Health and Technology (2022) 12:867–877
https://doi.org/10.1007/s12553-022-00668-5

ORIGINAL PAPER

An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic

E. Camargo[4] · J. Aguilar[1,2,5] · Y. Quintero[1] · F. Rivas[3,2] · D. Ardila[1]

Received: 21 January 2022 / Accepted: 5 April 2022 / Published online: 25 April 2022

Abstract
Several works have proposed predictive models of the SEIRD (Susceptible, Exposed, Infected, Recovered, and Dead) variables to characterize the pandemic of COVID-19. One of the challenges of these models is to be able to follow the dynamics of the disease to make more precise predictions. In this paper, we propose an approach based on incremental learning to build predictive models of the SEIRD variables for the COVID-19 pandemic. Our incremental learning approach is a dynamic ensemble method based on a bagging scheme that allows the addition of new models or the updating of incremental models. The article proposes an incremental learning architecture composed of two components. The first component carries out an analysis of the interdependencies of the SEIRD variables and the second component is an incremental learning model that builds/updates the predictive models. The paper analyses the quality of the predictive models of our incremental learning approach using data of the COVID-19 from Colombia, and shows interesting results about the predictions of the SEIRD variables. These results are compared with an incremental learning approach based on random forests.

Keywords Machine learning · COVID-19 · Prediction model

1 Introduction

The well-known and most used mathematical models to study the spread of infections/epidemics are the classical ordinary differential equations, such as the SIR (Susceptible, Infectious, Recovered) and SEIRD (Susceptible, Exposed, Infectious, Recovered, and Dead) models, which are widely used in epidemiology. Since the outbreak of COVID-19, the SIR model and its variations have increased their popularity. For the COVID-19, these models have been proposed to study its dynamic in different works [1–3]. For example, Zhong et al. [4] have proposed a SIR model for predicting the COVID-19 using China's first reported data. Also, Yang and Wang [5] presented an extended SEIR model for COVID-19 with time-varying transmission rates by considering the environmental effects. On the other hand, Zhu et al. [6] present a stochastic-based method for modeling and analysis of COVID-19 spread using a SEIR-Re-infected and Deceased-based Social Distancing model, called SEIR(R) D-SD, to consider the re-infection rate and social distancing factor into the traditional SEIRD model. In particular, the SEIRD model is the most adopted to characterize the COVID-19 pandemic because, from it, the effectiveness of various measures to attack the outbreak can be evaluated, which seems to be a difficult task for general statistical methods.

Other works have considered alternative approaches based on machine learning techniques to exploit the large amount of data that has been collected on COVID-19, with the aim of developing predictive models of the behavior of the SEIRD variables [7, 8]. For example, Quintero et al. [9] have proposed predictive models using classical machine learning techniques for the SEIRD variables based on historical data collected about them, and the contextual variables where the model was applied. Particularly, the contextual

This article is part of the COVID-19 Health Technology: Design, Regulation, Management, Assessment

✉ J. Aguilar
aguilar@ula.ve; aguilarjos@gmail.com

1 GIDITIC, Universidad EAFIT, Medellín, Colombia

2 CEMISID, Universidad de Los Andes, Mérida, Venezuela

3 Universidad Técnica Federico Santa María Valparaiso, Valparaíso, Chile

4 CobuildLab, Miami, EEUU, USA

5 Departamento de Automática, Universidad de Alcalá, Alcalá de Henares, Spain

variables considered in this work are: total population, the number of people over 65, poverty index, morbidity rates, average age and population density. For the construction of the SEIRD predictive models, in Quintero et al. [9], an analysis of the dependencies of these variables among themselves, and with the context variables, was carried out. A first work that uses deep learning related to COVID-19 is presented by Pelaez and Loayza [10]. They propose the use of a convolutional neural network for helping in the diagnosis of COVID-19 using chest X-ray images, giving as output the probability of identifying pneumonia caused by this virus. Finally, Chen et al. [11] develop a data-driven workflow to extract, process, and develop models of the COVID-19 epidemic using deep learning methods.

However, in the context of the pandemic, a relevant issue for data-based models is that they must adjust to the data that is being captured in real-time. In this sense, new online learning approaches are required in order to keep updating the models. Approaches such as incremental learning, online learning, active learning, are necessary. In this work, we focus on incremental learning techniques [12]. Incremental learning is a machine learning method where the current input data is continuously used to update the knowledge model. The aim of incremental learning is to adapt the model to new data without forgetting its existing knowledge. Many traditional machine learning algorithms inherently support incremental learning. Other algorithms can be adapted to support incremental learning. Examples of incremental algorithms include [12–14]: decision trees, decision rules, some artificial neural networks models (ex. Fuzzy ART and TopoART) or the incremental Support Vector Machine (SVM).

In the case of incremental learning works for the health area, an approach in medical diagnosis is proposed by Casalino, et al. [15], which consists in a Dynamic Incremental Semi-supervised Fuzzy Clustering to detect Bipolar Disorder Episodes. In the same direction of medical diagnosis using incremental learning, Braccioni, et al. [16] used a Forest-Tree Machine Learning approach for lung transplant recipients to study symptoms such as dyspnea, muscle effort and muscle pain, and their relationship with cardiac and pulmonary function parameters during an incremental exercise testing. Also, Schiel et al.[17] defined an approach for improving the prediction in electromyography signals for controlling prosthesis using incremental learning in a sparse Gaussian Process. Wang et al. [18] presented a proposal for the multiple percept detection problems in image sequences, in order to include new data from not previously identified categories, using the fast approximate nearest neighbor search tree-based filter using incremental learning. Another approach for using incremental learning in deep learning is presented by Su et al. [19]. They have proposed a framework for an incremental learning approach with a deep convolutional neural network in an anthropomorphic robot manipulator. Finally, Neto et al. [20] proposed an incremental LSTM model to tackle evolving data stream problems. Now, to the best of our knowledge, there are no works on incremental learning approaches applied in the field of prediction for COVID-19.

In this paper, we propose an incremental learning approach for prediction models based on a dynamic ensemble method that uses a bagging scheme that allows the addition of new models or the updating of incremental learning models. In this way, when the degradation of the ensemble method prediction is detected, then our approach prepares a training dataset with the current data to build new predictive models or update the incremental learning models. The new models are built with different machine learning techniques, and are considered by the ensemble method using a bagging scheme.

Specifically, this article defines an incremental learning architecture that has two components. A first component carries out an analysis of the interdependencies of the SEIRD variables similar to the work [9], and a second component is the incremental learning model that builds/updates the predictive models of the SEIRD variables for the COVID-19. Additionally, this work compares the prediction results of our approach with an incremental learning approach based on random forests. The remainder of the paper is organized as follows. Section 2 defines incremental learning and Sect. 3 describes our incremental learning architecture. Section 4 presents the experiments and results. Finally, Sect. 5 contains the conclusions.

