COVID-19 forecasting with deep learning: a distressing survey

L. Gutiérrez, R. de Medrano and J.L. Aznarte

***Abstract*—Building on the success of deep learning techniques in all sorts of classification and regression tasks, in the wake of the COVID-19 pandemic many researchers turned their tools and expertise to the task of predicting the evolution of the infection worldwide. This praiseworthy effort, based on a strong will to help, produced a panoply of models and applications aimed at helping health institutions to plan and decide on the mitigation measures that could control the spread of the pandemic, through forecasting the disease main indicators for public health.**

However, as we show in this paper, this emergency research endeavour has not necessarily been in line with common quality standards in research: it is indeed hard to find papers in which replicability and reproducibility are enabled, lest guaranteed.

After defining a set of quality criteria related to problem definition, dataset management and model identification and evaluation, we studied 96 papers in detail. None of the analysed papers scored positively in all the criteria, while only about one third scored positively in at least half of the defined criteria. These results show that, in the present case, emergency research has been prone to leave behind some of the basic requirements for quality scientific labour.

I. Introduction

Since the World Health Organisation proclaimed the COVID-19 outbreak as a pandemic in March, 2020[1], the spread of the disease has followed certain patterns based on dynamic transmission of the epidemic over time and exhibited a clearly non-linear behaviour. To try to foresee these patterns, during that period, different epidemiological models have been proposed. These models can be split into two wide categories: data-driven statistical models and classical mechanistic models based on epidemiological principles.

The classical epidemiological approach is based on developing compartmental or susceptible–infected–removed (SIR)-like models, which offer a clear epidemiological interpretation. However, predicting with them is sometimes difficult due to strong parameter value ambiguities, mathematical analytic complexity and the assumption that conditions for propagation will remain unchanged [1]. On the other hand, data-driven models use statistical regression practices and machine learning methods to predict how the disease spreads [2]. These machine learning methods are seen as particularly appropriate for predictions based on existing data, being sometimes considered as more accurate compared to common regression models, as they can capture complex and non-linear patterns in the data.

jlaznarte@dia.uned.es, corresponding author

All three authors are with the Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia – UNED

[1]https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic

Amongst the most successful ML flavours in recent years, deep learning (henceforth DL) has a prominent place in both scientific and newspaper articles. Despite being sometimes branded as a mere 'buzzword', DL models have been successfully applied to many problems, and are praised frequently amongst the most powerful AI tools. DL comprises complex artificial neural networks with many layers, including models such as deep belief networks, convolutional neural networks, auto-encoders, restricted Boltzmann machines, generative adversarial networks and recurrent neural networks, amongst others.

In the relatively short period since the start of the pandemic, many DL applications for COVID-19 forecasting have been presented, and their performance assessed with a wide variety of metrics. However, forecasting is a challenging and specialized task, especially when dealing with small datasets, and, as in any other scientific discipline, urgencies do not usually favour quality. Predictive models must be carefully evaluated not only on their ability to capture historical events but on their exactitude in forecasting future trends, fostering a stronger appreciation of the technology's capabilities and limitations [3]. Furthermore, this evaluation process must be standardized and then validated by the scientific community.

Notwithstanding, as we will show, most of the available applications of DL to COVID-19 forecasting are affected by common flaws. This worrisome fact raises serious concerns about the maturity of the field, its usefulness in the wake of emergencies and the common publish-or-perish academic career scheme, which has indeed been linked to the so-called replication crisis [4].

Concerning the novelty of our approach, there are indeed previous literature reviews which offer general comparisons of existing machine learning techniques applied to COVID-19 diagnosis and prognosis. To the best of our knowledge, none of them covers the prediction of the spread of the pandemic in an exhaustive manner, and none is focused on DL applications, as shown below.

Therefore, our analysis deals specifically with DL techniques applied to forecasting the number of COVID-19 infected cases, focusing on methodological difficulties and typical challenges that researchers confront. A rigorous quality screening is presented to highlight methodological concerns, emphasizing the weaknesses that can lead to issues about reproducibility and replicability of the results. In order to do so in an objective manner, a set of quality criteria is established beforehand, concerning the datasets used as well as the problem and model definition and evaluation. A set of 96 papers has been studied under these criteria: as we will

show, the results are all but flattering.

This document is structured as follows: in Section II the existing literature reviews and state-of-the-art papers about the application of DL techniques to forecasting the COVID-19 spread are summarized. In Section III the methodology employed in our work is explained and a set of quality criteria covering several aspects of the scientific process is defined. In Section IV a selection of papers on which DL approaches are applied to COVID-19 time series forecasting are reviewed in light of the aforementioned criteria, aiming at evaluating the replicability and reproducibility possibilities of each one. Additionally, in Section V, we examine the challenges found, and we discuss the most relevant findings. Section VI concludes with a summary of our findings.

II. Related works

Amidst the huge number of papers published since the start of the pandemic in the field of AI applications, a good number of review papers were already available when we decided to initiate our state of the art review. However, not many of them covered forecasting with DL methods (none of them was devoted exclusively to this issue), and thus our research questions were not really answered in the available literature.

When our work started, up to 11 state-of-the-art review papers concerning AI applications to the different aspects of the COVID-19 pandemic were already published [5–16]. Other review papers were released while we actually performed our analysis and were also considered [17–22]. However, most of these works had a broad-spectrum approach, making the target very general and inconclusive, covering any application of AI conceivable and reviewing only very few publications in each line of work. There are also other review papers [23–33], but their primary targets dealt with different applications (i.e. imaging and diagnostics, management, etc.).

For example, in the early work from [5], the fields of study were divided into i) early warnings and alerts, ii) tracking and prediction, iii) data dashboards, iv) diagnosis and prognosis, v) treatments and cures, and vi) social control. However, in that paper, mostly topical and opinion articles released in blogs and newspapers were cited, plus a pair of diagnosis papers and a couple of articles presenting compartmental epidemiological models used for forecasting. In any case, authors devoted little attention to forecasting with DL techniques.

The focus of [6] was divided between blockchain and AI in general, dividing the works into estimation of virus outbreaks sizes, detection and treatment. Within the scope of DL in forecasting, only one paper [34] was mentioned, emphasizing the lack of unified datasets while highlighting the possibility of developing adaptive AI models for predictions. However, this work is too general to answer our questions.

