applied sciences

MDPI

Article

A Feature-Based Analysis for Time-Series Classification of COVID-19 Incidence in Chile: A Case Study

Christopher Flores [1,†], Carla Taramasco [2,*,†], María Elena Lagos [3,†], Carla Rimassa [4,†] and Rosa Figueroa [1,*,†]

[1] Departamento de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de Concepción, Concepción 4070409, Chile; christopher.flores@biomedica.udec.cl
[2] Escuela Ingeniería Informática, Universidad de Valparaíso, Valparaíso 2361845, Chile
[3] Departamento de Salud Pública, Facultad de Enfermería, Universidad de Concepción, Concepción 4070409, Chile; mariaelagos@udec.cl
[4] Facultad de Medicina, Escuela Fonoaudiología, Universidad de Valparaíso, Valparaíso 2361845, Chile; carla.rimassa@uv.cl
* Correspondence: carla.taramasco@uv.cl (C.T.); rosa.figueroa@biomedica.udec.cl (R.F.)
† These authors contributed equally to this work.

Abstract: The 2019 Coronavirus disease (COVID-19) pandemic is a current challenge for the world's health systems aiming to control this disease. From an epidemiological point of view, the control of the incidence of this disease requires an understanding of the influence of the variables describing a population. This research aims to predict the COVID-19 incidence in three risk categories using two types of machine learning models, together with an analysis of the relative importance of the available features in predicting the COVID-19 incidence in the Chilean urban commune of Concepción. The classification results indicate that the ConvLSTM (Convolutional Long Short-Term Memory) classifier performed better than the SVM (Support Vector Machine), with results between 93% and 96% in terms of accuracy (ACC) and F-measure (F1) metrics. In addition, when considering each one of the regional and national features as well as the communal features (DEATHS and MOBILITY), it was observed that at the regional level the CRITICAL BED OCCUPANCY and PATIENTS IN ICU features positively contributed to the performance of the classifiers, while at the national level the features that most impacted the performance of the SVM and ConvLSTM were those related to the type of hospitalization of patients and the use of mechanical ventilators.

Keywords: COVID-19; incidence; machine learning; SARS-CoV-2; time series classification

Citation: Flores, C.; Taramasco, C.; Lagos, M.E.; Rimassa, C.; Figueroa, R. A Feature-Based Analysis for Time-Series Classification of COVID-19 Incidence in Chile: A Case Study. *Appl. Sci.* **2021**, *11*, 7080. https://doi.org/10.3390/app11157080

Academic Editor: Flavio Cannavò

Received: 1 July 2021
Accepted: 29 July 2021
Published: 31 July 2021

1. Introduction

Coronavirus Disease 2019 (COVID-19) is a disease caused by a type of coronavirus identified in 2019 in Wuhan, Hubei, China. This disease, appearing as pneumonia of unknown cause, and whose pathogen had not been previously identified in humans, was termed Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1]. In some cases, the disease progresses to a critical illness, such as acute respiratory failure, pneumonia, renal failure, and even death [2,3]. In March 2020, when there were more than 118,000 cases in 114 countries worldwide, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic [4]. During the first semester of 2020, Chile was one of the countries most affected by the COVID-19 pandemic, ranking sixth in the number of infected people among more than two hundred countries worldwide [5]. In this complex scenario, the Ministry of Health (MINSAL) is involved, as in many other countries, in the control of the pandemic following the indications of the WHO through testing, traceability, and isolation strategies [6].

COVID-19 is transmitted from person to person by the aerosols emitted by infected people when breathing, talking, sneezing, or coughing. This infectious mechanism, added

to the fact that there may be infected asymptomatic people, makes its detection a vital strategy to control the spread of this disease [7]. Like other countries globally, the Chilean government has promoted public health strategies based on self-care, mobility reduction, testing, traceability, and isolation. Self-care measures include the use of masks, hand washing, and social distancing. At the same time, public management actions have been promoted to reduce the mobility of the population, such as quarantines, online classes, teleworking, and curfews. Among the efforts for infection control, there has been an increase in testing, an expansion in the capacity of ICU beds in hospitals, the opening of health residences, free vaccination for the population, an increase in human resources to strengthen health care, traceability of suspected as well as confirmed cases, and greater control to ensure compliance with measures. Such actions are toughened or relaxed, depending on the behavior and evolution of the disease. Among the actions to increase and reinforce testing, the government has implemented community active case-finding interventions in crowded places such as workplaces, shopping malls, and plazas [8]. Active case-finding is a community screening approach to detect active cases of COVID-19 that have not been promptly detected by spontaneous consultation, especially in those people who do not have or recognize symptoms or who, for whatever reason, have not consulted health care centers [9]. In this approach, health teams bring Polymerase Chain Reaction (PCR) testing to strategic points of the community. Reverse Transcription Polymerase Chain Reaction (RT-PCR) is one of the most frequently used laboratory methods to detect and confirm COVID-19 cases in symptomatic and asymptomatic patients. This test detects the presence of SARS-CoV-2 Ribonucleic acid (RNA) from a respiratory tract secretion sample using the RT-PCR [10]. In active case finding, it is crucial to determine the community points to be tested to increase the effectiveness of detection. Therefore, the Chilean government has urged researchers to find new methods to predict the progression of COVID-19 including the prediction of incidence and its behavior in the different regions and communes or districts of the country. This research aims to predict the COVID-19 incidence in three risk categories: low risk (less than 10 new cases per day in 100,000 inhabitants), medium risk (between 10 and 25 new cases per day in 100,000 inhabitants), and critical risk (greater than 25 new cases per day in 100,000 inhabitants). To consider possible nonlinear relationships in the interaction of variables with the prediction of incidence, we propose a machine learning model, together with an analysis of the relative importance of the available features in predicting COVID-19 incidence. Taking the above into account, the main contributions of this study can be summarized as follows:

- The development of a three-level or three-category prediction model of COVID-19 incidence tested in a Chilean case study. In order to achieve this model of incidence levels, the time series problem was transformed into a classification problem. Four classifiers were then compared to predict these levels to select the one with the best performance;
- Support of the prediction of COVID-19 risk levels, considering an automatic analysis of epidemiological indicators (classifier features or variables) linked to the behavior of the disease in the community. This model is the initial attempt to create a national model that can guide the response and intensity of the active search effort to control COVID-19 according to the trends in communal, regional, and national features in Chile expressed in incidence risk levels.

The rest of the paper is organized as follows: Section 3 briefly describes the datasets, the available features, and the methods we used in this research. Section 4 presents the performances of the classification algorithms in terms of Accuracy (ACC), Precision (P), Recall (R), and F-measure (F1) together with an analysis of the relative importance of the available features in predicting the COVID-19 incidence. Section 5 shows an analysis of the results obtained and suggests possible future research to improve the present study.