2 Incremental learning

For a long time in the history of machine learning, there has been an implicit assumption that a “good” training set in a domain is available a priori. The training set then is “good” if it contains all necessary knowledge that once learned, can be reliably applied to any new examples in the domain. Consequently, emphasis is put on learning as much as possible from a fixed training set. Unfortunately, many real-world applications cannot match this ideal case, such as in dynamic control systems, web mining, and time series analysis, where the training examples are often fed to the learning algorithms over time, i.e., the learning process is continuous.

So, the incremental learning approaches can handle these problems and yield adjustments in the obtained models that can include newly collected data. The main idea of incremental learning approaches is to create models that can be dynamically adjusted without losing previously found knowledge. It can include new data sets and also new classes or other information that can be incorporated into

the models. This is a very good option for big data systems because it can include the data as time goes, without needing to stop and spend time creating new models. The three main conditions for using incremental learning are:

- The complete data set is not available at the moment of creating the model.
- The learning approach should be able to incorporate the new data without losing the knowledge previously obtained.
- It must find a balance between stability (the ability to not forget the already acquired knowledge) and plasticity (the ability to adjust the new data presented).

According to Chefrour [12], the diverse incremental learning algorithms can be classified into two classes:

- Adaptive systems: In this case, the structure of the system remains fixed, and the parameters are adjusted according to the new data received.
- Evolving systems: In this case, the structure can be modified by including new classes or behavior according to system changes over time.

There are some problems in incremental learning, Gepperth and Hammer [21] consider that the most important challenges are the online-adaptation of the model parameters, the concept drift (changes in data statistics that occur over time), the stability-plasticity dilemma, among others.

In general, there are three different kinds of incremental tasks:

- *Example-Incremental Learning Tasks (E-IL Tasks)*
- New training examples are provided after a learning system is trained. For example, a face recognition system can gradually improve its accuracy by incorporating new face images of registered users as they use it, without reconfiguring and/or retraining the entire system.
- *Class-Incremental Learning Tasks (C-IL Tasks)*: New output classes are provided after training a learning system. For example, if a new user is added to the group of registered users in the facial recognition system mentioned above, the system should be able to recognize the new user without reconfiguring or retraining the entire system.
- *Attribute-Incremental Learning Tasks (A-IL Tasks)*: New input attributes are provided after training a learning system. For example, if the camera used in the aforementioned facial recognition system is changed from a grayscale camera to a color camera, the system should be able to use the additional color features without reconfiguring or retraining the entire system.

A suitable incremental learning algorithm must be carefully designed. In some applications, incremental learning algorithms should automatically distinguish previous cases, and take actions to improve the representation of the current concepts as more examples become available, or respond to changes in the definition of the concepts when incoming examples become inconsistent with the learned concepts.

Schlimmer and Granger [22] proposed three criteria to measure the usefulness and effectiveness of an incremental learning method: (1) the number of observations (examples) needed to obtain a 'stable' concept description, (2) the cost of updating memory, and (3) the quality of learned concept descriptions. Some years later, [13] claimed that in order to measure an incremental learning algorithm, two new questions need to be answered: (1) How much better is a learned model at step $n+i$ than another model obtained before step n? (2) Can an incremental algorithm recover in the next incremental step(s), if it goes drastically off the "actual" concept at any stage? Consequently, they proposed three criteria for the evaluation of the robustness of incremental learning algorithms: (1) *Stability* – the prediction accuracy on the test set should not vary wildly at every incremental learning step; (2) *Improvement* – there should be an improvement in the prediction accuracy as the training progresses and the learning algorithm sees more training examples; and (3) *Recoverability* – the learning method should be able to recover from its errors, i.e., if the performance drops at a certain learning step, then the algorithm should be able to recover to the previous best performance.

Another frequently used criterion is the learning curve. An incremental algorithm may start learning from scratch and gain knowledge gradually with an increasing number of training examples. Consequently, the quality of the learned model shows a curve of gradual improvement over time, which is called the learning curve. Usually, the learned model is not very reliable at the early stage of the curve. Decisions can be made according to the learning curve on how valuable the output of the incremental learner might be at a certain stage. However, in practice, it is often difficult to determine the point at which the model has learned "enough" to be reliable. Generally, a typical "good" learning curve should increase rapidly to a relatively steady high level.

3 Our incremental learning approach

In this work, we propose an incremental learning architecture whose main objective is to predict the future behavior of SEIRD variables in a time window. This architecture allows updating the predictive models of the SEIRD variables with the most recent information (determined by the time window) on the behavior of COVID-19. The architecture consists of

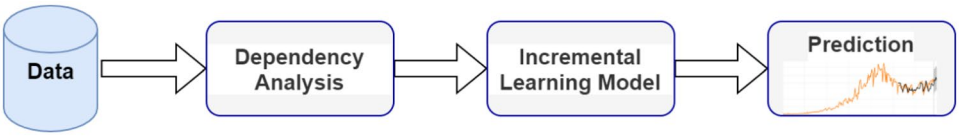

Fig. 1 Incremental Learning Architecture

two components (see Fig. 1): A dependency analysis process, which defines the variables that will be considered by the predictive models, and the incremental learning model that predicts the SEIRD variables.

Specifically, the first component performs a feature engineering process to select the variables to build the predictive models. In particular, this process consists of a dependency analysis to find the relationships between the SEIRD variables and the predictor variables, that is, to find the variables that are relevant for the construction of the predictive models. The second component builds the predictive models using our incremental learning model. The incremental learning process invokes the dependency analysis process from time to time to update the relationships between the SEIRD variables and the predictor variables.

Thus, this architecture that considers two vital processes for the construction of accurate predictive models, one of feature engineering and the other of incremental learning, is the first contribution of this article.

3.1 Variable dependence analysis for the SIERD model for incremental learning

The variables considered important as predictor variables for the incremental learning model were obtained through a time series analysis that considers other variables as predictors. In this work, a time series analysis is carried out to detect the relationships between the variables due to the data set has a chronological order. The model can be seen in Eq. 1.

$$Y_t = \beta_0 + \gamma_0 X_t + \gamma_1 X_{t-1} + \cdots + \gamma_k Z_t + \gamma_{k+1} Z_{t-1} + \cdots + \gamma_l W_t + \gamma_{l+1} W_{t-1} + \cdots + \gamma_m U_t + \gamma_{m+1} U_{t-1} + \cdots + \eta_t \quad (1)$$

where: β_o and γ_i are real constants, and η_I is an ARIMA process.