The study carried out by [7] covered the following fields: i) detection and diagnosis, ii) tracking and predicting the outbreak, iii) 'infodemiology' and 'infoveillance', and iv) biomedicine and pharmacotherapy. Authors propose some use cases and mention challenges and solutions, especially the lack of a standard data set, forcing each model to use its own dataset, and making comparisons difficult. They discuss lessons learned and give some recommendations, like the use official datasets from health authorities, the optimizing of algorithms or the integration with other methods. Again, the scope of this work is too wide, since it covers AI in general and all aspects related to the COVID-19 fight. Due to the colossal work that would be to cover all papers in such a wide field, the selection of papers is quite arbitrary. As a result of this, some of them are just mentioned, but none was particularly analysed, resulting in that only two works [34, 35] were related to forecasting cases using DL approaches, both included here.

In [8], the considered categories were i) quick pandemic alert, ii) tracking and diagnosing cases, iii) pharmacological treatment, and iv) public health interventions. A short set of papers were discussed, but only [36] dealt with forecasting with DL, and their final conclusions were very brief and too general.

From the various models analysed in [9], only six were related to DL, while again barely two of them [37, 38] were related to forecasting the spread or cases, and were just briefly commented. Their main conclusion was indeed brief and open: "there is a need of thorough assessment of these predictive analytic algorithm based on type of question to be answered" (sic).

In a more extensive work [10], data sources, classical TS methods, epidemiological models, forecasting, impact and decision-making tools were analysed. In the forecasting chapter, authors mention machine learning, DL, ARIMA and ensemble approaches. They highlighted [39, 40] for fully connected neural networks and [36, 41–43] for recursive neural networks, while approaches dealing with convolutional networks were all devoted to imaging and signal processing. The main conclusions on DL approaches were about the high amount of data required, the complexity of model hyper-parametrization, and the low interpretability of the results. But as the authors themselves admit, their purpose is just to "highlight effective data-driven methodologies that have been shown to be successful in other contexts and that have potential application in the different steps of the proposed roadmap".

In [11], models were divided in four categories: big data, social media/other communication media data, stochastic theory/mathematical models and data science/machine learning techniques. In the latter category, just two papers [37, 38] were relevant to our subject but they were just concisely mentioned. The main challenges they identified were the lack of quality and quantity in data, over-fitted models, overly clean data with eventual integrity loss, data abundance not always improving accuracy, wrong algorithm and attribute selection leading to misleading results, and model complexity that can affect the overall performance. While these questions are important, they are commonly inherent to every data-driven method. Their main conclusion ("it is important to analyse various forecasting models for COVID-19 to empower allied organizations with more appropriate information possible") justifies by itself the existence of our paper. In any case, the variety as well as the number of models that should be analysed must be higher in order to arrive at any sound conclusions.

In another brief paper [12], a few papers were merely enu-

merated and categorized in i) early detection and diagnosis of the infection ii), monitoring the treatment, iii) contact tracing of the individuals, iv) projection of cases and mortality, v) development of drugs and vaccines, vi) reducing the workload of healthcare workers, and vii) prevention of the disease. From the papers included therein, only [34] was relevant to our subject. With no identified challenge, their conclusions were both wide and general, so very little could be deducted from them.

The divisions in [13] were detection and diagnosis, virology, drug and vaccine development and epidemic. In the latter category, authors dedicated a section to outbreak detection, where a few papers were just described and summarized in a table [34, 36–39, 44–47]. The identified challenges were the lack of large-scale training data and the limited interaction between of computer science and medicine. Still, from this paper it cannot be elucidated which DL methods could be more useful for prediction, or even more, whether DL is useful at all or not.

Deep learning, edge computing and deep transfer learning were the focus of [14]. However, only two of the considered papers [37, 46] were related to our scope. No conclusions could be extracted regarding DL, as its presence was merely testimonial.

In a recent paper [15], only two new citations were added compared to the author's previously mentioned work [5], but they were related to position articles on a blog and a website.

For [16] the main topics were i) screening and treatment, ii) contact tracing, iii) prediction and forecasting, and vi) drugs and vaccination. Only four papers were reviewed for the third category, and only one of them was related to DL techniques. The descriptions and analysis were extensive, including the most important aspects and providing nice explanatory tables. However, the conclusions were brief: "deep learning algorithms [...] have more potential, robust, and advance among the other learning algorithms [while] most of the models are not deployed enough to show their real-world operation" (sic). Nevertheless, the only analysed paper within our scope [41] was insufficient to discard a more exhaustive analysis.

In [17] the domains covered were i) detection and diagnosis, ii) contact tracing, iii) forecasting, iv) vaccine development. While this paper is quite exhaustive about the role of AI in computerized tomography (CT) scans and X-Ray images, it only analyses one paper [41] in the forecasting field.

The central subjects for [18] were i) diagnosis using radiography images, ii) diagnosis using respiratory and coughing wave data, iii) severity and survival-mortality assessment, iv) outbreak forecasting models, v) virion sequence formation and drug discovery models. In the forecasting area they provided a list of 27 papers, 12 of them related to DL [35–37, 42, 44, 46, 48–53]. Unfortunately, only four of those papers were actually analysed, while the rest were just depicted in a table by their main features. The identified challenges were: model precision and reliability impacted by quickly constructed datasets and their limited real-world implementations. The final conclusions were that the utility of AI in predicting outbreak and forecasting the spread of COVID-19 is patent but further research is needed to identify real-world uses of AI for COVID-19.

The classification chosen by another exhaustive paper [19], was i) diagnosis, ii) treatment and vaccines, iii) epidemiology, iv) patient outcome and iv) infodemiology. Authors considered 82 studies out of the 435 retrieved, from which only a few [34, 35, 37, 38, 44, 45, 50, 54] were related to forecasting with DL. They analysed the most interesting aspects of the models, like employed techniques, features of the datasets, applications, and publishing countries. Unfortunately, the models were simply summarized in tables. Authors found that papers reported AI features and results inconsistently: for example, approximately one third of them did not disclose the type of validation or the data size, and a few of them did not even specify the type of AI used, thus hampering replicability.

In [20] the considered areas were i) clinical applications, ii) CT and X-ray image processing, iii) epidemiology, iv) pharmaceutical, v) text processing, vi) understanding the virus, and vii) dataset collection. It is in the epidemiology section where we find an exhaustive collection of papers related to forecasting [39, 55–80]. However, those were just described without any further analysis or criticism. Their main conclusions in our field of study were regarding the size of the data, the way they are collected and the variability of formats of these data, while authors propose global search algorithms for training the networks in order to avoid local optima. While those remarks were complete and sharp, they were given from a quite broad perspective.