2. Related Work

The SARS-CoV-2 coronavirus, which causes the COVID-19 disease, has caused a state of alert in all countries aiming to control this pandemic that has led to more than 2.84 million deaths in the world [11,12]. From an epidemiological point of view, one of the variables to be controlled is the incidence, a concept defined as the occurrence of new cases of the disease, generally measured in a population of 100,000 inhabitants [13,14]. The literature reveals research aimed at predicting the risk of COVID-19 using different epidemiological variables in the form of time series data [15–18]. Epidemiological, statistical, or machine learning models have been used to address the problem. Among the most widely used epidemiological models are the Susceptible, Susceptible, Infectious, or Recovered (SIR) model and its variations [19,20]. For example, Malavikaa et al. used the SIR model to estimate the maximum number of active cases and the peak period of COVID-19 in India [19]. The most commonly used statistical models are data correlation, probabilistic and autoregressive models such as Auto-Regressive Integrated Moving Average (ARIMA) [19,21–28]. Roy et al. used an ARIMA model to estimate the prevalence and incidence of COVID-19 in India [28]. Although authors using epidemiological or statistical methods for time series data reported good results, one of the main disadvantages of time series models is that they only perform well if there is a linear relationship between the variables analyzed [16,29–33]. In order to capture possible nonlinear relationships among the variables of interest, researchers have tried to use machine learning models [16,29–31,34–39]. For this purpose, some of the most commonly used algorithms are logistic regression, Support Vector Machine (SVM), Multilayer Perceptron (MLP) and more recently DL-based algorithms, such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). For example, Singh et al. used an ARIMA model and an SVM model to predict the daily confirmed cases of COVID-19 in the five most affected countries during the study period. The results obtained indicate that the SVM performed better than the ARIMA model in terms of ACC [30]. In another study, Shahid and Zameer used LSTM based models, Support Vector Regression (SVR), and ARIMA to predict confirmed cases, deaths and recovered cases in ten of the most affected countries during the study period. The results showed that models based on LSTM have a lower predictive error measured as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) than the SVR and ARIMA models [31]. In addition, Shastri et al. showed that the use of the Convolutional LSTM (ConvLSTM) improved the prediction of confirmed cases of COVID-19 in the USA and India, exhibiting performances of over 85% in terms of both ACC and F1 metrics [37].

Finally, one of the critical aspects to consider in an analysis of COVID-19 predictive models is the need to determine the importance of the features involved in the spread and risk of contagion of this disease. In this respect, Wilde et al. demonstrated an association between a high risk of mortality in patients admitted to the Intensive Care Unit (ICU) and increased occupancy of mechanically ventilated beds in England [40]. Other studies showed that mobility has a significant impact on the incidence of disease in the samples analyzed [34,41–45].

3. Materials and Methods

3.1. Datasets and Pre-Processing

The data for this study were obtained from the official COVID-19 repository of Ministry of Science, Technology, Knowledge and Innovation (MICITEC), which contains national, regional, and communal data associated with COVID-19. We used the commune of Concepción as a study case for feature importance analysis and incidence classification (refer to Figure 1). This commune is located in one of the most densely populated regions of the country: Biobío Region [46]. From the official COVID-19 repository, we used databases containing national summaries, regional summaries, and commune data for the period from April 2020 to March 2021.

At the commune level, we used the following features:

- Mobility: This feature was reported as the movements of mobile phones (antenna transitions) connected to Telefonica's national network, in a grouped and anonymous way. The internal mobility is measured as the evolution of travel occurring within the commune, and the external mobility is calculated as the movement from outside the commune into the commune. The mobility index corresponds to the number of trips within a specific commune normalized by the number of commune residents.
- Deaths: Number of deaths by residence and date in the commune of Concepcion.

At the regional level, we used the following features:

- Patients in ICU: Number of patients in an ICU who are COVID-19 confirmed cases by region, reported daily.
- Deaths: Number of COVID-19 diagnosed deaths per day, according to the region of residence, reported daily.
- Critical bed occupancy: Number of beds in an adult ICU by region, total occupation, and occupation by COVID-19.

Finally, at the national level, we used the following features:

- Patients in ICU (age group): Number of patients in an ICU by age groups (<=39; 40–49; 50–59; 60–69; and >=70) who are COVID-19 confirmed cases, reported daily.
- Deaths (age group): Number of people who died of COVID-19, grouped by age ranges (<=39; 40–49; 50–59; 60–69; 70–79; 80–89; and >=90) reported daily.
- Mechanical ventilator availability: Number of mechanical ventilators available and the number of ventilators occupied for each date reported.
- Patient hospitalization (bed type): Number of hospitalized patients with a COVID-19 diagnosis according to the type of bed they occupy: basic, medium, ITU and ICU.
- Patients on mechanical ventilation: Number of hospitalized patients in the ICU. The number of patients who require invasive mechanical ventilation, patients without mechanical ventilation and those connected to noninvasive mechanical ventilation who are COVID-19 confirmed cases.

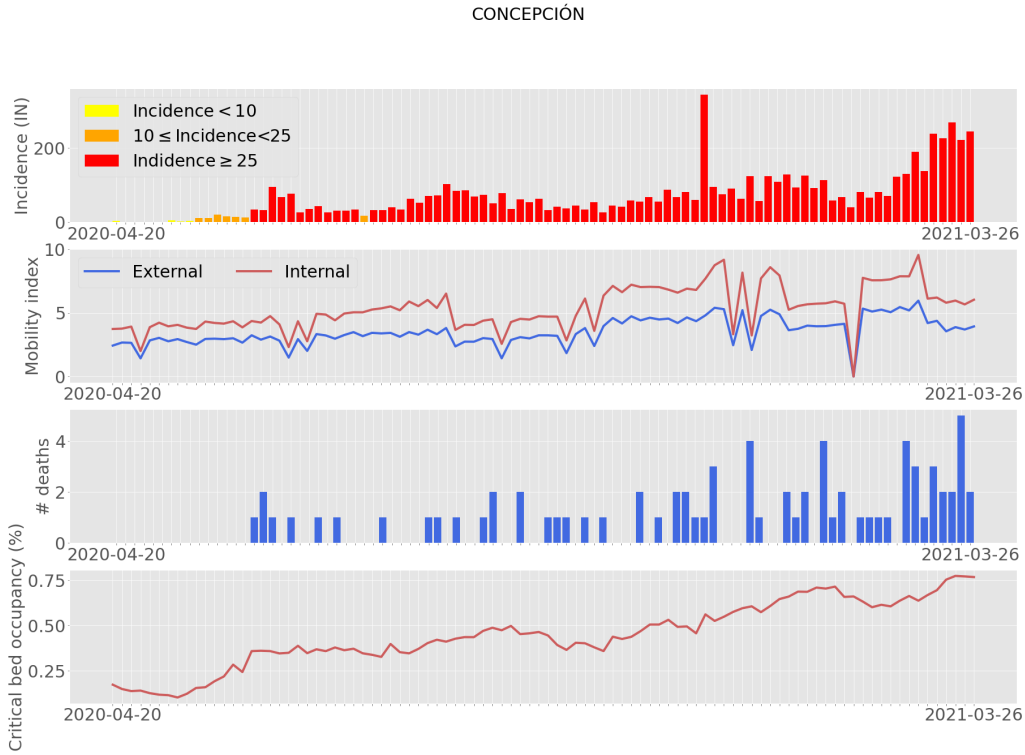

Figure 1. Evolution of some COVID-19 features together with the incidence behavior over the study time.