The models were adjusted with the 5 variables of the SEIRD model as predictor variables, to which the delayed effects of up to 7 days were considered. In this way, for the 5 variables is analyzed the dependence with the other variables and themselves over time. To evaluate the method, the dataset is divided into training and testing, and to evaluate the predictions made is used the Mean Absolute Percentage Error (MAPE) of the predictions of the dataset (see Eq. 2).

$$\text{MAPE} = \frac{1}{N} \sum_{j=n+1}^{n+N} \frac{\left| Y_i - \hat{Y}_i \right|}{\left| Y_i \right|} \quad (2)$$

where: Y_i is the real value of the variable in time i, $\hat{Y}_i$ is the predicted value of the variable in time i, n is the total amount of data in the training set, and N is the total amount of data in the test set.

To determine the best predictors, a genetic algorithm (GA) was implemented, which searches for the best descriptors for each target variable. Figure 2 shows the process followed by the method.

Figure 2 describes the general procedure of detection/selection of the best descriptors that are highly associated with the SEIRD variables, using the GA with ARIMA time series models. The inputs of the GA are: the variable of interest (SEIRD) and the set of predictors (SEIRD lagging). Within the GA there is a selection and variation process in which different subsets of predictors are obtained. With each combination of predictors, several ARIMA models are adjusted, varying the parameters (p, d, q), and the best model is chosen according to the MAPE metric (see Eq. 2). At

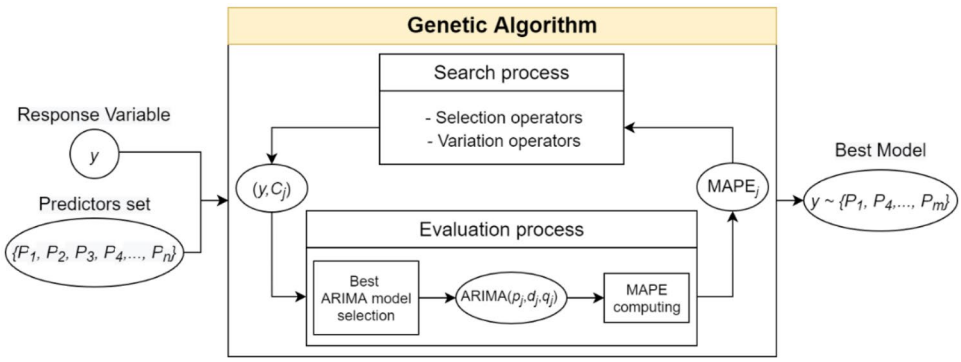

Fig. 2 General procedure for selection/extraction of the best features

the end, the GA provides the best adjusted ARIMA model and the subset of predictors. Thus, GA determines the best combination of lagging variables as descriptors, optimizing MAPE.

3.2 Incremental machine learning model

We propose an incremental learning approach based on the Ensemble Learning paradigm. Our goal is to create models that constantly learn from COVID-19 data, to generate more accurate predictions of the SEIRD model of the COVID-19. The general process is shown in Fig. 3. In general, our approach uses a bagging scheme like the ensemble method [23, 24]. It initially trains a set of predictive models, which use different machine learning techniques (gradient boosting, random forest, etc.) and are trained using a random subset of the data. These subsets of data (random samples) are defined using a 0.632 Bootstrap strategy. During the operation, our approach, in the first stage, makes online predictions using the predictive models that have been previously built. Then, the MAPE of each model is calculated and the model with the lowest MAPE is selected. If the prediction of the best predictive model is very bad, then it builds new prediction models (maybe with new machine learning techniques) and/or updates the current incremental learning models. For that, it builds a new dataset with the most recent data, and using a 0.632 Bootstrap strategy, defines the training data to update the incremental learning models and/or create new predictive models. The newly trained models are included in the bagging scheme to become part of the set of predictive models. Thus, our approach keeps its predictive capability.

Our approach uses powerful machine learning techniques for building predictive models, like Gradient boosting, Random Forest, Bootstrap Aggregation, linear regression, and the Limited-memory BFGS (L-BFGS or LM-BFGS) neural network [23, 25]. Also, our approach proposes a dynamic ensemble method that adds new predictive models but also, can update the incremental learning models. In this way, when the degradation of the ensemble method prediction is detected, then our approach prepares a training dataset with the current data to build new predictive models or update the incremental learning models. Our approach has two parameters: the threshold beyond which the predictions are considered not good,and the backward time interval that is used to calculate the prediction error (deferred prediction evaluation). Both must be optimized according to the dataset used.

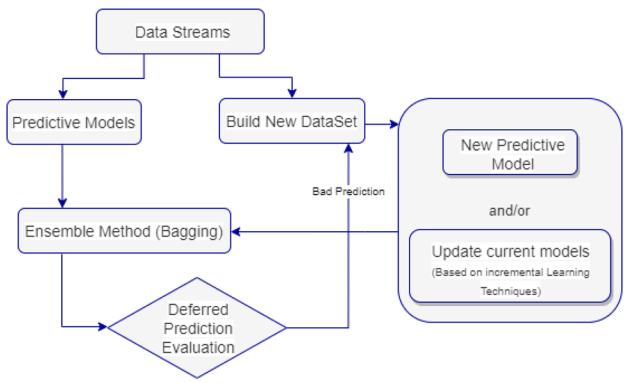

Fig. 3 Our incremental learning approach

4 Experimentation

4.1 Experimental context

For the development of this work, the dataset used to predict the SEIRD variables is from Colombia, and it contains the daily information of these variables for the COVID-19. This dataset can be obtained from the official website of the National Institute of Health of Colombia (INS). In the INS dataset can be found the next variables: *Date*: timestamp; *Exposed*: number of people who have been exposed, which is estimated with the number of tests performed; *Infected*: number of people infected; *Recovered*: number of people who have recovered; and *Dead*: number of people who have died. The *Susceptible* variable is calculated as the total number of inhabitants in Colombia minus the variables exposed, recovered, infected and deaths. For the variables exposed, infected, recovered and dead, a summary is presented in Table 1, which shows the average, minimum (min), maximum (max) and standard deviation (sd) of each of these variables. The susceptible variable is not in Table 1, since it is calculated as the difference between the total number of inhabitants of the country and the other variables.

The dataset contains information for 300 days for each of the variables mentioned above. For the susceptible variable, there is no meaning in calculating these statistics. Table 1 shows that the total number of infected persons per day is between 5 and 13.056, with an average of 3.929 infected persons per day. On the other hand, there were days in which there were no deaths from COVID-19, and sometimes a maximum of 438 deaths, daily on average there were 135 deaths, while on average the daily total of people recovered was 3.284, and on some days this figure rose to 12.295

Table 1 Variables statistics

	Exposed	Infected	Recovered	Dead
Average	16.097	3.929	3.284	135
min	258	5	5	0
max	41.434	13.056	12.295	438
sd	12.328	4.010	3.670	132

people recovered. In addition, an average of 16.097 people were exposed to the virus daily, reaching a maximum of 41.434.