In [21], the considered applications were protein and drug development, diagnosis and outcome predictions, epidemiology and 'infodemiology'. In the latter category, we can find some modelling and forecasting papers such as [38, 44, 45, 54]. Authors found that "very few of the reviewed systems have operational maturity" and identified three main issues: the need of open global repositories, the creation of multidisciplinary teams, and the need for open science so that solutions can be shared globally and adapted to other contexts.

By the time of finishing this document, a systematic review of the papers covering image-related DL techniques applied to COVID-19 was released [81]. Since time series forecasting and image recognition are entirely different fields, the purpose of that work might be in a similar line to the conclusions extracted here, but there is no overlap.

Summarizing, most of the analysed review papers focus on all the fields related to the fight against COVID-19, or on the variety of AI disciplines available, but, in particular, none is precisely focused on forecasting with DL. As we have seen, the main trend is to describe the methods employed and highlight the overall challenges in a general manner. The lower number of forecasting methods analysed, as well as the predominance of compartmental models, traditional statistical techniques, and conventional machine learning methods versus DL ones, adds up enough evidence to justify the existence of this document.

III. Methodology

A. Paper Selection

As stated above, we focus on DL forecasting approaches related to the prediction of the COVID-19 pandemic outbreak.

Thus, this review focus on works that are using artificial neural networks and more precisely DL techniques to forecast the spread of the COVID-19 pandemic.

According to the European Centre for Disease Prevention and Control (ECDC) [82], the most accurate indicators of epidemic intensity are the absolute number of newly confirmed cases and their notification rate per 100,000 population. Hence the output of the considered models must be, at least but not limited to, the number of newly confirmed cases. This indicator is usually complemented with the number of total cases, active cases, recovered cases, deceases, and other measures. On the other hand, the inputs will usually be the number of total recorded (confirmed) cases, but they may be accompanied by the recorded number of total cases, active cases, recovered cases, deceases etc.

For the sake of simplicity and standardisation, the models proposed in the reviewed papers were sorted amongst one of the following categories:

- Artificial neural networks (ANN) [83, 84]: multilayer perceptron [85] (MLP) or feed-forward multilayer neural network (FFNN) [86], Autoregressive Networks [87], Auto-encoders [88, 89], Adaptive Networks [90].
- Recurrent Neural Networks (RNN) [91]: Long Short-Term Memory units (LSTM) [92], Gated Recurrent Units (GRU) [93], Bidirectional RNNs (BRNN) [94], Multi-head attention (ATT) [95].
- Convolutional Neural Networks (CNN) [96].
- Extreme learning machines (ELM) [97].
- Ensemble methods.

Other denominations, such as Deep Neural Networks (DNN) [98], could have been ascertained into any of the previous categories, being the 'deep' characteristic an arbitrary boundary.

We consider studies published in English between 1 January 2020 and 10 May 2021, including conference proceedings, dissertations, peer-reviewed articles, and preprints. Any other publications such as blogs, topical papers, opinion essays or commentaries, were discarded. We did not contemplate any limitations regarding the origin of publication, study design, or outcomes. Out of the several hundred titles retrieved through a systematic search and independent screening by titles and abstracts, 97 studies were retained for full text reading. The selected ones were crosschecked with the cited bibliography of the reviews already discussed in the previous sections, resulting in the addition of a few more papers to our study.

The search was performed in well-known databases like ResearchGate, SpringerLink, Elsevier, IEEE Xplore, ACM Digital Library, arXiv, medRxiv, and Google Scholar, excluding terms like 'sentiment', 'drug', 'X-Ray', 'Computer Tomography', 'Imaging', 'RNA' etc. or any of its variants. For an example of the queries used, see Figure 1.

B. Assessment Criteria

In order to assess the quality of every considered paper, following the lead of previous meta-analysis as explained in Section II, in order to make our analysis as fair as possible, we need to define a set of criteria. These criteria or key quality indicators must represent concrete, measurable features of the papers, and must be as objective as possible. In this section, the set of key quality indicators that have been chosen for comparison of the selected papers is described. These indicators aim to assess the information that quality papers must provide to the reader, in order to evidence the robustness of the model, to elucidate the conditions of the study, to explain how uncertainty is managed, and to guarantee future replicability.

In relation to concerns expressed in previous works about how AI, ML or DL are applied in the field of medicine [99–104], our work is rooted in existing paper evaluation frameworks [105, 106], which we have adapted to the specific needs of the chosen field. Despite the sharp and useful recommendations from [104], it is mainly focused on clinical trials, and thus its main purpose is to be a guideline for developing studies rather than a literature review. From the list of items described in [105], while some of them are common to any kind of AI study, and hence applicable to our problem, the majority is exclusively applicable to medical imaging. Therefore, while the medical imaging items were not considered here, the general principles were assumed in order to elaborate our list of criteria. Finally, specific criteria related to forecasting were also added to the list.

Below we describe the set of considered criteria, which are classified according to their focus.

1) Criteria related with the problem description: In any case, to be considered as a quality paper, any article must include a specific and clear description of the problem to be solved, stating the dependent and independent variables that are considered, the area of study, the forecasting horizon, the period of study, and the employed techniques (i.e., type of ANN). Authors should avoid ambiguous assertions like 'predicting the curve', 'forecasting the spread', 'foresee the evolution', etc., favouring clear statements about measurable variables.

1) Object of study. The paper must clearly indicate what is the goal of the study, the type of predictive modelling to be performed, the target variables to be predicted, and the characteristics of the variables which are inherent to the problem description and have a direct effect on the replicability of the experiment: area of study (province, state, region, country), variables to predict (cases, deaths, recoveries), etc.
2) Model identification. The chosen forecasting models must be properly identified and presented, citing previous works in case the models are not new.
3) Forecast horizon. The study must specify the time lag into the future for which forecasts are to be prepared. In the COVID-19 forecasting case, this will vary from short-term forecasting horizons (weeks) to long-term horizons (years) [107]. The chosen forecasting horizon may have a direct impact on the prediction error [108] as well as on the usability of the results.

2) Criteria related with the datasets: Any good paper in this context must contain a clear description of the dataset and the data curation procedures applied, including availability and any transformations in the ETL process. This is especially

Search examples:

COVID forecasting "deep learning" -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph
COVID forecasting Convolutional -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph
COVID forecasting LSTM -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph
COVID forecasting Recurrent -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph
COVID forecasting RNN -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph
COVID forecasting CNN -images -drug -sentiment -RNA -symptoms -X-Ray -CT -Radiograph

Figure 1: Example of some search constraints employed.

important in the COVID-19 forecasting framework since the data are far from consolidated.