We standardized these datasets, i.e., we centered the values around the mean with unit standard deviation [47]. Standardization is one of the feature scaling techniques that aim to make the gradient descent converge faster to minimum values in neural network-based algorithms or make all features contribute equally in the case of distance-based algorithms such as SVM [48]. In addition to describing the variables provided in this section, we will perform a descriptive statistical analysis for a better understanding of the study.

3.2. Problem Definition

This article presents an exploratory analysis of the possibility of predicting risk categories or classes of COVID-19 using freely available data in Chile as a way to guide the active search for cases (refer to Figure 2). The classes were defined based on the incidence risk level model proposed by Harvard Global Health Institute (HGHI) [14] and the availability of case study data. Therefore, machine learning models classify the incidence rate at each time step between low risk (less than 10 new cases per 100,000 people), medium risk (10 to 25 cases per 100,000 people), and critical risk (more than 25 cases per 100,000 people).

Consider a dataset containing a multivariate time series of n training samples $X = \{x_{tr}\}_{t=1}^{n}$, where $x_{tr} \in R^{s \times f}$ represents a sample of s historical time steps from the current value for each of the f features of the problem. Consider also that each i-th sample is assigned a class $y_t \in Y$, where $y_t \in L = \{0, 1, 2\}$ and $Y = \{y_t\}_{t=1}^{n}$ represent the set of all labels, as a function of the incidence value for a given time step (refer to Figure 1). The aim of supervised training is to find a decision function $\delta(\cdot)$ that allows the assignation of a class y_t to a test sample $x_t \in X_T$ according to $\delta(x_t) : x_t \subseteq X_T \rightarrow y_t \subseteq L$. Finally, consider a baseline model of each classification algorithm trained only with the communal features whose decision function is denoted as $\delta_b(\cdot)$. This baseline model will be used to evaluate the importance of each of the features in the classification problem.

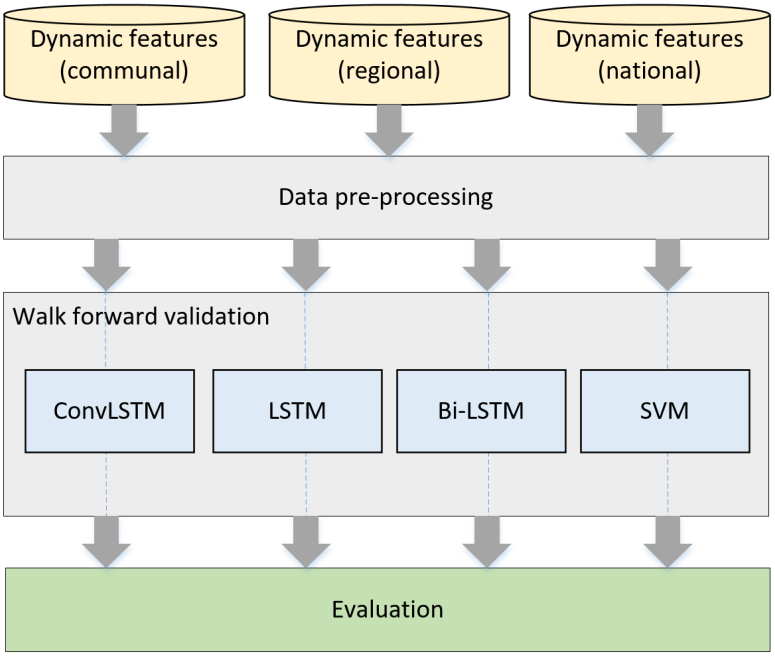

Figure 2. Proposed workflow.

3.3. Classification Algorithms

For classification, we considered two of the most widely used supervised algorithms in time series classification: LSTM-based classifiers and SVM [16,30,31,36–39,49]. In the case of LSTM-based classifiers, we have employed models based on simple LSTM, Bidirectional Long Short-Term Memory (BiLSTM) and ConvLSTM with typical parameters (refer to Figure 3, and Table 1). BiLSTM provides additional training by analyzing the data from

left to right and right to left [50]. On the other hand, ConvLSTM is a type of LSTM that includes a convolution operation inside the cell [51]. Thus, such models are useful to analyze data sequences with spatio-temporal information [37,52,53]. The decision function of the LSTM-based classifiers relies on the value of the softmax function of the classification layer according to:

Table 1. Parameters of the classification algorithms.

Classifier	Parameter	Value
LSTM models	filters	64
	kernel size	(1,2)
	epochs	40
	activation function (units)	ReLU
	dropout	0.2
	optimizer	Adam
	learning rate	10^{-3}
	batch size	8
	hidden units	128
	activation function (class)	Softmax
	loss function	Categorical crossentropy
SVM	kernel	rbf
	C	1.0
	γ	escale *

* Multiplicative inverse between the number of features and the variance of the data.

$$\delta(x_i) = \underset{j \in [1,\ldots,|L|]}{\operatorname{argmax}} \frac{e^{z_j}}{\sum_{k=1}^{l} e^{z_k}}, \tag{1}$$

where $z = \{z_1, \ldots, z_l\}$ is the intermediate output of the softmax layer for the test data point x_i. With regard to SVM, the decision function depends on n_s vectors that help to form the class separation hyperplane (support vectors) with their respective weights α_j and classes y_j and a kernel function that makes it possible to transform the input data $K(\cdot)$ to another dimension, where b is a scalar value and $\text{sign}(\cdot)$ is the sign function [54–56]:

$$\delta(x_i) = \text{sign}(\sum_{j=1}^{n_s} \alpha_j y_j K(\vec{x}_i, \vec{x}_j) + b), \tag{2}$$

In SVM classifiers, we apply kernel rbf to capture the nonlinear patterns generally present in temporal data such as COVID-19. The selected kernel is defined as $K(\cdot) = \exp(-\gamma||\vec{x}_i - \vec{x}_j||)$ with γ a hyperparameter that controls the tradeoff between error due to bias and variance in the model, keeping the other parameters as default (refer to Table 1) [33,57]. Finally, in both classifiers we considered $s = 4$ time steps for each of the i-th sample features [58,59]. In the case of SVM, an average of the four time steps was considered to form an f-dimensional feature vector at each time instant.