For this study, the target variables were Susceptible, Exposed, Infected, Recovered and Dead. The predictive models of the target variables were built with the Gradient Boosting Regressor, with the following hyperparameters: the loss function to be optimized was least squares regression, n_estimators = 500, learning_rate = 0.01, criterion = "MSE" and test_size = 0.2; Random Forest with the following hyperparameters:n_estimators = 500, criterion = "MSE", max_depth = 100, min_samples_leaf = 2 and test_size = 0.2; Linear regression with the following hyperparameters:test_size = 0.2 and random_state = 10;and L-BFGS with the following hyperparameters: Relu was used as the activation function, the optimization function was Adam and the loss function was the MSE,hidden_layer_sizes = (25), solver = 'lbfgs' and alpha = 1e-5.

All the experimentation was done with data that are between March and December of 2020. For the analysis of dependencies, the MAPE was used as the metric to evaluate the predictive quality of the model. The quality metrics used to evaluate each model were MAPE, Mean Square Error (MSE), and coefficient of determination (R^2).

4.2 Experimental cases

4.2.1 Variable dependence analysis

Two experiments were carried out for the analysis of dependencies, one for August and the other for September. With the development of this study, it was not only determined the best model with the set of variables that compose it, but it was also possible to find the time window to predict the future behavior. Table 2 shows the average MAPE for the 4 weeks of August for the intervals of 1, 2, 3 and 4 days. The reason for using a 4-day interval to make predictions is that up to this number of days is where we have obtained the best performance of the predictive models for the interval of delayed of 7 days analyzed.

The MAPE values presented in Table 2 are the mean of the MAPE values calculated for each week in the prediction intervals (1, 2, 3 and 4 days). According to the results of the susceptible and dead variables, the best prediction is obtained on days 1 and 4, i.e., on average on days 1 and 4 the predictions of these variables are closer to reality. The infected and recovered variables are closer to reality, on average, on days 1 and 2, while the exposed variable is closer to reality on days 1 and 3.

Table 2 Average MAPE predictions for August 2020

Variables	1 day (%)	2 day (%)	3 day (%)	4 day (%)
Susceptible	0.025	0.038	0.097	0.009
Exposed	5.484	7.775	7.703	7.931
Infected	6.816	6.529	7.162	8.738
Recovered	6.306	6.891	7.714	7.956
Deaths	11.651	17.566	18.348	16.565

Table 3 Average MAPE predictions for September 2020

Variables	1 day (%)	2 day (%)	3 day (%)	4 day (%)
Susceptible	0.763	1.046	0.726	1.020
Exposed	6.932	6.766	6.546	7.077
Infected	6.613	7.700	7.478	9.128
Recovered	8.005	8.146	8.554	7.585
Dead	13.619	12.925	14.210	13.469

The average MAPE for the 4 weeks of September is shown below, at intervals of 1, 2, 3 and 4 days (see Table 3).

In the second case, the quality of the models and the time window for September 2020 are shown in Table 3. As for August 2020, the values presented are the mean of the MAPE values calculated for the 4 weeks in the prediction intervals (1, 2, 3 and 4 days). In this case, 3 of the 5 predicted variables are closer to reality on day 3 (susceptible, exposed and infected); however, susceptible and infected are also closer to reality on day 1, while exposed on day 2. On the other hand, the recovered and dead variables have in common that on day 4 the predictions made for them are closer to reality, and the second-best predicted day for them are 1 and 2, respectively.

Tables 2, 3 show the average MAPE values for the 4 weeks of each month, for 4 days (time window), since predictions beyond these days are very far from reality (MAPE values higher than 50%), and result in unreliable values in the predictions of the machine learning models. Now, the prediction of the death variable is not so good, regardless of the time window.

Considering the dependency analysis performed, the variables that help to better predict the behavior of the SEIRD variables are given in Table 4, and are the ones considered to build the machine learning models.

Table 4 Target and Predictor variables for machine learning models

Target	Predictor Variables
Susceptible	exposed(t-5), exposed(t-6), infected(t-5), recovered(t-5), recovered(t-7)
Exposed	susceptible(t-6), infected(t-7), deaths(t-5), dead(t-7)
Infected	exposed(t-6), exposed(t-7), deaths(t-5), dead(t-6)
Recovered	susceptible(t-6), susceptible(t-7), exposed(t-5), exposed(t-6), exposed(t-7), infected(t-5), infected(t-7), dead(t-5)
Dead	susceptible(t-5), exposed(t-5), exposed(t-6), infected(t-5), recovered(t-5), recovered(t-7)

Table 5 Quality of the models used to predict learning incremental the SEIRD variables for Colombia

Target Variable	**Machine Learning Techniques in our incremental approach**	$t-1$			$t-2$			$t-3$		
		$\mathbf{R^2}$	**MSE**	**MAPE**	$\mathbf{R^2}$	**MSE**	**MAPE**	$\mathbf{R^2}$	**MSE**	**MAPE**
S	Gradient boosting	0.990	0.001	71.554	0.999	3.1e-05	70.271	0.993	0.001	70.770
	Random forest	0.922	0.003	71.581	0.822	0.009	71.415	0.892	0.008	72.180
	Linear regressor	0.998	0.001	76.664	0.996	0.001	76.761	0.911	0.002	75.890
	LM-BFGS	0.933	0.001	77.105	0.994	0.001	77.054	0.922	0.001	78.200
E	Gradient boosting	0.788	0.006	77.424	0.481	0.015	70.271	0.651	0.021	76.770
	Random forest	0.842	0.006	71.786	0.987	0.001	71.502	0.887	0.07	70.118
	Linear regressor	0.788	0.006	76.064	0.899	0.003	76.681	0.809	0.003	75.889
	LM-BFGS	0.950	0.005	77.905	0.846	0.006	77.054	0.801	0.011	78.020
I	Gradient boosting	0.872	0.004	74.564	0.990	0.001	70.071	0.810	0.002	70.077
	Random forest	0.858	0.005	75.681	0.443	0.010	71.102	0.343	0.021	71.618
	Linear regressor	0.860	0.005	76.664	0.824	0.004	76.781	0.840	0.004	75.989
	LM-BFGS	0.833	0.006	77.803	0.863	0.004	77.154	0.791	0.007	78.120
R	Gradient boosting	0.979	0.001	71.554	0.995	0.001	70.971	0.845	0.013	70.177
	Random forest	0.955	0.005	70.923	0.229	0.030	71.102	0.246	0.031	71.118
	Linear regressor	0.884	0.006	76.664	0.980	0.001	76.782	0.980	0.001	75.189
	LM-BFGS	0.938	0.002	78.302	0.991	0.001	77.154	0.981	0.004	78.120
D	Gradient boosting	0.919	0.007	70.432	0.968	0.001	70.871	0.955	0.003	70.177
	Random forest	0.933	0.008	70.581	0.188	0.036	71.502	0.231	0.030	71.118
	Linear regressor	0.829	0.009	75.342	0.959	0.002	76.681	0.940	0.003	75.189
	LM-BFGS	0.863	0.006	76.903	0.935	0.002	77.054	0.912	0.003	78.120

