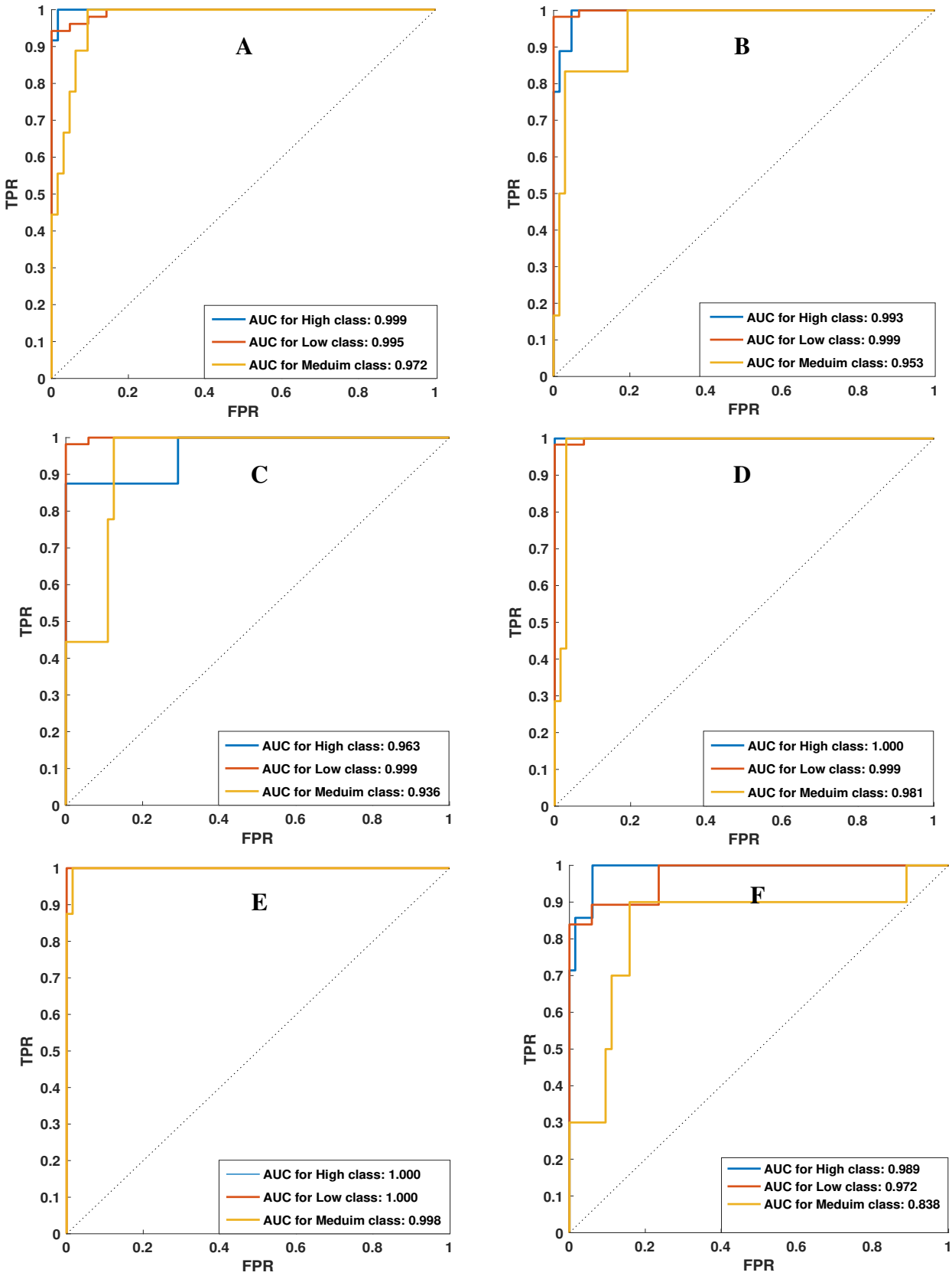
The results of classification accuracy for Detection True Classifier (DTC) and Artificial Neural Networks (ANN) algorithms registered 98.6% and 97.4% of validation datasets describing the degree of possibility of Coronavirus disease infection. However, the classification accuracy is worse for the Naïve Bayes Classifier (NBC) algorithm with 94.6 % of validation datasets correctly classified, the Support Vector Machine (SVM) algorithm with 85.1 % of validation datasets correctly classified and Linear Discriminant Analysis (LDA) algorithm with 89.8 % of validation datasets correctly classified. These values, presented in Table 3, confirm that the DTC and ANN algorithm allows a better discriminate within the degrees of possibility of Coronavirus disease infection according to days of their evolution than state of art methods of supervised classifications such as NBC, SVM, EG, NAC and LDA algorithms.