4) Data sources. The paper must clearly state the sources of the data, providing links to them and/or depositing the data tables used for modelling in a publicly accessible repository.
5) Features. Variables contained in the dataset (cases, deaths, recoveries, etc.) and the area where the data is circumscribed to (province, state, region, country, hospital) must be properly described in the document.
6) Study interval. The paper must explicitly include the initial and final date for the considered dataset, providing a clear view of the dataset size and the period analysed.
7) Missing data handling. The paper must specify how inconsistent, missing and/or wrong data points were handled.
8) Data preprocessing. How raw data from various sources was converted into a time series must be clearly specified, as well as any use of normalization, rescaling and/or standardization.

3) Criteria related with the model description:

9) Software. The paper must specify the names, version numbers and configuration settings used in any software, libraries, frameworks and packages used in the experiments.
10) Accessibility. The paper must state a publicly accessible repository where the full code of the modelling process can be found, in order to allow replication and a better interpretation of the study.
11) Initialization. The paper must indicate how the initial parameters of the models were fixed, specifying the distribution from which random values were drawn for any randomly initialized parameters, as well as any random seeds necessary. If transfer learning is employed, the source of the starting weights must be clarified, or the weights provided. When there is a combination of random initialization and transfer learning, it must be clear which portions of the model were initialized with which strategies.
12) Topology. The number of layers and how they are connected must be clearly and fully specified in the paper.
13) Activation functions. The paper must specify the number and type of cells on every layer, and the type of activation function selected in every one of them [109].
14) Objective function and optimizer. The paper must precisely describe the function to be optimized, also called the cost function, loss function, or error function in minimization problems [110], as well as the chosen optimizer and how it has been parametrized [111].

4) Criteria related with the model evaluation: Cross-validation and bootstrapping are validation methods that are typically used for the evaluation of model performance or for fine-tuning. Alternatively, hold-out validation may address the internal validity of a model but would not accurately assess its generalizability [99]. Moreover, using hold-out in small datasets may lead to biased predictions, and in that case results will be dependent on how the data is split into train and test sets. Cross-validation can provide a better indication of how well the model will perform on unseen data, as it gives the opportunity to train on multiple train-test splits [112]. An honest validation procedure should reveal the optimism that is associated with the full modelling procedure, since model uncertainty usually is more important for optimism in model performance than parameter uncertainty [113].

Despite statistical testing for calibration is not without pitfalls [114–116], when p-values are reported with sensible precision (i.e., $p = 0.023$, instead of the conventional $p < 0.05$), together with 95% confidence intervals, the consistency between the results obtained and pure chance can be measured, thus providing a better understanding of the results.

15) Validation. The papers must clearly specify how the results were validated (hold-out, cross-validation, rolling validation, etc.) and how data were assigned into training, validation, and testing partitions.
16) Error metrics. The papers must clearly describe the error metrics employed to assess the model's performance and choose appropriate and well-known metrics for forecasting problems [117].
17) Benchmark comparison. The performance of the AI model must be compared against state-of-the-art models and naïve models.
18) Statistical inference. The papers must state what kind of hypothesis tests have been applied in order to decide whether experimental results contain enough information to cast doubt on conventional wisdom.

5) Final score:

19) Final score. Meant as a summary of the set of criteria described above, this score will be computed as the sum of the number of criteria that each paper meets completely and explitly. Only in case of a draw, we will recourse to comparing the number of criteria that are met in an implicit way (see † in the following section), and then those which are just partially met (see ‡below).

Table I: Summary of scores per field: N (no), Y†‡(implicit and partially yes), Y‡(partially yes), Y† (implicitly yes), Y (yes).

	N	Y†‡	Y†	Y‡	Y
Object of Study	0	1	30	5	**60**
Forecast horizon	23	0	0	0	**73**
Data Sources	5	0	0	0	**91**
Features	4	1	14	11	**66**
Dataset Interval	13	0	0	1	**82**
Missing data handling	**80**	0	2	0	14
Data Pre-Processing	**45**	0	0	8	43
Software	**43**	3	1	31	18
Accessibility	**88**	0	0	0	8
Initialization	**76**	0	4	8	8
Topology	19	0	0	1	**76**
Activation Functions	28	0	0	29	**39**
Objective Function & Optimizer	25	0	0	33	**38**
Validation	26	0	0	**57**	13
Error Metrics	14	0	0	0	**82**
Benchmark Comparison	18	0	0	**45**	33
Statistical Inference	**88**	0	0	0	8

IV. Analysed models

At the time of writing, several papers about DL applications to COVID-19 have been retracted [118], in yet another hint to worrisome flaws in the quality of science in emergency times. However, none of them dealt with forecasting except one, which was indeed withdrawn on 10 Nov., 2020 [119], leaving the total amount of considered papers in 96.

All those works were evaluated against each of the criteria defined above. Papers were marked with an "N" when they did not meet the criterion, and with a "Y" when it was fully satisfied. Papers were awarded a "Y with reservations", when criteria were partially met, for example in cases when the required information could be only found implicitly throughout the text (†), or only partially (‡) or both (†‡).

A. Problem Description

To stress the potential novelty of their models, certain authors tend to give imaginative or elaborate names to them, sometimes difficulting the identification. Nevertheless, all the models found in the considered papers were classified according to the model taxonomy detailed in **Section III-A**. Amongst the 95 analysed papers, a total amount of 143 models were employed. The most popular model was LSTM, followed by FFN, while GRU and CNN ranked in 3rd and 4th position (see **Figure 2**).

All the considered papers explicitly state the object of study, albeit with different fortune. For example, 12 of them [37, 41, 47, 48, 59, 62, 69, 120–124] did this only in an implicit way, by distributing the information throughout the text. Only [38] and [125] did this in a partial manner, not mentioning the variables to predict. The information provided by the former was implicit.

Surprisingly, from all the analysed papers, 23 of them did not explicitly state the forecast horizon employed [44, 52–56, 60, 69, 70, 73, 80, 120–122, 125–133], while the rest were found to do so in one way or another.