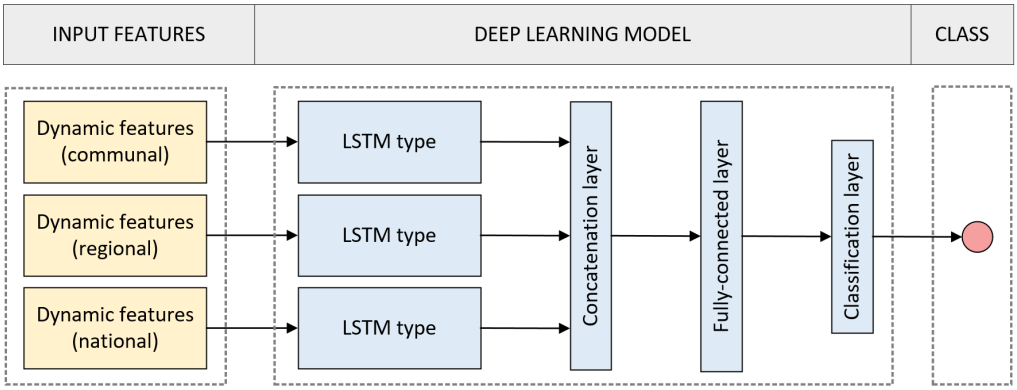

Figure 3. Proposed deep learning model to estimate the risk of COVID-19 incidence.

3.4. Evaluation of Algorithms

Owing to constraints in terms of the amount of available data, we considered the technique known as "walk forward validation" to evaluate the performance of the classification algorithms [60]. This technique allows us to perform a progressive evaluation of the performance using temporal data as new time-dependent observations are available (refer to Figure 4). In this case, in each iteration, the classifier's performance at the i-th sample was evaluated by our algorithm according to its features at that time instant. The metrics for evaluating the test samples are the ACC, P, R, and F1:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{3}$$

$$P_i = \frac{TP_i}{TP_i + FP_i}, \tag{4}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}, \tag{5}$$

$$F1_i = \frac{2P_iR_i}{P_i + R_i}, \tag{6}$$

where TP and FP correspond to the true and false positives, while TN and FN correspond to the true and false negatives of classification. Note that P, R, and $F1$ are calculated for each one of the i-th classes. In addition, our algorithm evaluates the training error of the classifiers in terms of the zero-one-loss metric ($\mathcal{L}$) as follows [61]:

$$\mathcal{L}(y_i, y_i') = \begin{cases} 1 & y_i \neq y_i' \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where y_i and y_i' represent the predictions and actual classes, respectively. To measure the relative importance of the available features in the prediction of COVID-19 incidence, we proposed an iterative wrap-around method that considers the following steps (refer to Figure 5). First, our approach creates a baseline model $\delta_b(\cdot)$ that considers only the communal feature subset to obtain an initial performance. Then, each of the regional and national features is added to the model individually (one at a time) to create a feature space. For each feature added to the model, the algorithm measures the performance difference compared to the performance of the baseline model. Finally, the performance differences are sorted from highest to lowest. A larger difference indicates that this variable is more relevant to the problem because it contributes more to the classifier's performance.

Figure 4. Walk forward validation scheme using an expanding window.

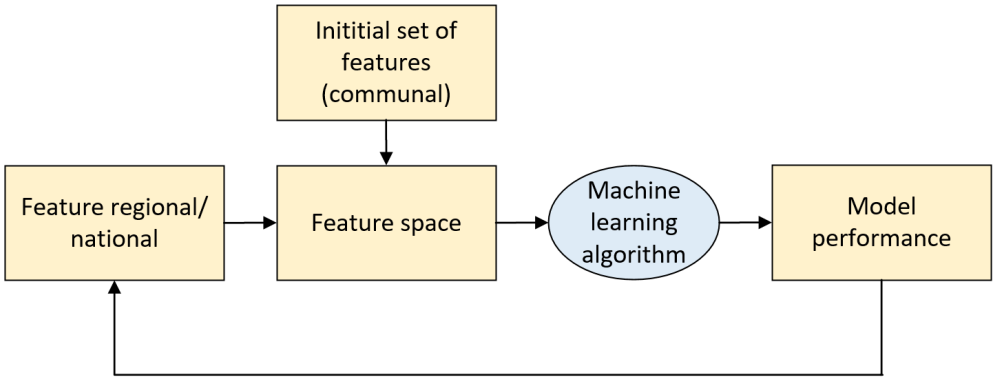

Figure 5. Proposed workflow to determine the relevance of the features.

4. Results

Figure 6 shows the boxplot of the commune and regional features. This graph mainly shows internal mobility indexes that are larger than external ones, outliers in communal deaths ($\geq$3), and a maximum occupancy of critical beds close to 80%. On the other hand, Figure 7 shows the correlation matrix of the communal and regional features, including incidence per 100,000 inhabitants. From this Figure, we observed a high correlation between incidence and variables related to ICU patients, a moderate correlation between incidence and deaths (communal), and a low correlation between incidence and mobility. As mentioned in the previous section, we used the HGHI model as a basis for defining incidence risk levels. This model classifies incidence into four levels: green (less than one new case per 100,000 people), yellow (one to nine cases per 100,000 people), orange (ten to twenty-five new cases per 100,000 people), and red (more than 25 new cases per day per 100,000 people). In this research, the distribution of the incidence data proposed by Harvard is shown in Figure 8. From this Figure, we can observe poor data availability in the green and yellow levels; since this study uses machine learning models, we reclassified the incidence data keeping the thresholds proposed by Harvard but combining the green and yellow categories into a single category named low risk. Then, the machine learning models classify the incidence into three classes: low risk (less than 10 new cases per day in 100,000 inhabitants), medium risk (between 10 and 25 new cases per day in 100,000 inhabitants), and critical risk (greater than 25 new cases per day in 100,000 inhabitants).

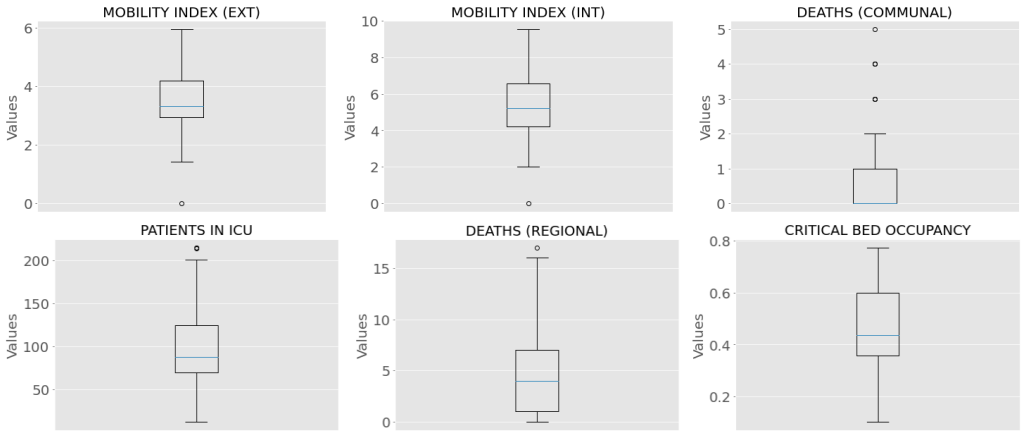

Figure 6. Boxplot for commune and regional features.