4.2.2 Prediction models

Table 5 shows the performance of our approach predicting the SEIRD variables based on the analysis of the time dependence that determined a 4-day predictive interval, where each SEIRD variable has the dependence defined in Table 4. Table 5 presents an example of the performance of each technique used by our incremental learning approach for the SEIRD variables

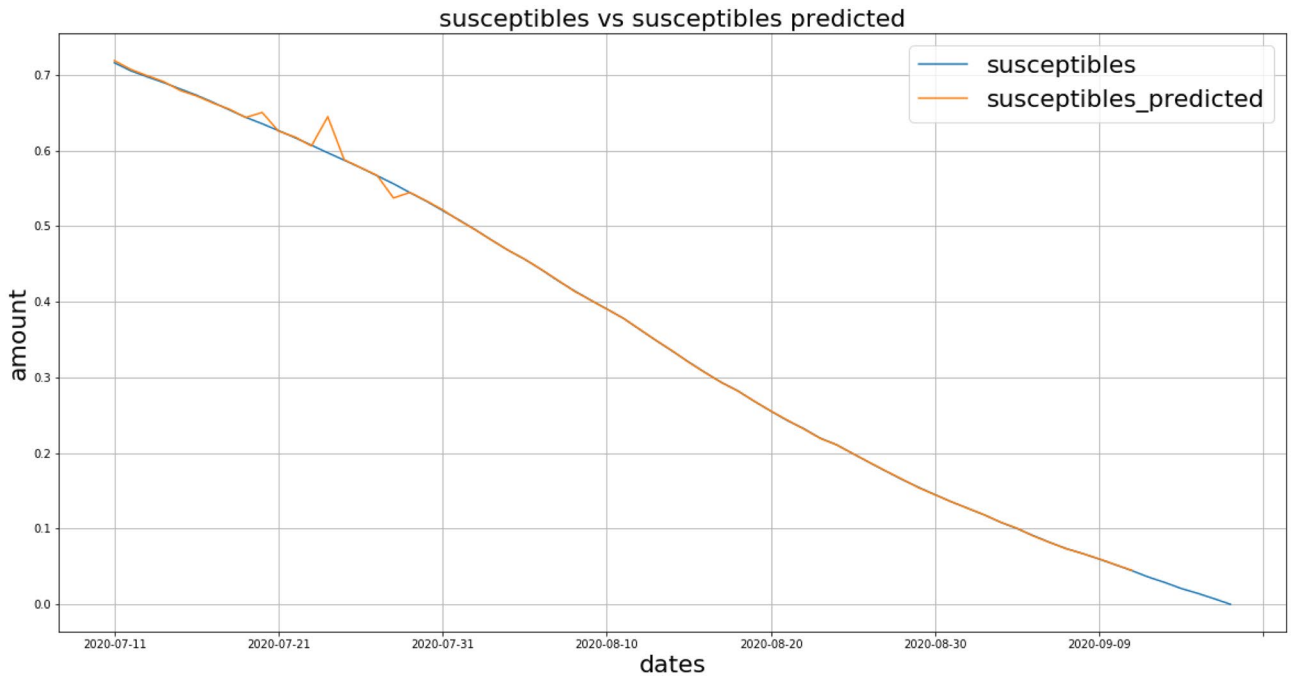

Fig. 4 Prediction of the Susceptible variable with our Incremental learning approach

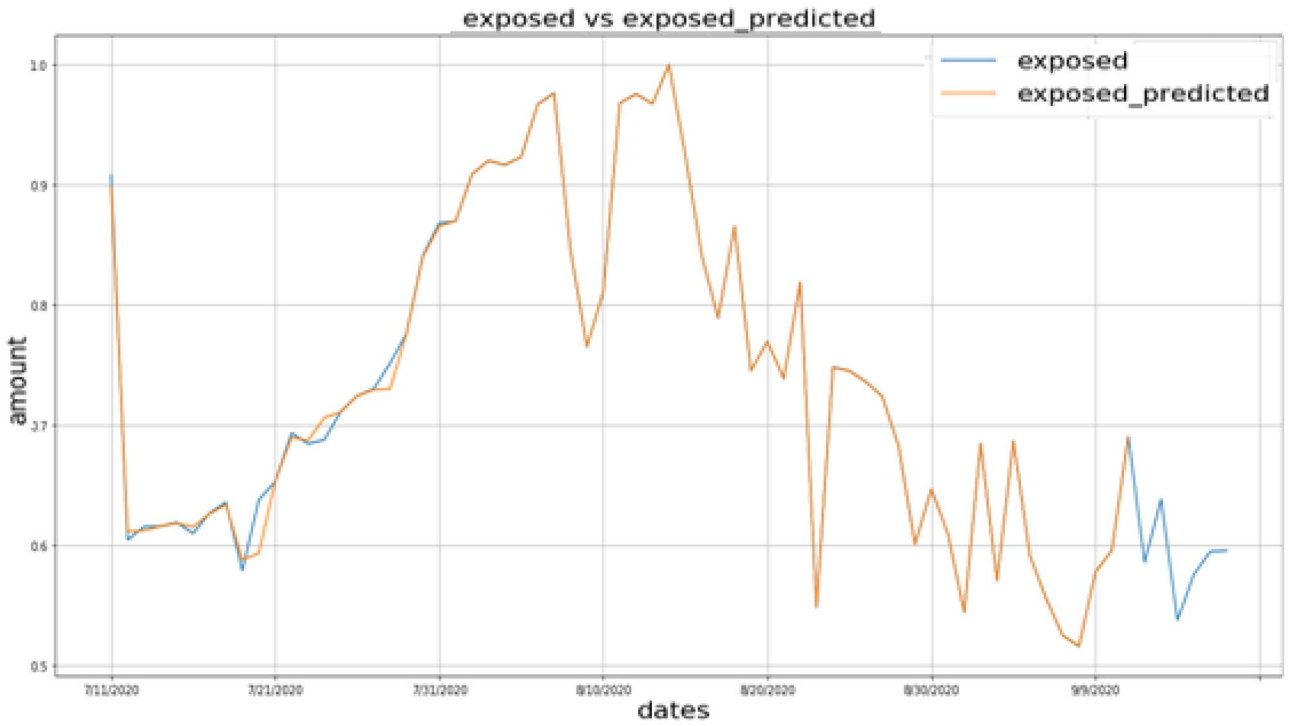

Fig. 5 Prediction of the Exposed variable with our Incremental learning approach

for three different moments ($t-i, i=1,2,3$), using MAPE, MSE and R^2 metrics. $t-i$ is the backward time interval (deferred prediction evaluation), which was determined as 7-day after a hyperparameter optimization process (we have used a Bayesian hyperparameter optimization approach). Also, during the hyperparameter optimization process was determined the threshold values equal to 0.9. Finally, we have used an incremental learning version of random forest.