A ROC (Receiver Operating Characteristic) curve summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) by varying the threshold for assigning observations to a given class. A ROC curve is a performance measurement for classification problem at various thresholds settings. The plot of True Positive Rate (TPR; sensitivity) versus False Positive Rate (FPR; 1-specificity) across varying cut-offs generates a curve in the unit square called a ROC curve. ROC curves are used to compare different supervised classifiers. In our case we have used the multi-class ROC curves, which is a kind of multi-objective optimization to compare the supervised classification algorithm; LDA, NBC, SVM, ANN, DTC, EG and NAC. As in several multi-class problem, the idea is generally to carry out pairwise comparison: one class

vs. all other classes, one class vs. another class [69]. If the curve has high values, it leads to the greater of area under the curve obtained, and the less error the classifier makes. ROC curve corresponding to progressively greater discriminant capacity are located progressively closer to the upper left-hand corner in "ROC space". To numerically assess the discrimination of the classes, the Area Under the Curve (AUC) was used to measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. Higher the AUC, better the model is at distinguishing between classes of degrees of possibility of Coronavirus disease.

The figure below represents the comparisons between ROC curves by LDA, NBC, SVM, ANN and DTC for each class of degrees of possibility of Coronavirus disease infection according to days of their evolution for validation datasets (1/3 samples). The ROCs of each model are plotted using the average of 100 runs. We can see that the classification threshold in ANN and DTC methods (figures 3D and 3E) are very close to 1 and very far from diagonal for all the classes of degrees of possibility of Coronavirus disease compared to other methods. But the classification threshold is relatively reduced for the LDA and SVM methods (figures 3A and 3B). The area under the curves of each class of degrees of possibility of Coronavirus disease for the ANN and DTC methods are higher than that for NBC, SVM, LDA, EG and NAC methods. This also confirms that the measures of sensitivity and specificity of ROC curves for ANN and DTC algorithms are generally stronger than the other ROC curves for NBC, SVM, LDA EG and NAC algorithms for each class of degrees of possibility of Coronavirus disease in all figures 3(A-G), which indicates better in-sample accuracy than the other classifier methods. The AUC values, presented in Figures 3 (A-G), confirm that the ANN and DTC supervised methods allows a better classifier of degrees of possibility of Coronavirus disease.

622


A
AUC for High class: 0.999
AUC for Low class: 0.995
AUC for Meduim class: 0.972
B
AUC for High class: 0.993
AUC for Low class: 0.999
AUC for Meduim class: 0.953
C
AUC for High class: 0.963
AUC for Low class: 0.999
AUC for Meduim class: 0.936
D
AUC for High class: 1.000
AUC for Low class: 0.999
AUC for Meduim class: 0.981
E
AUC for High class: 1.000
AUC for Low class: 1.000
AUC for Meduim class: 0.998
F
AUC for High class: 0.989
AUC for Low class: 0.972
AUC for Meduim class: 0.838
TPR
FPR