B. Data

From all the reviewed papers, only 5 failed to state the source of the datasets used [54, 77, 120, 134, 135]. The other

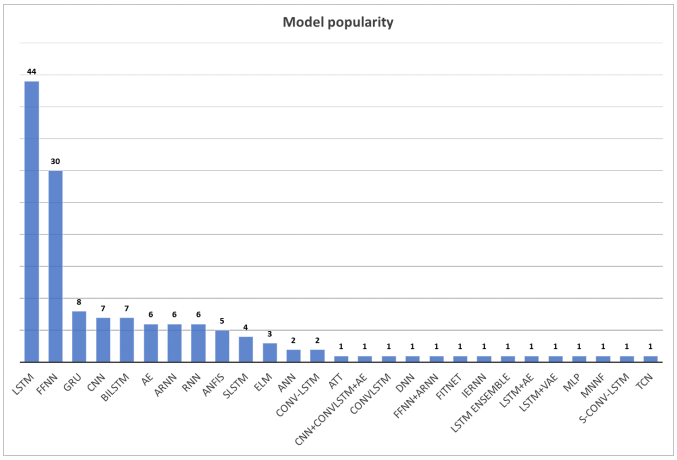

Figure 2: Number of times each type of models has been proposed in the set of considered papers.

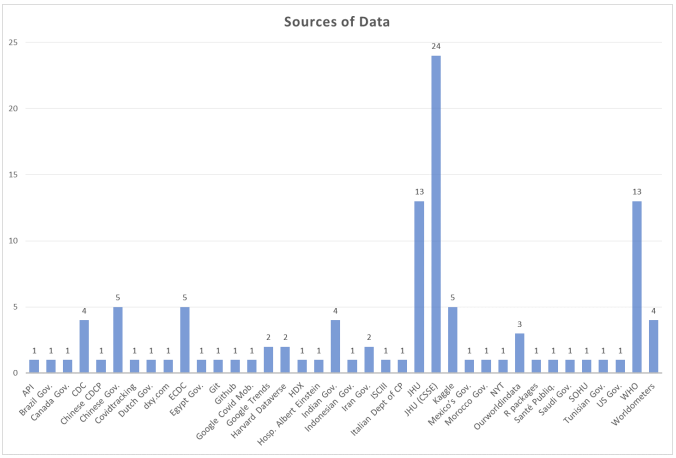

Figure 3: Number of times each source of data has been used in the considered papers.

89 mentioned up to 110 data sources in total. As can be seen in Figure 3, 21 of those employed governmental data, while twelve more used data from local or regional health agencies, such as Centers for Disease Control and Prevention (CDC), European Center for Disease Control and Prevention (ECDC), Chinese Center for Disease Control and Prevention CCDC, etc. The most popular data source was the repository of the Johns Hopkins University (JHU), mentioned 37 times, whereas The Center for Systems Science and Engineering (CSSE) at JHU was specifically mentioned in only 24 of them. The main international organisation mentioned was World Health Organisation (WHO) (30 times), while publicly accessible data repositories were relatively popular: Kaggle was mentioned 5 times, Worldometers 4 times and OurWorldInData 3 times. Only two private repositories were found to be considered: an API with authorised access from [79] and hospital data from [80]

When describing the features present in the dataset, the results are more heterogeneous: 4 papers failed to report any detail at all [77, 133, 134, 136], while 10 of them only described the features partially [39, 41, 50, 52, 54, 57, 126, 131, 137, 138], while [139] did it only in an implicit manner. Other 14 provided this information implicitly and distributed

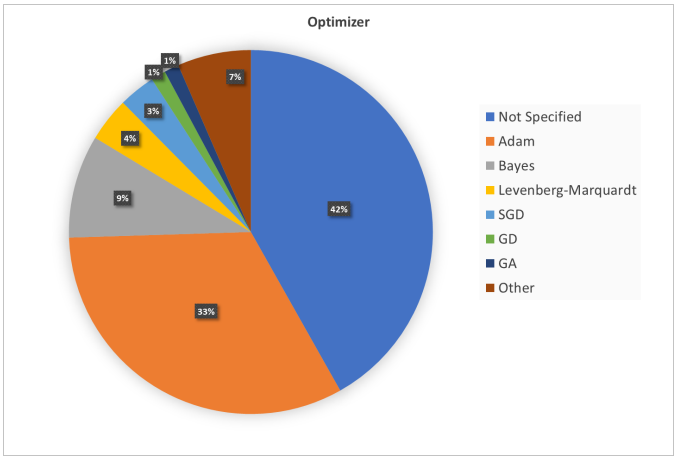

Figure 4: Number of times each optimizer has been used in the set of considered papers.

throughout the whole document [42, 53, 59, 61, 63, 64, 75, 122, 124, 132, 140–143]. The most common reason for this is that the variables are not specified (new cases, accumulated deaths, etc.) or even the area where the data belongs to is not declared.

The considered time interval (and thus the length of the dataset) was not stated in up to 13 of the analysed papers [58, 61, 72, 77, 121, 128, 130, 132, 133, 136, 137, 139, 144]. As can be seen in Appendix **??**, this size varies from only 14 days used by [38] up to two hundred and eighty-four days from [145]. The average size of the dataset was 100.36 days with a standard deviation of 56.19.

With respect to missing data, only 14 studies stated how they dealt with this problem [43, 53, 61, 68, 71, 123, 130, 140, 146–151], while 2 others did this implicitly [22, 135]. The rest of the papers did not mention anything about this aspect, which does not mean that they failed to approach the issue. The lack of missing data might be behind this, but it is always a good practice to explicitly state it. Amongst the papers which dealt with missing data, the approaches are heterogeneous. For example, missing data was just left blank [53], simply eliminated [71, 130, 146], or no missing data found [150]. Others replaced the missing data by the average of five previous and posterior data points [61], or by an average of one week of data [140] or by reversed order values from the sequence [123] or by using linear weighted moving average [149].

Up to 45 of the papers did not mention anything regarding what kind of data pre-processing was applied [35, 39, 42, 44, 46, 49–52, 54–57, 61–65, 67, 69, 71, 77, 80, 121, 124, 130, 132–135, 138, 142, 144–146, 148, 149, 152–158], while 8 only acknowledged this partially [53, 72, 75, 122, 123, 143, 159, 160]. While this does not necessarily mean that data was not pre-processed, these kind of inscrutabilities obviously hinder replicability. The most widespread practice was minimax normalization.

C. Model description

Only 18 works fully documented the software packages and libraries employed, including the versions [37–39, 46, 52, 55, 73, 74, 77, 128, 134, 142, 151, 154, 156, 159, 161, 162]. While [80] did not make it explicit, it was possible to infer it from the source code. Only the name of the software could be implicitly extracted from the repository in [62, 72, 155], which is not a recommended practice. From the rest, 31 papers only revealed the software name [22, 40, 42, 47, 48, 58, 61, 64, 68–71, 75, 79, 122–125, 130–133, 147–149, 157, 160, 163–166], while the others did not include any mention at all. This practice leads to difficulties in reproducibility.