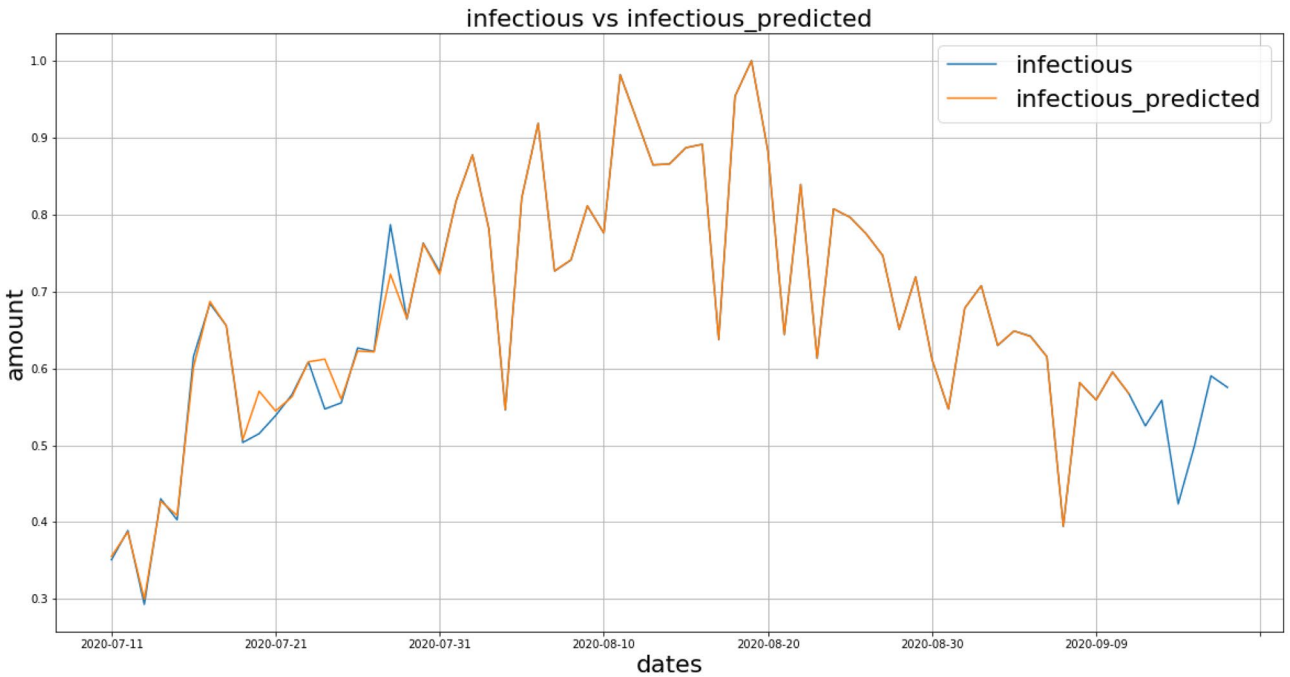

Fig. 6 Prediction of the Infectious variable with our Incremental learning approach

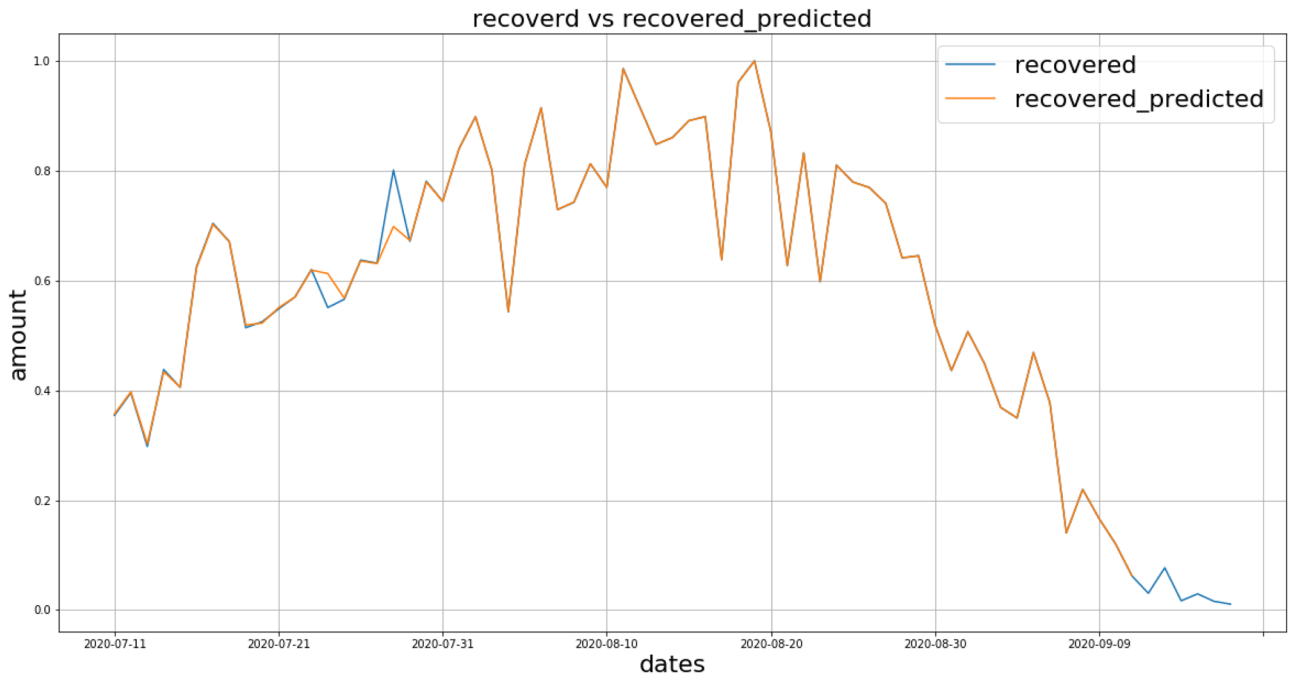

Fig. 7 Prediction of the Recovered variable with our Incremental learning approach

In general, our approach has created a new predictive model for E and I variables because the initial predictive models are not good (see Table 5, column.