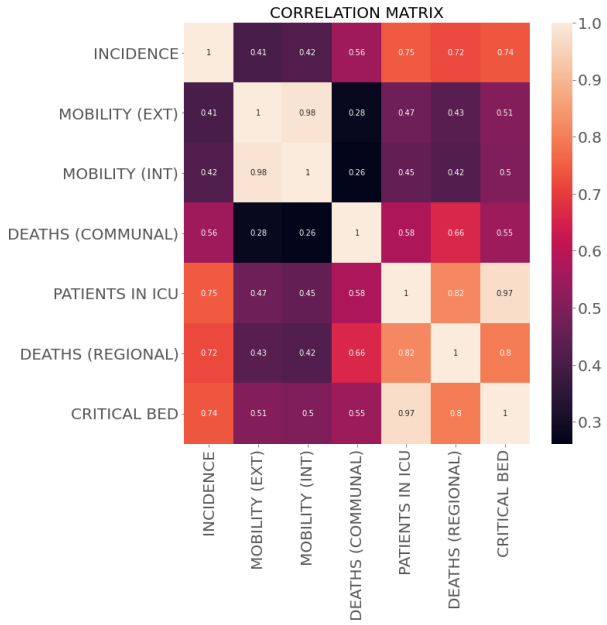

Figure 7. Correlation matrix (pearson) for commune and regional features.

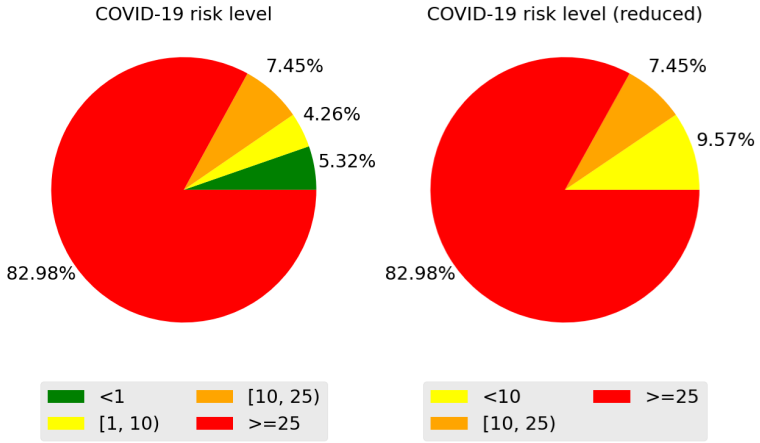

Figure 8. Class distribution of the dataset. Left: HGHI model. Right: HGHI model reduced according to the number of available examples.

Table 2 shows the classification results of the different models implemented, based on an SVM and LSTM. Both types of classifier had a good overall performance in classifying

the incidence classes with results above 93% in all reported metrics. In all cases the F1 values were greater than ACC (≥94%). In this sense, the Precision values were higher than Recall values, meaning that classifiers were more able to correctly detect the positive class (i.e., each of the classes in the problem). In all cases, LSTM-based classifiers performed better than or equal to the SVM. Additionally, we observed that the performance of ConvLSTM was better than that of the rest of the classifiers in all performance metrics.

Table 2. Weighted average classification results.

Classifier	ACC (%)	P (%)	R (%)	F1 (%)
SVM	93.98	95.66	93.98	94.81
LSTM	95.18	96.05	95.18	95.61
BiLSTM	93.98	95.66	93.98	94.81
ConvLSTM	**96.39**	**96.71**	**96.39**	**96.55**

Values in bold indicate better performances.

Table 3 shows the average results of the classifier training error curves (refer to (7)) as shown in Figure 9 according to the validation scheme used [62,63]. From Table 3, we can see that in all cases the training error of SVM was lower than all LSTM-based classifiers. On the other hand, the largest error was obtained in the ConvLSTM-based classifier. These results indicate that LSTM-based classifiers could be less prone to overfitting than SVM because although they presented a higher training error, they performed better when classifying test examples. In addition, the training error decreases noticeably as the number of training examples increases.

Table 3. Average classification error results.

Classifier	Classification Error (Zero-One Loss)
SVM	**0.0180 ± 0.0178**
LSTM	0.0241 ± 0.0197
BiLSTM	0.0218 ± 0.0188
ConvLSTM	0.0250 ± 0.0210

Values in bold indicate better performances.

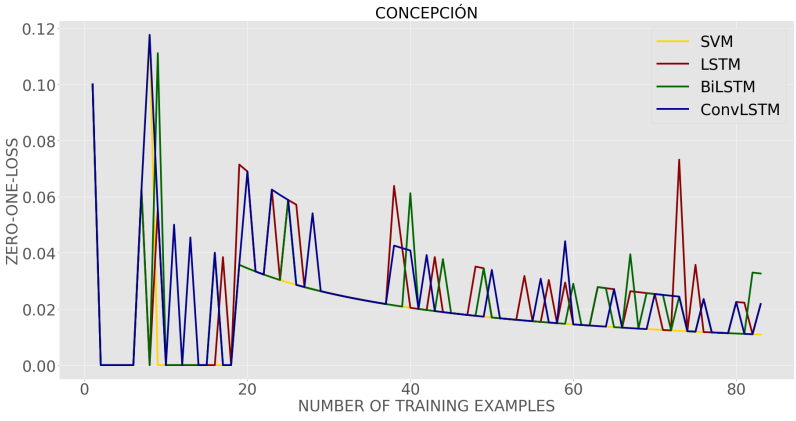

Figure 9. Error curves of the classifiers.

To measure feature significance in the performance of the classifiers, our approach individually incorporates regional and national features into the set of commune features, as mentioned in Section 3.4. Figure 10 shows the performance difference in terms of the zero-one-loss metric considering the set of commune features as the baseline. We note that, at the regional level, the CRITICAL BED OCCUPANCY feature ranks first in error difference, followed by the PATIENTS IN ICU and DEATHS features. On the other hand, at the national level, the PATIENTS IN ICU, PATIENT HOSPITALIZATION, BED TYPES

features and features related to the type of mechanical ventilators rank first in both SVM- and LSTM-based classifiers.