Only 8 papers decided to provide a repository where the full experimental protocol could be accessed [48, 55, 62, 80, 120, 155, 156, 162]. This is an opaque practice that does not favour replicability.

Concerning the initialization of the model, 8 of the articles undisclosed the chosen way for initializing the weights in an unambiguous manner [38, 62, 79, 123, 128, 131, 134, 151], while 4 more mentioned this just in an implicit way [55, 80, 155, 162]. Other 8 decided to provide this information just in a partial way, only declaring that this was done randomly, but without specifying which distribution was used [22, 43, 51, 63, 120, 133, 150, 156]. The rest did not make any mention at all about this aspect.

Regarding network topology, up to 19 of the works failed to mention the number and type of layers from which the network was made up [38, 41, 42, 56, 57, 64, 67, 121, 128, 130, 133, 138, 140, 142, 148, 152, 153, 155, 164]. Only [120] partially did this task, while the rest clearly stated this fact.

From all of the analysed papers, up to 27 failed to explain the number of units employed in each layer and their activation functions [13, 38, 41, 42, 45, 54, 56, 57, 60, 61, 64, 69, 77, 121, 124, 128, 130, 133, 140, 142, 148, 152–155, 161, 164]. Other 29 did it only in a partial way [34, 36, 37, 39, 44, 46, 47, 49, 52, 53, 62, 63, 74, 76, 79, 120, 122, 125, 126, 131, 132, 136, 138, 144, 156, 157, 159, 162, 166], while the rest made this information explicit and complete. The most popular was ReL followed by Sigmoid + Tanh (due to the popularity of LSTMs) and standalone Tanh.

Up to 25 of the works failed to describe the selected objective function, and/or the optimizer applied to minimise it [42, 44, 52–54, 57, 61, 63, 67, 70, 74, 78, 128–130, 134, 138, 140, 142–144, 154–156, 166]. Another 32 succeeded in this task only partially [37, 38, 41, 45–49, 55, 56, 59, 62, 64, 69, 71, 73, 76, 120–125, 127, 131, 136, 139, 145, 148, 152, 153, 164]. As can be seen in Figure 4, the most frequently chosen optimizer was Adam, followed by far by Bayesian optimizer.

D. Evaluation

Concerning evaluation, only 13 of the analized papers informed about a full cross-validation method [37, 43, 46, 55, 63, 71, 80, 124, 125, 140, 151, 159, 166], while other 25 did not mention if any kind of validation was performed at all [34–36, 38, 47, 53, 54, 57–59, 61, 72, 77, 121, 132, 133, 137, 138, 142, 144, 146–148, 158, 165]. This worrisome fact undoubtedly makes the interpretation of the results an exercise of faith. The 57 remaining papers only performed a 1-split hold-out validation, which of course can introduce some bias in their conclusions, especially when the size of the

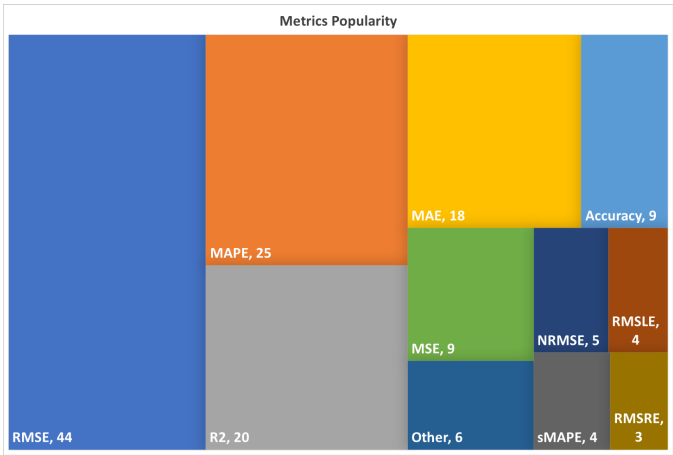

Figure 5: Number of times each error metric has been used in the set of considered papers.

dataset is small. The most common split rate was 80/20 for test and validation, respectively, followed by 90/10 and 75/25. Only [153, 156] applied a 50/50 split, and [73] even went for 40/60. This practice is not advisable, especially with such small datasets, as models will find more difficulties to learn the general principles and will show a poor validation and test performance.

Unfortunately, up to 14 papers failed to provide any error metric at all [36, 50, 56, 57, 62, 64, 67, 77, 121, 122, 126, 129, 142, 144]. Furthermore, while the problem at hand is clearly a regression one, 8 of the studies employed only metrics for classification, making difficult to understand how far the predictions were from the actual values [80, 136, 139–141, 148, 154, 157]. Particularly, [136, 139] used an own formula in an effort to 'adapt' accuracy metric to prediction problems. Another five articles used a mix of classification and regression metrics, leaving some room for comparisons [41, 58, 125, 128, 145]. The rest provided only regression metrics. From Figure 5 it is clear that the most common metric was RMSE, followed by MAPE, R^2 and MAE, while up to 6% of the times, accuracy was chosen.

As mentioned above, comparison with naive and state-of-the-art models is key to prove the goodness of any forecast attempt. Of all the analized papers, 18 papers did not include any kind of benchmark comparison against any other model [34, 41, 42, 51, 53, 55–57, 62, 66, 77, 129, 130, 134, 141, 158, 163, 164]. Another 31 only compared their proposals against complex algorithms, assuming that all of them are thus better than basic persistence or random approaches, which may lead to problematic conclusions [22, 37, 38, 46, 47, 59, 61, 63, 64, 67, 70, 75, 76, 78, 126–128, 131, 132, 136, 138, 139, 142, 143, 145, 151–155, 157].

Less than 10% of the papers reported the application of some kind of statistical inference to their results, and thus, for the rest, it is difficult to assess that the true gain of the model is not due to simply chance [22, 46, 65, 76, 127, 133, 140, 155].

Regarding confidence intervals, only 18 papers [22, 36, 38, 41, 42, 57, 62, 65, 67, 69, 76, 120, 123, 140, 142, 144, 145, 152] employed them to communicate their results, while [156] mentioned this during the training phase only. From those ones mentioned, solely a few of them [57, 65, 123, 144] employed the intervals for accompanying the numerical results, while the rest only applied them for the charts. As an example, [38, 42] only used them for only one out of the several curves provided.

In particular, only 11 of those [36, 42, 62, 65, 76, 123, 140, 142, 144, 145, 152] provided a 95% confidence interval, while [22, 57] employed a threshold of 80% for their uncertainty intervals, but not in the article, but in a website that supports their paper.