623

624

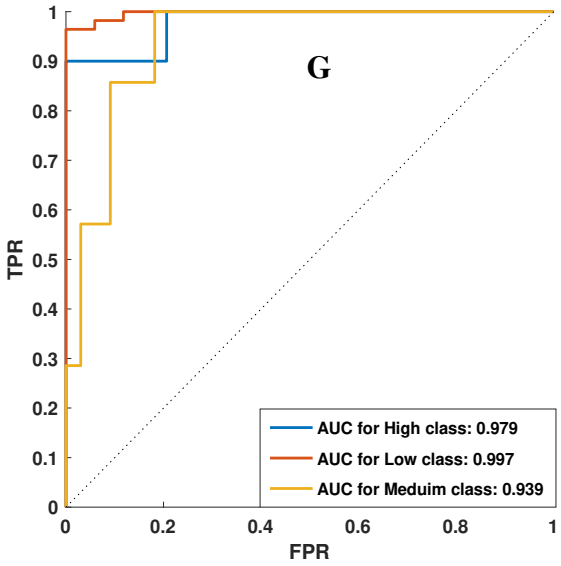


Figure 3. The average of 100 iterations of Receiver operating characteristic (ROC) curve of A) LDA, (B) NBC, (C) SVM, (D) ANN and (E) DTC (F) EG (G) NAC supervised classifications algorithms for classify the degrees of possibility of Coronavirus disease infection according to days of their evolution using the validation datasets.

5.3.Effect of the size of testing and training datasets

As we already mentioned, cross validation is the most important part of machine learning models, and the results might depend on it. K-fold cross-validation is extremely useful with the appropriate K which is the number of folds chosen.

Unfortunately, there is no theoretically perfect procedure of determining the appropriate value of K in K-fold cross validation. Many researchers used K = 10, for example, see Bengio and Grandvalet (2004), Vehtari and Lampinen (2004). Davison and Hinkley (1997) recommend K = min ($n^{1/2}$, 10) in practice. Clark (2003) compares different values of K for his data and suggested that the choice of K = 4 will probably be good in general, though not necessarily optimal. Duan et al. (2003) used K = 5 for the number of folds. Anderson et al (2006) suggested, leaving 20% of the n-samples at a time for final model validation. For this, we have

tested the method of supervised learning classification LDA, NBC, SVM, ANN and DTC on different partitions of testing and training dataset depending on different values of K between 2 and 10. To examine the influence of the cross-validation method, we have computed the ratio of classification (%) obtained for each method of the supervised learning classification LDA, NBC, SVM, ANN and DTC on COVID-19 data. The values in Table 3, indicate that the detection tree classifier (DTC) algorithm gives the best results of classification for different sizes of testing and training datasets for the samples of infected patients by Covid-19 according to the time of evolution (in days) of coronavirus disease in different countries. This sampling analysis is done to assess the importance of K-fold cross-validation method in determining the best classifier among the classifiers studied.

Table 3. The Means and Standard Deviation of classification performance (expressed as %) for 100 runs of LDA, NBC, SVM, ANN, DTC, EG and NAC supervised methods on different sizes of testing and training dataset of coronavirus diseases or different values of K in K-fold cross validation.

	K-fold	LDA	NBC	SVM	ANN	DTC	EG	NAC
1/2 training & 1/2 testing	2	87.4 ± 2.7	91.4 ± 2.2	82.3± 4.5	91.2± 2.8	96.9 ± 1.7	83.1 ± 5.2	91.8 ± 3.3
2/3 training & 1/3 testing	3	89.5 ± 3.4	94.5 ± 2.5	85.6 ± 5.2	97.3± 1.8	98.6 ± 0.7	89.2 ± 4.2	93.2 ± 2.6
3/4 training & 1/4 testing	4	89.5 ± 3.2	94.5± 1.9	87.1 ± 4.4	98.1± 0.8	98.9 ± 0.5	90.4 ± 3.8	93.3 ± 2.6
4/5 training & 1/5 testing	5	90.2 ± 3.4	95.4 ± 2.6	87.4 ± 3.8	99.1 ± 0.5	99.2 ± 0.3	91.2 ± 3.8	93.5 ± 2.4
9/10 training & 1/10 testing	10	91.1 ± 2.8	95.8 ± 2.2	88.9 ± 3.8	99.5 ± 0.2	99.5 ± 0.3	91.4 ± 3.2	94.5 ± 2.2

To quantify the results of the AUROC of each model performed 100 runs, we calculated in The Table 4 the mean and standard deviation of the AUROC for three classes of degrees of possibility of infection with Coronavirus disease infection and the proposed supervised learning methods.