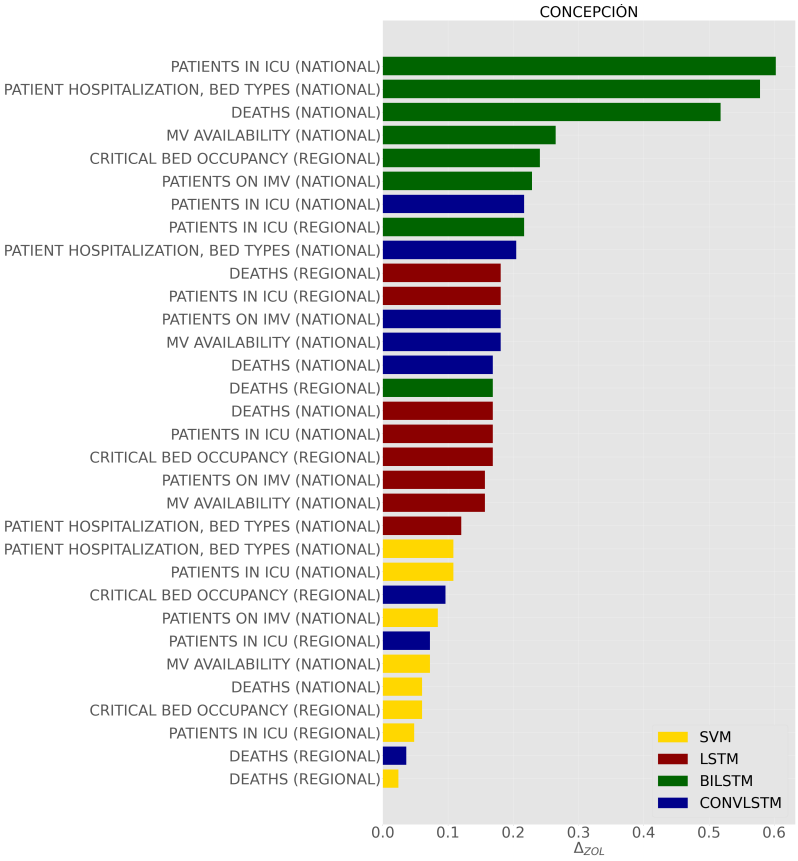

Figure 10. This bar plot shows the relative importance of the features in terms of zero-one-loss error in predicting risk categories.

5. Conclusions and Future Research

This article presents a model for predicting COVID-19 incidence in three risk categories: low, medium, and critical. Predicting the incidence in categories responds to the need to support the active search strategies carried out in the country for the detection and control of COVID-19. The risk levels of the model can be interpreted following the HGHI framework recommendations as follows: (i) the yellow level means a low transmission rate of the disease and orients decision makers to continue with testing and traceability measures; (ii) the orange level means that the virus transmission speed has increased, so that it is necessary to evaluate measures to increase active search efforts in communities at this risk level; and (iii) the red level means the speed of virus transmission is high, with it therefore being necessary to evaluate more aggressive measures to contain the spread of the disease in the community.

The SVM- and LSTM-based classifiers tested in the case study showed good performance ranging between 93% and 96%. On the other hand, the zero-one-loss error decreases as more data are incorporated, with the SVM model having the lowest value. This result is explained by the small amount of data available to train and test the models since LSTM-based classifiers usually require considerably more data. As demonstrated, these models could be potentially helpful to evaluate the risk of incidence in a community; however, they do not replicate the structure proposed by the original HGHI classification, so that further investigation and the incorporation of more data are recommended.

Taking into consideration the dynamic commune features (DEATHS and MOBILITY) as the baseline in each classifier, the importance of each of the regional and national features (refer to Figure 10) was analyzed. In this regard, it was observed that, in most

cases, each of the commune and regional features positively impacted the performance of the classifiers. For example, at the regional level, CRITICAL BED OCCUPANCY and PATIENTS IN ICU rank first in terms of error difference with respect to those of the commune. On the other hand, in the case of national features, a more significant influence was found on the performance of classifiers in the use of features related to the type of hospitalization of patients and the use of mechanical ventilators. These results are consistent with the results of previous studies related to COVID-19 that show an association between COVID-19 and risk factors associated with patients admitted to the Intensive Care Unit (ICU) and occupancy of mechanically ventilated beds [34,40–45,64].

We would like to point out that this study has the following limitations: (i) The data utilized correspond to free reports from MICITEC, so that there is no control over the frequency and quality of the information reported, which resulted in low availability of data for analysis (approximately two weekly reports for the period of study). (ii) The progression of COVID-19 in Chile has not allowed us to have a reasonable amount of data to train a model that considers the low incidence categories proposed by HGHI; in fact, after evaluating several Chilean communes, only Concepcion presented a data distribution that allowed us to identify at least three categories. We therefore aim to continue researching as more data become available to confirm this study's results and to extend the model to more communes. Consequently, in future work, we plan to pursue the following research directions: (i) In this paper, we grouped the incidence data by keeping the thresholds proposed by Harvard but combining the green and yellow categories into a single category that is referred to as low risk. As a future research direction, we propose studying, both technically and epidemiologically, the effect of regrouping the Harvard model categories to obtain a larger number of cases per class (e.g., considering only a binary problem for nonhigh and high incidence levels with a threshold of 25 cases per 100,000 population, or redefining new thresholds). (ii) We propose developing a software tool that allows decision makers to direct active search efforts using this predictive model, since at the present time they are only directed towards crowded places. In this regard, the idea of exploring new COVID-19 incidence levels according to the availability and distribution of data is also under consideration. (iii) We would like to analyze the personal risk of COVID-19 infection in Chile by examining the commune, regional, and national data of this study, as well as health and socio-demographic data as soon as these data become available to us. (iv) Finally, we consider the possibility of extending the definition of our research problem to combine classification models with epidemiological models that consider the individual prediction of COVID-19 incidence, concerning confirmed cases, deaths and recovered cases, using MAE and RMSE as statistical loss functions.

Author Contributions: Conceptualization, C.F., C.T., M.E.L., C.R. and R.F.; formal analysis, C.F., C.T., M.E.L. and R.F.; funding acquisition, C.T.; methodology, C.F., C.T., M.E.L., C.R. and R.F.; supervision, M.E.L. and R.F.; writing—original draft, C.F., M.E.L. and R.F.; writing—review and editing, C.T. and C.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Agency for Research and Development (ANID) under the COVID0739 grant, the National Center on Health Information Systems through the CORFO Project under Grant 16CTTS-66390, and FONDECYT Regular 1201787.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data for this study were obtained from the official COVID-19 repository of the Ministry of Science, Technology, Knowledge and Innovation (MICITEC) located at https://github.com/MinCiencia/Datos-COVID19, accessed on 27 July, 2021.

Acknowledgments: The authors would like to thank the Agency for Research and Development (ANID) under COVID0739 and the National Center on Health Information Systems (CORFO-CENS 16CTTS-66390).

Conflicts of Interest: The authors declare that there are no conflict of interest.