Some singular practices were found, for example in [148], where predictions were made with ±50% of the predicted value, and some of the charts depicted an interval whose level of confidence was not defined. On the other hand, [132] provided the metric values with their mean and their variance, which at least provide some additional information about the fitness of the model. In an attempt to capture uncertainty, in [136, 139] metrics were delivered for different error margins (from 0.05 to 0.5, in steps of 0.05).

Finally, [38, 41, 42, 67, 69, 120] did not mention any numerical indication for the confidence threshold. This practice in particular, together with the absence of any confidence intervals at all, makes more difficult to interpret the uncertainty in their predictions, as the estimated probability of capturing the truth is ambiguous. The rest did not employ any kind of confidence interval, or at least, failed to mention it.

E. Final score

After applying all the criteria, only a maigre 35 of the 96 studied papers scores in *at least half* of the fields. The best score overall was 15 over 17, obtained by [151], only failing in the statistical inference and the accessibility fields, as can be seen in Figure 6. It is followed by [22, 43], both with a positive score in 12 of the fields, and both failing in accessibility. But [22] provides information about how missing data was handled in an implicit way, while [43] totally fails in managing statistical inference and in providing information about the software.

On the other side of the ranking, [121] scores only in the data source and features description, while the object of study is only available implicitly throughout the whole document.

V. Discussion and Recommendations

The outstanding efforts to model and forecast the COVID-19 pandemic using deep learning techniques are obvious and should be praised. Nonetheless, the predominance of methodological and reporting insufficiencies has been also patent throughout our analysis. In fact, none of the papers fulfills all of the proposed quality criteria. Remarkably, [22] was the only one failing in a single criterion, followed by [123, 151, 162] which failed in only two of them. The paper fulfilling more of the criteria without any reservations is [151].

As can be seen in Figure 7, the most common weaknesses are related to the lack of application of statistical inference (in 87 articles), poor definition of the experiments (again

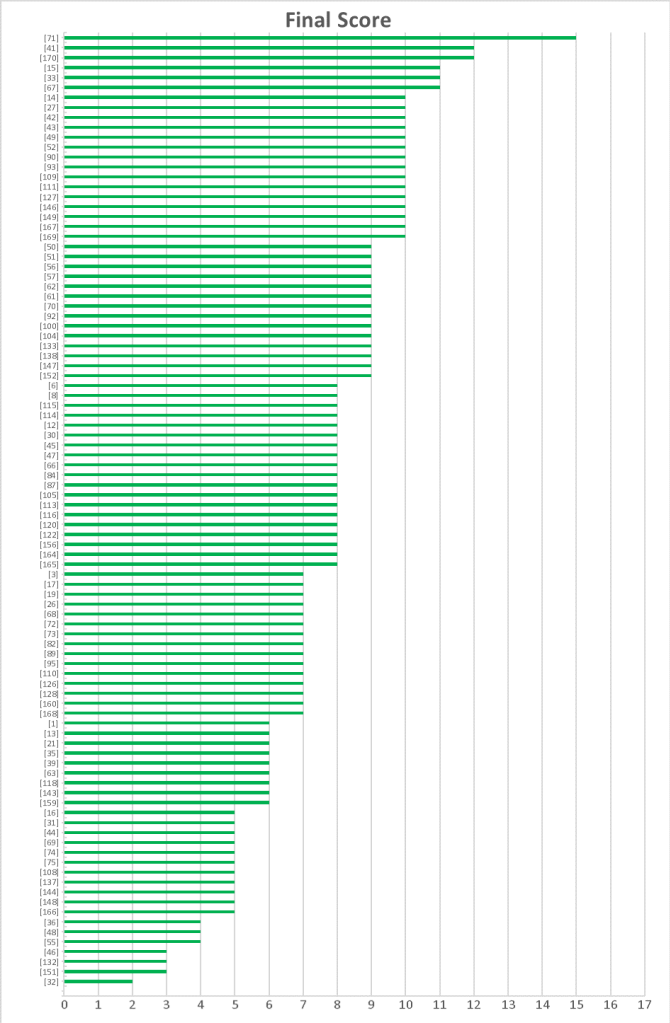

Figure 6: Final scores recorded for the considered papers.

in 87 of the papers), missing data handling (in 79 of the works), missing model initialization details (in 76 papers), no information on data pre-processing (in 44 of them) and lack of software information (in 43 of the articles). These issues may lead to excessively enthusiastic performance estimations and reduced replicability.

When dealing with criteria related to the problem definition, our findings reveal that sometimes it is not clear what the target of a paper is due to the use of ambiguous terms or incomplete assertions, such as "predict COVID-19 infection" or "forecast COVID-19 outbreak". The enunciation of a forecasting problem, in opposition to other types of AI endeavours, should not be necessarily a difficult task. It should be enough to explicitly state the goal of the study, the target variables (i.e., "number of COVID-19 confirmed cases"), the region the predictions are being made for (i.e., "China", "Emilia-Romagna" or "Hospital Albert Einstein"), the forecasting horizon (i.e., "in the next ten days") and the model employed (i.e., "a stacked LSTM model"). So, simple and clear statements explicitly establishing these factors should be a requirement for any paper describing such an application.

Concerning the data-related criteria, the small volumes of COVID-19 data, the different dataset sizes, the diverse kind of variables collected within the datasets, and the way this data is collected by the different organizations and governments remains a huge challenge for an accurate model comparison. We agree with [7] in suggesting that the use of big collaboratively and high-quality datasets provided by governments and healthcare organizations (i.e., WHO, CDC, JHU, etc.) may help to overcome this issue. The surveillance on the quality of the aggregated data by renowned organizations can help to avoid 'retrodden' datasets and may reduce over-fitting, derived from the fact that the community is focused on outperforming benchmarks on a single public dataset.

However, in our opinion, there are no excuses for not explicitly stating the data sources, for example, as well as a clear description of the variables and the intervals considered, or the decisions taken about missing data or the pre-processing stages.

With respect to model description, in a research environment in which open science is becoming more and more encouraged, and for the sake of interpretability and replicability, it is common sense to reveal as much details from the model as possible, so the experiments can be reproduced, and models can be compared to future research.

The disclosure of the software packages, frameworks, and libraries employed, as well as its versions, can certainly enhance the understanding of the performance and conclusions derived from any experiment, while enabling replicability.

Similarly, when dealing with neural networks, revealing the number of layers, number of units, and the activation functions, together with the objective and optimizer function, becomes essential to understand the developed model and its eventual advantages and drawbacks.