Table 4. The mean and SD of the AUROC for three classes of degrees of possibility of Coronavirus disease infection.

	Low	Medium	High
LDA	0.995 ± 0.004	0.972 ± 0.024	0.999 ± 0.001
NBC	0.999 ± 0.001	0.953 ± 0.042	0.993 ± 0.006
SVM	0.999 ± 0.001	0.936 ± 0.064	0.963 ± 0.032
DTC	0.999 ± 0.001	0.981 ± 0.018	1.000 ± 0.000
ANN	1.000 ± 0.000	0.998 ± 0.001	1.000 ± 0.000
EG	0.972 ± 0.027	0.838 ± 0.158	0.989 ± 0.001
NAC	0.979 ± 0.026	0.939 ± 0.058	0.979 ± 0.028

In application of machine learning algorithms for classification tasks, we need to apply approximate statistical test for determining whether one learning algorithm out-performs another on a particular learning task. For this we propose to apply McNemar's test that is based on a X2 test for goodness of fit. So, we may reject the null hypothesis in favor of the hypothesis that the two algorithms have different performance if P-value is greater than 0.05. We found that the application of McNemar test between seven used supervised algorithms, gives the existence of performance difference between DTC and ANN algorithms and other algorithms, and there is no performance difference between DTC and ANN algorithms.

6. Conclusions and Future Research

COVID-19 has shocked the world due to the speed of its spread and its non-availability of vaccine or drug. Various researchers are working for conquering this deadly virus. We used 224 reports recorded on specified dates in four countries which are labelled in three classes of degrees of infections for each day: Low, medium and Hight possibility of Coronavirus disease infection. Various direct and indirect factors like the economical level and governmental aids, medical resources, previous experience, Governance reactions and the used technology are being extracted from these COVID-19 samples according to the time of evolution (in days) of coronavirus disease in different countries.

The machine learning algorithms are used for classifying daily samples in four countries into three different classes degrees of possibility of Coronavirus disease infection. After performing classification, it was revealed that decision trees classifier (DTC) and artificial neural networks (ANN) gives excellent results by having 98.6% and 97.3% accuracy well classified. Other machine learning algorithm that showed better results were Naive Bayesian classifier (NBC) by having 94.5% accuracy well classified. The efficiency of models can be improved by increasing the amount of data.

Further studies should be performed to investigate some issues, for example, a study that takes into account a priori information such as the degrees of possibility of Coronavirus disease into the objective function through Partial Least Square (PLS) models. Furthermore, it would be very interesting to apply of new technology in the search for the degrees of possibility of Coronavirus disease infection such as the deep learning technique which allows to extract the spectral features to classify the state of COVID-19.

7. Declarations

Acknowledgements

This study was carried out by Faculty of Engineering-Islamic University of Lebanon in partnership with Lebanese University, Statistics and Informatics Department - Faculty of Science in Beirut.

Funding

This project has been funded with the support of the National Council for Scientific Research in Lebanon CNRS-L.

Ethics approval and consent to participate

The study does not involve human participants. Not applicable

Consent for publication

Not applicable

Availability of data and material

All data, models, and code generated or used during the study appear in the submitted article and are provided upon request by contacting Abbas Rammal via email: rammal_abbass@hotmail.com

Competing interests

The authors declare that there are no competing interests.

Authors' contributions

AR: Conceptualization, Methodology, Software, Data curation, Visualization, Validation, Investigation, Original and revised draft preparation, Writing-Reviewing, and Editing. EY: Data curation, Investigation, Original and revised draft preparation, Writing-Reviewing, and Editing. WF and RK: Conceptualization, Methodology, Data Collection, and Writing. All authors read and approved the final manuscript.

8. References

[1] Wu, F., Zhao, S., Yu, B., Chen, Y., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z. 2020. A new coronavirus associated with human respiratory disease in china. Nature. 44(59), 265–269.

[2] Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T., Atkinson, P.M. 2020. COVID-19 Outbreak Prediction with Machine Learning http://dx.doi.org/10.2139/ssrn.3580188

[3] Ivanov, D. 2020. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp. Res. Part E Logist. 136, https://doi:10.1016/j.tre.2020.101922

[4] Vaishya, R., Javaid, M., Khan, I.H., Haleem, A. 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr. 14(4), 337–339.