$t-2$), R^2s are smaller than 0.9). Also, we can see that the best predictive model of our approach can change over time. For example, for the variable S, Gradient boosting is the best at the beginning ($R^2 = 0.993$, MAPE = 71.554 and MSE = 0.0001), and after, the linear model is the best. Finally, the only incremental learning technique implemented, besides our approach, was random forest, and sometimes it gives very bad results after retraining. For example, for the variable S, but it improves for the variables I, R and

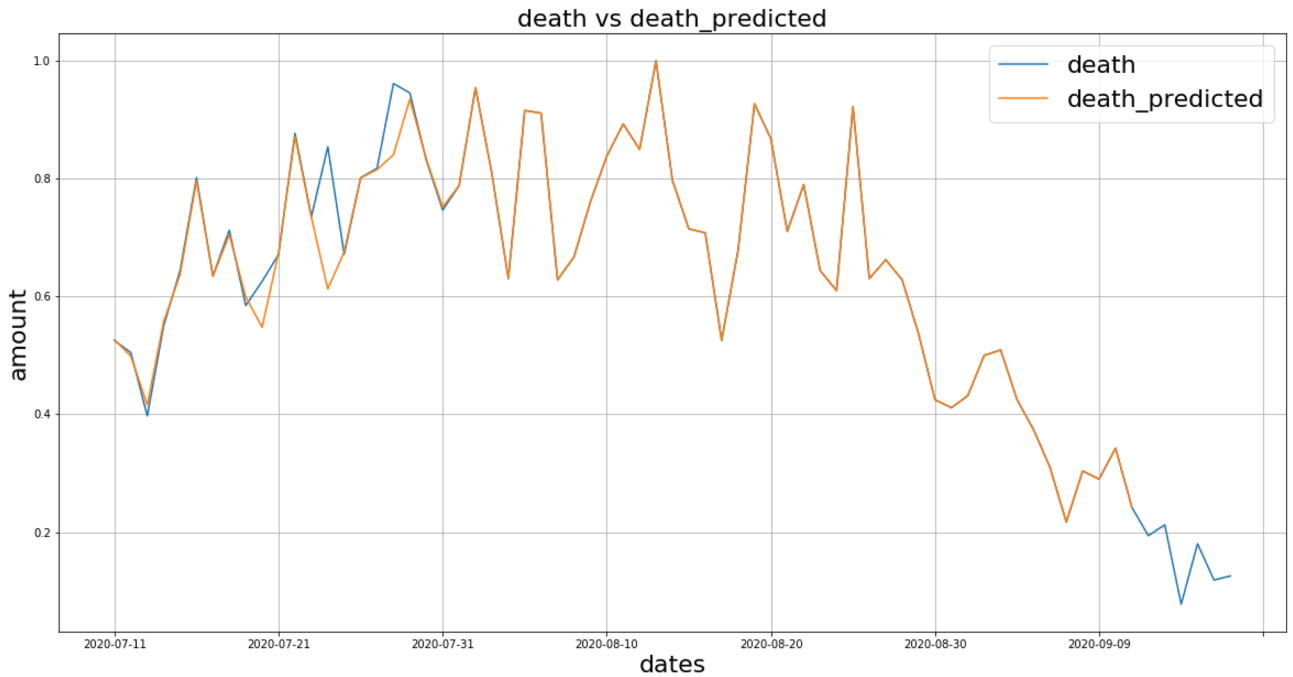

Fig. 8 Prediction of the Death variable with our Incremental learning approach

D. For the other techniques, in each step $t - i$, our approach adds randomly new predictive models based on them to the set of predictive models, trained with a new dataset using a 0.632 Bootstrap strategy.

Figures 4, 5, 6, 7, 8 show the prediction of our incremental learning model (remember that it selects the best prediction as output) of the SEIRD variables for the interval t-4. In these Figures, the predictions of the SEIRD variables are made with a 95% confidence interval, and the predicted values by our incremental learning approach are the blue line. Based on these results, each variable has a low error predicting its value. Thus, our incremental learning approach follows all variables quite well. In some abrupt behaviors, it has problems to predict them (but very slight, see Figs. 6, 7, 8), while when it is very homogeneous the behavior, then it can follow it perfectly (see susceptible variable).

Finally, we consider three metrics of the incremental learning area (see section II.2) to assess the quality of our approach to incremental learning: Stability, Improvement and Recoverability. With respect to the *Stability* metric, our approach keeps the prediction accuracy on the test set because it does not vary in every incremental learning step. Concerning the *Improvement* metric, also our approach keeps the improvement in the prediction accuracy as the training progresses and the learning algorithm receives more training examples; Finally, for the *Recoverability* metric, our learning method can recover very well because it can keep the performance drops at a certain time using the bagging approach such as it selects the best prediction model, and when it determines that the result is very bad, then it retrains (for the predictive models based on incremental learning techniques) and/or learns new models.

5 Conclusions and future works

The SEIRD model is a mathematical model based on dynamical equations, which has been widely used for characterizing the pandemic of COVID-19. Several works have proposed to predict these variables using the large amount of data that has been collected on the disease. However, these models need to be updated online considering the dynamics of the disease, particularly, to be able to characterize each of the waves that the disease has had.

In this work, we have proposed a novel incremental learning architecture, which can predict the SEIRD variables. This architecture has a first component to analyze the temporal inter-dependence and intra-dependence of the SEIRD variables (it carries out a feature engineering process). Also, the architecture has a second component, an incremental learning model based on the Ensemble Learning paradigm composed of two steps: In the first step, the incremental learning model follows a bagging scheme to make predictions based on different predictive models that have been built using different machine learning techniques. Then, this model selects the best prediction as output. If the prediction is not good, then it builds new prediction models (maybe with new machine learning techniques) or/and it updates the current models (for the predictive models developed with incremental learning techniques). The incremental learning model is the second contribution of this work.

With respect to the obtained results, they are of different types. A first result is about the analysis of dependencies, which defines the relationships between the SEIRD variables and the predictor variables, and the interval of delay to be used. The first component of our architecture determines both of these things quite efficiently. A second result is the predictive models of the SEIRD variables. In general, our approach can create new predictive models or update the current models developed with incremental learning techniques. The second component of our architecture carries out these tasks quite well according to the values obtained in the performance metrics. Finally, our incremental learning architecture accomplishes the typical criteria of an incremental learning approach, namely, stability, improvement and recoverability criteria.

In general, the predictive results of our incremental learning approach are very good because, by the bagging approach, we can obtain good predictions with the predictive models for the current data; and when it considers necessary a new predict model, then it generates it with the new online data. By combining individual models, the ensemble model tends to be very flexible (less biased) and less data-sensitive (less variance). Thus, our approach can follow the different behaviors of the SEIRD variables with a low error. In each case, it considers a different predictive model to be used.