References

1. Honein, M.A.; Christie, A.; Rose, D.A.; Brooks, J.T.; Meaney-Delman, D.; Cohn, A.; Sauber-Schatz, E.K.; Walker, A.; McDonald, L.C.; Liburd, L.C.; et al. Summary of Guidance for Public Health Strategies to Address High Levels of Community Transmission of SARS-CoV-2 and Related Deaths, December 2020. *Morb. Mortal. Wkly. Rep.* **2020**, *69*, 1860. [CrossRef]
2. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
3. Toniati, P.; Piva, S.; Cattalini, M.; Garrafa, E.; Regola, F.; Castelli, F.; Franceschini, F.; Focà, E.; Andreoli, L.; Latronico, N.; et al. Tocilizumab for the treatment of severe COVID-19 pneumonia with hyperinflammatory syndrome and acute respiratory failure: A single center study of 100 patients in Brescia, Italy. *Autoimmun. Rev.* **2020**, *19*, 102568. [CrossRef]
4. Shah, K.; Kamrai, D.; Mekala, H.; Mann, B.; Desai, K.; Patel, R.S. Focus on mental health during the coronavirus (COVID-19) pandemic: Applying learnings from the past outbreaks. *Cureus* **2020**, *12*, e7405. [CrossRef] [PubMed]
5. Benavides, G.A.; Larach, F.; Marchant, V.; Fernández, J.; Montoya, F.; Domínguez, S.; Mejías, C. The evolution of the COVID-19 pandemic in Chile during 2020: A data perspective. *arXiv* **2021**, arXiv:2102.11841.
6. Canals, M.; Cuadrado, C.; Canals, A.; Yohannessen, K.; Lefio, L.A.; Bertoglia, M.P.; Eguiguren, P.; Siches, I.; Iglesias, V.; Arteaga, O. Epidemic trends, public health response and health system capacity: The Chilean experience in four months of the COVID-19 pandemic. *Rev. Panam. Salud Públic.* **2020**, *44*, e99. [CrossRef] [PubMed]
7. Jayaweera, M.; Perera, H.; Gunawardana, B.; Manatunge, J. Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environ. Res.* **2020**, 188, 109819. [CrossRef]
8. Li, Z.; Chen, Q.; Feng, L.; Rodewald, L.; Xia, Y.; Yu, H.; Zhang, R.; An, Z.; Yin, W.; Chen, W.; et al. Active case finding with case management: The key to tackling the COVID-19 pandemic. *Lancet* **2020**, *396*, 63–70. [CrossRef]
9. Ministerio de Salud. *Manual Operativo para la Búsqueda Activa de Casos en la Comunidad y Diagnóstico Precoz Covid-19*; 2021. Available online: https://www.minsal.cl/wp-content/uploads/2020/10/201005-MANUAL-BAC.pdf (accessed on 30 June 2021).
10. Corman, V.M.; Landt, O.; Kaiser, M.; Molenkamp, R.; Meijer, A.; Chu, D.K.; Bleicker, T.; Brünink, S.; Schneider, J.; Schmidt, M.L.; et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **2020**, *25*, 2000045. [CrossRef]
11. Fontanet, A.; Autran, B.; Lina, B.; Kieny, M.P.; Karim, S.S.A.; Sridhar, D. SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet* **2021**, *397*, 952–954. [CrossRef]
12. Roser, M.; Ritchie, H.; Ortiz-Ospina, E.; Hasell, J. Coronavirus Pandemic (COVID-19). *Our World Data* **2020**. Available online: https://ourworldindata.org/coronavirus (accessed on 30 June 2021).
13. Dalsgaard, S.; Thorsteinsson, E.; Trabjerg, B.B.; Schullehner, J.; Plana-Ripoll, O.; Brikell, I.; Wimberley, T.; Thygesen, M.; Madsen, K.B.; Timmerman, A.; et al. Incidence rates and cumulative incidences of the full spectrum of diagnosed mental disorders in childhood and adolescence. *JAMA Psychiatry* **2020**, *77*, 155–164. [CrossRef] [PubMed]
14. Safra, E.J. *Key Metrics for COVID Suppression: A Framework for Policy Makers and the Public*; 2020. Available online: https://ethics.harvard.edu/files/center-for-ethics/files/key_metrics_and_indicators_v4.pdf (accessed on 30 June 2021).
15. Qi, H.; Xiao, S.; Shi, R.; Ward, M.P.; Chen, Y.; Tu, W.; Su, Q.; Wang, W.; Wang, X.; Zhang, Z. COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci. Total. Environ.* **2020**, *728*, 138778. [CrossRef]
16. Chimmula, V.K.R.; Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **2020**, *135*, 109864. [CrossRef]
17. Hu, Z.; Ge, Q.; Li, S.; Jin, L.; Xiong, M. Artificial intelligence forecasting of covid-19 in china. *arXiv* **2020**, arXiv:2002.07112.
18. Bertozzi, A.L.; Franco, E.; Mohler, G.; Short, M.B.; Sledge, D. The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 16732–16738. [CrossRef] [PubMed]
19. Malavika, B.; Marimuthu, S.; Joy, M.; Nadaraj, A.; Asirvatham, E.S.; Jeyaseelan, L. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. *Clin. Epidemiol. Glob. Health* **2021**, *9*, 26–33. [CrossRef] [PubMed]
20. Calafiore, G.C.; Novara, C.; Possieri, C. A modified SIR model for the COVID-19 contagion in Italy. In Proceedings of the 2020 59th IEEE Conference on Decision and Control (CDC), Jeju Island, Korea, 14–18 December 2020; pp. 3889–3894.
21. Giuliani, D.; Dickson, M.M.; Espa, G.; Santi, F. Modelling and predicting the spatio-temporal spread of COVID-19 in Italy. *BMC Infect. Dis.* **2020**, *20*, 1–10. [CrossRef]
22. Deb, S.; Majumdar, M. A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. *arXiv* **2020**, arXiv:2003.10655.
23. Hamidouche, M. COVID-19 outbreak in Algeria: A mathematical model to predict the incidence. *medRxiv* **2020**. [CrossRef]
24. Panuganti, B.A.; Jafari, A.; MacDonald, B.; DeConde, A.S. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. *Otolaryngol. Head Neck Surg.* **2020**, *163*, 491–497. [CrossRef]
25. Yuan, X.; Xu, J.; Hussain, S.; Wang, H.; Gao, N.; Zhang, L. Trends and prediction in daily new cases and deaths of COVID-19 in the United States: An internet search-interest based model. *Explor. Res. Hypothesis Med.* **2020**, *5*, 1. [CrossRef]
26. Stübinger, J.; Schneider, L. Epidemiology of coronavirus covid-19: Forecasting the future incidence in different countries. In *Healthcare*; Multidisciplinary Digital Publishing Institute: Basel, Switzerland, 2020; Volume 8, p. 99. [CrossRef]
27. Paul, M.; Held, L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat. Med.* **2011**, *30*, 1118–1136. [CrossRef] [PubMed]