[5] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L. 2020 Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. 296(2), 32-40.

[6] Pham, Q., Nguyen, D.C., Huynh-The, T., Hwang, W., Pathirana, P.N. 2020. Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts. IEEE Access. 8, 130820-130839.

[7] Luengo-Oroz, M., Pham, K.H., Bullock, J., Kirkpatrick, R., Luccioni, A., Rubel, S., Wachholz, C., Chakchouk, M., Biggs, P., Nguyen, T., Purnat, T., Mariano, B. 2020. Artificial intelligence cooperation to support the global response to COVID-19. Nature Machine Intelligence. 2, 295–297.

[8] Robson, B. 2020. COVID-19 Coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance. Comput.

751 Biol. Med. 121, 103749.
752 [9] Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Bonten,
753 M.M.J., Damen, J.A.A., Debray, T.P.A., De Vos, M., Dhiman, P., Haller, M.C., Harhay, M.O.,
754 Henckaerts, L., Kreuzberger, N., Lohman, A., Luijken, K., Ma, J., Andaur, C.L., Reitsma, J.B.,
755 Sergeant, J.C., Shi, C., Skoetz, N., Smits, L.J.M., Snell, K.I.E., Sperrin, M., Spijker, R.,
756 Steyerberg, E.W., Takada, T., van Kuijk, S.M.J., van Royen, F.S., Wallisch, C., Hooft, L.,
757 Moons, K.G.M., van Smeden, M. 2020. Prediction models for diagnosis and prognosis of covid-
758 19 infection: systematic review and critical appraisal. BMJ. 369.
759 [10] Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C. 2020 Data-based analysis,
760 modelling and forecasting of the COVID-19 outbreak. PLoS ONE. 15(3).
761 [11] Agrebi, S., Larbi, A. 2020. Use of artificial intelligence in infectious diseases. Artificial
762 Intelligence in Precision Health. 415-438.
763 [12] Koike, F., Morimoto, N. 2018. Supervised forecasting of the range expansion of novel
764 non-indigenous organisms: Alien pest organisms and the 2009 H1N1 flu pandemic. Global
765 Ecol. Biogeogr. 27, 991-1000.
766 [13] Agarwal, N., Koti, S.R., Saran, S., Kumar, S. 2018. Data mining techniques for predicting
767 dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Curr.
768 Sci. 114, 2281-2291.
769 [14] Raja, D.B., Mallol, R., Ting, C.Y., Kamaludin, F., Ahmad, R., Ismail, S., Jayaraj, V.J.,
770 Sundram, B.M. 2019. Artificial intelligence model as predictor for dengue outbreaks. Malays.
771 J. Public Health Med. 19, 103-108.
772 [15] Anno, S., Hara, T., Kai, H., Lee, M.A., Chang, Y., Oyoshi, K., Mizukami, Y., Tadono, T.
773 2019. Spatiotemporal dengue fever hotspots associated with climatic factors in taiwan including
774 outbreak predictions based on machine-learning. Geospatial Health. 14, 183-194.
775 [16] Raja, D.R., Mallol, R., Ting, C. Y., Kamaludin, F., Rohani Ahmad, Suzilah Ismail, Vivek