Future works require testing the model to see how it behaves in countries that have already had more than two waves, in such a way as to determine if the proposed approach can perform well in inter-wave predictions, understanding that the intra-wave transition will require of training phases to generate the appropriate models for these new waves, which our approach allows perfectly. Also, our model is based on the MAPE metric for the data dependency analysis, and in the MAPE, MSE and R^2 metrics for the estimation of the quality of the prediction. Here, again, it is possible to test with other metrics, in order to test the sensibility of our approach. Moreover, it is possible to extend the machine learning algorithms used in other areas [26, 27], to develop specialized incremental learning approaches for specific domains. Finally, the addition of ontological information during the prediction process will be analyzed to introduce contextual information like reasoning [28], or the introduction of an online feature engineering process [29, 30] inside of our incremental learning approach.

Funding This research has been carried out in the framework of the project "Plataforma web para la recolección de datos, visualización, análisis, predicción y evaluación de estrategias de control de la enfermedadproducida por SARS-CoV-2 medianteherramientas de modelación matemática, simulación e inteligencia artificial", which has been funded by the program MinCienciaTón (Covid-19 2020) of MinCiencias Colombia and EAFIT University through the agreement number 1216101576695.

Declarations

Ethical approval The paper complies with ethical Requirements.

Consent to participate and to publish The paper does not require consent to Participate and to Publish.

Research involving human participants and/or animals NA.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Lopez L, Rodo X. A Modified SEIR Model to Predict the COVID-19 Outbreak in Spain and Italy. Simulating Control Scenarios and Multi Scale Epidemics. The Lancet 2020;1:1-21
2. Prem K, Liu Y, Russell T, Kucharski A, Eggo R, Davies N. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health. 2020;5(5):e261–327.
3. Yang Z, Zeng Z, Wang K, Wong, S-S, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020;12(3).
4. Zhong L, Mu L, Li J, Wang J, Yin Z, Liu D. Early prediction of the novel coronavirus outbreak in the Mainland China based on simple mathematical model. IEEE Access. 2020;8:51761–9.
5. Yang C, Wang J. A mathematical model for the novel coronavirus epidemic in Wuhan, China. Math Biosci Eng. 2020;17(3):2708–24.
6. Zhu X, Gao B, Zhong Y, Gu C, Choi K. Extended Kalman filter based on stochastic epidemiological model for COVID-19 modelling. Comp Biol Med. 2021;137
7. Alimadadi A, Aryal S, Manandhar I, Munroe P, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics. 2020;52:200–2.
8. Vaishya R, Javaid M, Haleem I, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr. 2020;14(4):337–9.
9. Quintero Y, Ardila D, Camargo E, Rivas F, Aguilar J. Machine Learning Models for the Prediction of the SEIRD variables for the COVID-19 pandemic based on a Deep Dependence Analysis of Variables. Comp Biol Med 2021;134
10. Pelaez E, Loayza F. A deep learning model to screen for Corona Virus Disease (COVID-19) from X-ray chest images. Proc 2020 IEEE Andescon. 2020;1–6.
11. Chen S, Rajib P, Janies D, Murphy K, Feng T, Thill J. Exploring Feasibility of Multivariate Deep Learning Models in Predicting COVID-19 Epidemic. Front Public Health. 2021;9.
12. Chefrour A. Incremental supervised learning: algorithms and applications in pattern recognition. Evol Intel. 2019;12:97–112.
13. Syed NA, Liu H, Sung KK. Handling concept drifts in incremental learning with support vector machines. Proc ACM Int Conf Knowledge Discov Data Mining. 1999;317–321.
14. Luong M, Pham C. Incremental Learning for Autonomous Navigation of Mobile Robots based on Deep Reinforcement Learning. J Intell Robot Syst. 2021;101.
15. Casalino G, Castellano G, Galetta F, Kaczmarek-Majer K. Dynamic Incremental Semi-supervised Fuzzy Clustering for Bipolar Disorder Episode Prediction. Lect Notes Comput Sci. 2020;12323:79–93.
16. Braccioni F, Bottigliengo D, Ermolao A, et al. Dyspnea, effort and muscle pain during exercise in lung transplant recipients: an analysis of their association with cardiopulmonary function parameters using machine learning. Respiratory Res. 2020;21.
17. Schiel F, Hagengruber A, Vogel J, Triebel R. Incremental learning of EMG-based control commands using Gaussian Processes. Proc 4th Conf Robot Learning. 2020.
18. Wang X, Wang X, Wilkes M. A Nearest Neighbor Classifier-Based Automated On-Line Novel Visual Percept Detection Method. In: New Developments in Unsupervised Outlier Detection. Springer, Singapore. 2021.
19. Su H, Qi W, Hu Y, Karimi HR, Ferrigno G, De Momi E. An Incremental Learning Framework for Human-like Redundancy Optimization of Anthropomorphic Manipulators. IEEE Trans Industrial Inform. 2020.
20. Neto AC, Coelho RA, de Castro CL. An Incremental Learning approach using Long Short-Term Memory Neural Networks. J SociedadeBrasileira de Automática. 2020;2(1).
21. Gepperth A, Hammer B. Incremental learning algorithms and applications. Proc European Symp Artif Neural Networks (ESANN), Bruges, Belgium. 2016;hal-01418129
22. Schlimmer JC, Granger RH. Incremental learning from noisy data. Mach Learn. 1986;1(3):317–54.
23. Lu H, Wang H, Yoon S. A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. Expert Syst Appl. 2019;116:340–50.
24. Waissman J, Sarrate R, Escobet T, Aguilar J, Dahhou B. Wastewater treatment process supervision by means of a fuzzy automaton model. Proce IEEE Int Symp Intelligent Control. 2000;163–168.
25. Livieris I. An advanced active set L-BFGS algorithm for training weight-constrained neural networks. Neural Comput Appl. 2020;32:6669–84.
26. Aguilar J. A Fuzzy Cognitive Map Based on the Random Neural Model. Lect Notes Comput Sci. 2001;2070:333–8.
27. Puerto E, Aguilar J, López C, Chávez D. Using Multilayer Fuzzy Cognitive Maps to diagnose Autism Spectrum Disorder. Appl Soft Comput. 2019;75:58–71.
28. Aguilar J., Jerez M., Exposito E., Villemur T. CARMiCLOC: Context Awareness Middleware in Cloud Computing. Proc Latin American Comp Conf (CLEI). 2015.
29. Jiménez M, Aguilar J, Monsalve-Pulido J, Montoya E. An automatic approach of audio feature engineering for the extraction, analysis and selection of descriptors. Int J Multimedia Information Retrieval. 2021;10:33–42.
30. Pacheco F, Rangel C, Aguilar J, Cerrada M, Altamiranda J. Methodological framework for data processing based on the Data Science paradigm. Proceedings XL Latin American Computing Conference. 2014.