28. Roy, S.; Bhunia, G.S.; Shit, P.K. Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Model. Earth Syst. Environ.* **2020**, *7*, 1385–1391. [CrossRef] [PubMed]
29. Mollalo, A.; Rivera, K.M.; Vahedi, B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4204. [CrossRef] [PubMed]
30. Singh, S.; Parmar, K.S.; Makkhan, S.J.S.; Kaur, J.; Peshoria, S.; Kumar, J. Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. *Chaos Solitons Fractals* **2020**, *139*, 110086. [CrossRef]
31. Shahid, F.; Zameer, A.; Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* **2020**, *140*, 110212. [CrossRef]
32. Harb, A.M.; Harb, S.M. Corona COVID-19 spread-a nonlinear modeling and simulation. *Comput. Electr. Eng.* **2020**, *88*, 106884. [CrossRef]
33. Rohith, G.; Devika, K. Dynamics and control of COVID-19 pandemic with nonlinear incidence rates. *Nonlinear Dyn.* **2020**, *101*, 2013–2026. [CrossRef]
34. Ramchandani, A.; Fan, C.; Mostafavi, A. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access* **2020**, *8*, 159915–159930. [CrossRef]
35. Haarhaus, M.; Santos, C.; Haase, M.; Mota Veiga, P.; Lucas, C.; Macario, F. Risk prediction of COVID-19 incidence and mortality in a large multi-national hemodialysis cohort: Implications for management of the pandemic in outpatient hemodialysis settings. *Clin. Kidney J.* **2021**, *14*, 805–813. [CrossRef] [PubMed]
36. Ayyoubzadeh, S.M.; Ayyoubzadeh, S.M.; Zahedi, H.; Ahmadi, M.; Kalhori, S.R.N. Predicting COVID-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study. *JMIR Public Health Surveill.* **2020**, *6*, e18828. [CrossRef] [PubMed]
37. Shastri, S.; Singh, K.; Kumar, S.; Kour, P.; Mansotra, V. Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Chaos Solitons Fractals* **2020**, *140*, 110227. [CrossRef]
38. Singh, V.; Poonia, R.C.; Kumar, S.; Dass, P.; Agarwal, P.; Bhatnagar, V.; Raja, L. Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. *J. Discret. Math. Sci. Cryptogr.* **2020**, *23*, 1583–1597. [CrossRef]
39. Rustam, F.; Reshi, A.A.; Mehmood, A.; Ullah, S.; On, B.W.; Aslam, W.; Choi, G.S. COVID-19 future forecasting using supervised machine learning models. *IEEE Access* **2020**, *8*, 101489–101499. [CrossRef]
40. Wilde, H.; Mellan, T.; Hawryluk, I.; Dennis, J.M.; Denaxas, S.; Pagel, C.; Duncan, A.; Bhatt, S.; Flaxman, S.; Mateen, B.A.; et al. The association between mechanical ventilator availability and mortality risk in intensive care patients with COVID-19: A national retrospective cohort study. *medRxiv* **2021**. [CrossRef]
41. Nouvellet, P.; Bhatia, S.; Cori, A.; Ainslie, K.E.; Baguelin, M.; Bhatt, S.; Boonyasiri, A.; Brazeau, N.F.; Cattarino, L.; Cooper, L.V.; et al. Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **2021**, *12*, 1–9. [CrossRef] [PubMed]
42. Zhou, Y.; Xu, R.; Hu, D.; Yue, Y.; Li, Q.; Xia, J. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: A modelling study using mobile phone data. *Lancet Digit. Health* **2020**, *2*, e417–e424. [CrossRef]
43. Shi, Z.; Fang, Y. Temporal relationship between outbound traffic from Wuhan and the 2019 coronavirus disease (COVID-19) incidence in China. *MedRxiv* **2020**. [CrossRef]
44. Mazzoli, M.; Mateo, D.; Hernando, A.; Meloni, S.; Ramasco, J.J. Effects of mobility and multi-seeding on the propagation of the COVID-19 in Spain. *MedRxiv* **2020**. [CrossRef]
45. Andersen, M.S.; Bento, A.I.; Basu, A.; Marsicano, C.; Simon, K. College openings, mobility, and the incidence of covid-19 cases. *medRxiv* **2020**. [CrossRef]
46. Prada, J. Understanding studentification dynamics in low-income neighbourhoods: Students as gentrifiers in Concepcion (Chile). *Urban Studies* **2019**, *14*, 2863–2879. [CrossRef]
47. Stajkowski, S.; Zeynoddin, M.; Farghaly, H.; Gharabaghi, B.; Bonakdari, H. A Methodology for Forecasting Dissolved Oxygen in Urban Streams. *Water* **2020**, *12*, 2568. [CrossRef]
48. Dick, G.; Owen, C.A.; Whigham, P.A. Feature standardisation and coefficient optimisation for effective symbolic regression. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference, Cancún, Mexico, 8–12 July 2020; pp. 306–314. [CrossRef]
49. Parbat, D.; Chakraborty, M. A python based support vector regression model for prediction of COVID19 cases in India. *Chaos Solitons Fractals* **2020**, *138*, 109942. [CrossRef] [PubMed]
50. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The performance of LSTM and BiLSTM in forecasting time series. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 3285–3292. [CrossRef]
51. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.-k.; Woo, W.-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *2015*, 802–810.
52. Song, H.; Wang, W.; Zhao, S.; Shen, J.; Lam, K.M. Pyramid dilated deeper convlstm for video salient object detection. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 October 2018; pp. 715–731. [CrossRef]
53. Kim, S.; Hong, S.; Joh, M.; Song, S.k. Deeprain: Convlstm network for precipitation prediction using multichannel radar data. *arXiv* **2017**, arXiv:1711.02316.

54. Karim, F.; Majumdar, S.; Darabi, H. Insights into LSTM fully convolutional networks for time series classification. *IEEE Access* **2019**, *7*, 67718–67725. [CrossRef]
55. Yildirim, O.; Baloglu, U.B.; Tan, R.S.; Ciaccio, E.J.; Acharya, U.R. A new approach for arrhythmia classification using deep coded features and LSTM networks. *Comput. Methods Programs Biomed.* **2019**, *176*, 121–133. [CrossRef]
56. Sharif, O.; Hossain, E.; Hoque, M.M. TechTexC: Classification of Technical Texts using Convolution and Bidirectional Long Short Term Memory Network. *arXiv* **2020**, arXiv:2012.11420.
57. Yuan, J.; Wu, Y.; Jing, W.; Liu, J.; Du, M.; Wang, Y.; Liu, M. Non-linear correlation between daily new cases of COVID-19 and meteorological factors in 127 countries. *Environ. Res.* **2021**, *193*, 110521. [CrossRef] [PubMed]
58. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; Volume 4, pp. 2047–2052. [CrossRef]
59. Nabavi, S.S.; Rochan, M.; Wang, Y. Future semantic segmentation with convolutional lstm. *arXiv* **2018**, arXiv:1807.07946.
60. Hu, M.Y.; Zhang, G.; Jiang, C.X.; Patuwo, B.E. A cross-validation analysis of neural network out-of-sample performance in exchange rate forecasting. *Decis. Sci.* **1999**, *30*, 197–216. [CrossRef]
61. Charuvaka, A.; Rangwala, H. *HierCost: Improving Large Scale Hierarchical Classification with Cost Sensitive Learning*; Springer: Berlin/Heidelberg, Germany, 2015. [CrossRef]
62. Zhao, J.; Li, Y.; Yu, X.; Zhang, X. Levenberg-Marquardt algorithm for Mackey-Glass chaotic time series prediction. *Discret. Dyn. Nat. Soc.* **2014**, *2014*. [CrossRef]
63. Cao, L.J.; Tay, F.E.H. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* **2003**, *14*, 1506–1518. [CrossRef]
64. Mena, G.E.; Martinez, P.P.; Mahmud, A.; Marquet, P.; Buckee, C.; Santillana, M. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *medRxiv* **2021**. [CrossRef]