776 Jason Jayaraj, & Bala Murali Sundram. (2019). ARTIFICIAL INTELLIGENCE MODEL AS
777 PREDICTOR FOR DENGUE OUTBREAKS. Malaysian Journal of Public Health Medicine,
778 19(2), 103-108.
779 [17] Agarwal N, Koti SR, Saran S, Kumar AS. Data mining techniques for predicting dengue
780 outbreak in geospatial domain using weather parameters for New Delhi, India. Current Science
781 journal. 2018; 114: 2281-2291.
782 [18] Titus Muurlink, O., Stephenson, P., Islam, M.Z., Taylor-Robinson, A.W. 2018. Long-term
783 predictors of dengue outbreaks in Bangladesh: A data mining approach. Infectious Disease
784 Modelling. 3, 322-330.
785 [19] Chenar, S.S., Deng, Z. 2018. Development of artificial intelligence approach to forecasting
786 oyster norovirus outbreaks along Gulf of Mexico coast. Environment International. 111, 212-
787 223.
788 [20] Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao,
789 Z., Mai, Z. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in
790 China under public health interventions. Journal of Thoracic Disease.
791 [21] Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Rajendra Acharya, U.
792 2020. Automated detection of COVID-19 cases using deep neural networks with X-ray images.
793 Computer Biol Med. 121, 103792.
794 [22] Wang, L., Wong, A. 2020. COVID-Net: a tailored deep convolutional neural network
795 design for detection of COVID-19 Cases from chest radiography images.
796 [23] Hemdan, E.E.D., Shouman, M.A., Karar, M.E. 2020. COVIDX-Net: A Framework of
797 Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images.
798 https://arxiv.org/abs/2003.11055
799 [24] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng,
800 X., Xu, B. 2020. A deep learning algorithm using CT images to screen for Corona Virus Disease

(COVID-19).

[25] Khanday, A.D., Rabani, S.T., Khan, Q.R. 2020. Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology. 12, 731–739.

[26] Dutta, P., Paul, S., Kumar, A. 2021. Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19. Electronic Devices, Circuits, and Systems for Biomedical Applications, 521–540. https://doi.org/10.1016/B978-0-323-85172-5.00020-4

[27] Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., Peng, E., Huang, J., Zhang, Y., Xu, X., Xu, H., Zhou, F., Wang, G. 2020. Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. Front. Cell Dev. Biol. 8, 683. doi: 10.3389/fcell.2020.00683

[28] Tamal, M., Alshammari, M., Alabdullah, M., Hourani, R., Abu Alola, H., Hegazi, T. 2020. An Integrated Framework with Machine Learning and Radiomics for Accurate and Rapid Early Diagnosis of COVID-19 from Chest X-ray. Expert Systems with Applications, https://doi.org/10.1101/2020.10.01.20205146

[29] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., Cabitza, F. 2020. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: a Feasibility Study. Journal of Medical Systems, https://doi.org/10.1101/2020.04.22.20075143

[30] Huerta, E.B., Duval, B., Hao, J. K., 2010. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. Neurocomputing. 73 (13-15), 2375–2383.

[31] Rammal, A., Perrin, E., Vrabie, V., Assaf, R., Fenniri, H. 2017. Selection of discriminant mid-infrared wavenumbers by combining a naïve Bayesian classifier and a genetic algorithm: Application to the evaluation of lignocellulosic biomass biodegradation. Mathematical Biosciences. 289, 153-161.

[32] Punia, M., Joshi, P.K., Porwal, M.C. 2010. Decision tree classification of land use land

826 cover for Delhi, India using IRS-P6 AWiFS data. Expert Systems with Applications. 38 (5),
827 5577–5583.
828 [33] Safavian, S.R., Landgrebe, D. 1991. A survey of decision tree classifier methodology.
829 IEEE Transactions on Systems, Man, and Cybernetics. 21(3), 660-674.
830 [34] Liu, M., Wang, M., Wang, D. 2013. Comparison of random forest, support vector machine
831 and back propagation neural network for electronic tongue data classification: Application to
832 the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical.
833 177, 970-980
834 [35] Kotsiantis, S.B. 2007. Supervised Machine Learning: A Review of Classification
835 Informatica. 31, 249-268.
836 [36] Villuendas-Rey, Y., Yáñez-Márquez, C., Antón-Vargas, J., López-Yáñez, I. 2019. An
837 Extension of the Gamma Associative Classifier for Dealing with Hybrid Data. Access IEEE, 7,
838 64198-64205.
839 [37] Villuendas-Rey, Y., Rey-Benguría, C., Ferreira-Santiago, Á., Camacho-Nieto, O., Yáñez-
840 Márquez, C. 2017. The naïve associative classifier (NAC): A novel, simple, transparent, and
841 accurate classification model evaluated on financial data. Neurocomputing. 265, 105–115
842 [38] Kotsiantis, S.B., 2007. Supervised Machine Learning: A Review of Classification
843 Informatica. 31, 249-268.
844 [39] Rish I. 2001. An Empirical Study of the naive Bayes Classifier, Artificial Intelligence
845 Journal IJCAI. 3: 41-46.
846 [40] Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition. Academic Press,
847 second edition.
848 [41] Garber, F.D., Djouadi, A. 1988. Bounds on the Bayes classification error based on
849 pairwise risk functions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 10,
850 281–288.

851 [42] Park, C.H., Lee, M. 2008. On applying linear discriminant analysis for multi-labeled
852 problems. Pattern Recognition Letters. 29 (7), 878–887.
853 [43] Sharma, R., Ghosh, A., Joshi, P.K. 2013. Decision tree approach for classification of
854 remotely sensed satellite data using open-source support. Journal of Earth System Science, 122,
855 1237–1247.
856 [44] Bressan, G.M., De Azevedo, B.C., Ap, E., Lizzi, S. 2017. A Decision Tree Approach for
857 the Musical Genres Classification. Applied Mathematics & Information Sciences. 11 (6), 1703-
858 1713.
859 [45] Rokach, L., Maimon, O. 2008. Data mining with decision trees: Theory and applications,
860 World Scientific. 69, doi.org/10.1142/6604
861 [46] Tsang, S., Kao, B., Yip, K.Y., Ho, W.S., Lee, S.D. 2009. Decision trees for uncertain data.
862 IEEE 25th International Conference on Data Engineering, 441-444.
863 [47] Luts, J., Ojeda, F., Plas, R., Van De Moor, B., De Huffel, S. 2010. A tutorial on support
864 vector machine-based methods for classification problems in chemometrics, Analytica Chimica
865 Acta. 665, 129–145.
866 [48] Cortes, C., Vapnik, V. 1995. Support-vector network. Machine Learning. 20, 273–297,
867 doi.org/10.1007/BF00994018
868 [49] Jain, A.K., Mao, J., Mohuiddin, K. 1996. Artificial Neural Networks: A Tutorial. IEEE
869 Computer. 29(3), 31-44.
870 [50] Simon, H. 1999. Neural Networks and Learning Machines. Third Edition. McMaster
871 University Hamilton, Ontario, Canada. Rev. ed of: Neural networks. 2nd ed.
872 [51] Sildir, H., Aydin, E., Kavzoglu, T. 2020. Design of Feedforward Neural Networks in the
873 Classification of Hyperspectral Imagery Using Superstructural Optimization. Remote Sensing.
874 12, 956, doi:10.3390/rs12060956
875 [52] ItzamA, L.Y., Amadeo, J.A.C., Oscar, C.N., Cornelio Y.M. 2011. Pollutants Time-Series

876 Prediction using the Gamma Classifier, International Journal of Computational Intelligence
877 Systems. 4(4), 680-711, DOI: 10.1080/18756891.2011.9727822
878 [53] Rangel-Diaz-de-la-Vega, A., Villuendas-Rey, Y., Yanez-Marquez, C., Camacho-Nieto,
879 O., Lopez-Yanez, I. 2020. Impact of Imbalanced Datasets Preprocessing in the Performance of
880 Associative Classifiers. Applied Sciences. 10(8), 2779; doi:10.3390/app10082779
881 [54] Rangel-Diaz De La Vega, A., Villuendas-Rey, Y., Yanez-Marquez, C., Camacho-Nieto,
882 O. 2020. The Naïve Associative Classifier With Epsilon Disambiguation. IEEE Access. 8,
883 51862-51870.
884 [55] Villuendas-Rey, Y., Hernandez-Castano, J., Camacho-Nieto, O., Yanez-Marquez, C.,
885 Lopez-Yanez, I. 2019. NACOD: A Naïve Associative Classifier for Online Data. IEEE Access,
886 7, 117761-117767.
887 [56] Sun, Z., Zhang, H., Yang, Y., Wan, H., Wang, Y. 2020. Impacts of Geographic Factors
888 and Population Density on the COVID-19 Spreading under the Lockdown Policies of China.
889 Science of the Total Environment. 746(666).
890 [57] Dowd, J. B., Andriano, L., Brazel, D., Rotondi, V., Block, P., Ding, X., Liu, Y., Mills, M.
891 2020. Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-
892 19. Proceedings of the National Academy of Sciences of the United States of America 117(18),
893 9696–98
894 [58] Emanuel, E.J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C.,
895 Boyle, C., Smith, M., Phillips, J.P. 2020. Fair Allocation of Scarce Medical Resources in the
896 Time of Covid-19. N Engl J Med. 382(21), 2049-2055. doi: 10.1056/NEJMsb2005114.
897 [59] Lai, C.C., Wang, C.Y., Wang, Y.H., Hsueh, S.C., Ko, W.C., Hsueh, P.R. 2020. Global
898 epidemiology of coronavirus disease 2019 (COVID-19): disease incidence, daily cumulative
899 index, mortality, and their association with country healthcare resources and economic status.
900 Int J Antimicrob Agents. 55(4), 105946. doi: 10.1016/j.ijantimicag.2020.105946.

[60] Safavi, P.E., Rahimian, K., Doustmohammadi, A., Safari dastjerdei, M., Rasouli, A., Zahiri, J. 2021. A prediction model for COVID-19 prevalence based on demographic and healthcare parameters in Iran. https://doi.org/10.1101/2021.01.27.21250551

[61] Payam, S., Bordbar, N., Ghanbarzadegan, A., Najibi, M., Bastani. P. 2020. Ranking of Iranian Provinces Based on Healthcare Infrastructures: Before and after Implementation of Health Transformation Plan. Cost Effectiveness and Resource Allocation 18(1).

[63] Nie, P., Ding, L., Chen, Z. 2021. Income-related health inequality among Chinese adults during the COVID-19 pandemic: evidence based on an online survey. Int J Equity Health 20, 106. https://doi.org/10.1186/s12939-021-01448-9

[63] Buja, A., Paganini, M., Cocchio, S., Scioni, M., Rebba, V., Baldo, V. 2020. Demographic and socio-economic factors, and healthcare resource indicators associated with the rapid spread of COVID-19 in Northern Italy: An ecological study. PLoS ONE. 15(12), https://doi.org/10.1371/journal.pone.0244535

[64] National Institute of Statistics (ISTAT). Main database. [cited 2020 April 13]. Available from: http://dati.istat.it/Index.aspx

[65] National Institute of Statistics (ISTAT). Health for All database. [cited 2020 April 13]. Available from: https://www.istat.it/it/archivio/14562.

[66] Bizri, A.R., Khachfe, H.H., Fares, M.Y., Musharrafieh, U. 2021. COVID-19 Pandemic: An Insult Over Injury for Lebanon. J Community Health. 46(3), 487-493.

[67] Sydow, J., Windeler, A., Muller-Seitz, G. 2012. Path Constitution Analysis: A Methodology for Understanding Path Dependence and Path Creation. Bus Res, 5, 155–176.

[68] Rammal, A., Perrin, E., Vrabie, V., Bertrand, I., Chabbert, B. 2017. Classification of lignocellulosic biomass by weighted-covariance factor fuzzy C-means clustering of mid-infrared and near-infrared spectra, Journal of Chemometrics. 31(2), 1–10.

[69] Everson, R., Fieldsend, J. 2006. Multi-class ROC analysis from a multi-objective

optimization perspective. Pattern Recognition Letters 27(8), 918-927.

Supplementary Information

Table 3: Correlation between the degrees of possibility and the predictors

Degrees of possibility		
Pearson Correlation		**Sig (2-tailed)**
Technology used	**-0.052**	**0.442**
Special event	**0.061**	**0.362**
Procedure	**0.027**	**0.69**
Experience	**-0.363**	**000**
Density	**-0.22**	**0.001**
Family Nb	**0.152**	**0.022**
Temperature	**-0.009**	**0.893**
Lockdown	**-0.232**	**000**
Medical Resources	**-0.224**	**0.001**
Economy Resources	**-0.081**	**0.23**