Another potential source of obscurity is the randomization of the weights [137], being one of the main sources of stochasticity of the model. Unveiling the distribution from which the weights are being initialized, as well as the employed seeds, is crucial in enabling the reproducibility of any investigation in this field.

Finally, access to the source code and the original dataset employed enhances the comprehension of the model itself and eases the endeavour of repeating someone else's experiments. In such sense, making the complete experiment framework available in a public repository is a practice that boosts the progress of science, especially in these challenging times.

Regarding the evaluation of the proposals, there is no unique appropriate metric for model errors. Using RMSE leads to large errors having a relatively greater influence on the total compared to the smaller ones [167]. This makes MAE better for discriminating among models. Despite its robustness against outliers, MAE is more sensitive to variance, fluctuating its value between several errors sets with the same RMSE [168].

RMSE might be selected to minimize cost function because it helps to calculate the gradient of absolute errors. It is known that with a low number of samples (i.e. 100), giving the values of the errors themselves is probably better than any statistics. Otherwise, large outliers might be excluded from the RMSE

Ref	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Final Score
[1]	Y†	Y	Y	Y†	Y	N	N	Y‡	N	N	Y	N	Y‡	Y	Y	Y‡	N	6
[3]	Y†	Y	Y	Y	Y	N	N	Y‡	N	N	Y	Y‡	Y	Y‡	Y‡	Y	N	7
[6]	Y‡	N	Y	Y	Y	N	N	Y†	Y	Y†	Y	Y	Y	Y	Y‡	Y‡	N	8
[8]	Y†	N	Y	Y	Y	N	Y	Y‡	N	N	Y	Y	N	Y‡	Y	Y	N	8
[12]	Y	N	Y	Y‡	Y	N	Y	Y‡	N	Y	Y	Y‡	Y‡	Y‡	Y	Y	N	8
[13]	Y‡	Y	Y	Y‡	Y‡	N	Y	N	N	N	Y	N	Y‡	N	Y	Y	N	6
[14]	Y	Y	Y	Y	Y	Y	N	Y‡	N	N	Y	Y	Y	Y‡	Y	Y‡	N	10
[15]	Y	Y	Y	Y	Y	N	N	Y	N	N	Y	Y‡	Y‡	Y	Y	Y	Y	11
[16]	Y	Y	Y	Y†	Y	N	N	Y‡	N	N	N	N	Y‡	Y‡	N	Y	N	5
[17]	Y	Y	Y	Y	Y	N	Y	Y‡	N	N	N	N	Y‡	Y‡	Y	N	N	7
[19]	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	Y‡	Y‡	Y	Y	N	7
[21]	Y‡	N	Y	Y	Y	N	Y	Y‡	N	N	Y	Y‡	Y‡	Y	Y†	Y‡	N	6
[26]	Y†	Y	Y	Y	Y	N	N	Y†‡	Y	Y	Y	Y‡	Y‡	Y‡	N	N	N	7
[27]	Y	N	Y	Y	Y	N	N	Y	Y	Y†	Y	Y	Y‡	Y	Y	N	N	10
[30]	Y	Y	N	N	Y	N	N	Y	N	Y	Y	Y	N	Y‡	Y	N	N	8
[31]	Y†	N	Y	Y†	N	N	N	Y‡	N	N	Y	Y‡	Y	N	Y	Y	N	5
[32]	Y†	N	Y	Y	N	N	N	N	N	N	N	N	Y‡	N	N	Y‡	N	2
[33]	Y	Y	Y	Y	Y	Y	Y	Y‡	N	N	Y	Y	Y	Y‡	Y	Y‡	N	11
[35]	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	Y‡	N	Y	N	6
[36]	Y†	Y	Y	Y‡	Y	N	Y	N	N	N	N	N	Y‡	Y‡	Y†	N	N	4
[39]	Y†	Y	Y	Y‡	Y	N	N	N	N	N	Y	Y	Y	Y‡	N	Y‡	N	6
[41]	Y	Y	Y	Y	Y	Y†	Y	Y‡	N	Y‡	Y	Y	Y	Y‡	Y	Y	Y	12
[42]	Y	Y	Y	Y	Y	Y	N	N	N	N	Y	Y	Y	N	Y	Y‡	N	10
[43]	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	Y	Y	Y‡	Y	N	N	10
[44]	Y	Y	Y	Y	N	N	N	N	N	N	Y	Y‡	N	N	N	Y‡	N	5
[45]	Y†	Y	Y	Y	N	N	Y	Y‡	N	N	Y	Y	Y	N	Y	Y‡	N	8
[46]	Y	N	N	Y‡	Y	N	N	N	N	N	Y	N	N	N	Y†	Y‡	N	3
[47]	Y	Y	Y	Y	Y	N	N	N	N	N	Y	Y	Y‡	Y‡	Y†	Y	N	8
[48]	Y	N	Y	N	N	N	N	Y‡	N	Y‡	N	N	Y	N	Y	Y	N	4
[49]	Y†	Y	Y	Y	Y	N	Y	Y	N	N	Y	Y‡	Y‡	Y	Y	Y	N	10
[50]	Y†‡	Y	Y	Y	Y	N	Y	Y	N	Y	N	N	Y‡	N	Y	Y	N	9
[51]	Y†	Y	Y	Y	Y	N	Y	Y‡	N	N	Y	Y	Y	N	Y	Y‡	N	9
[52]	Y	Y	Y	Y	Y	N	Y	Y‡	N	N	Y	Y	Y	Y‡	Y	N	N	10
[55]	Y	N	Y	Y	Y	N	N	N	N	N	N	N	Y‡	Y‡	N	N	N	4
[56]	Y‡	Y	Y	Y	Y	N	Y	N	N	N	Y	Y	N	Y‡	Y	Y	N	9
[57]	Y	Y	Y	Y	Y	N	N	Y	Y	Y‡	Y	Y‡	N	Y‡	Y	Y‡	N	9
[61]	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	Y‡	Y	N	Y	N	N	9
[62]	Y	Y	Y	Y	Y	N	N	N	N	N	Y	Y	Y	N	Y	Y‡	N	9
[63]	Y	N	Y	Y	Y	N	N	N	N	N	Y	Y‡	N	Y‡	Y	Y‡	N	6
[66]	Y‡	Y	Y	Y	Y	N	N	N	N	Y‡	Y	Y	Y	Y‡	Y	N	N	8
[67]	Y	Y	Y	Y	Y	Y	Y	N	N	Y‡	Y	Y	Y	Y‡	Y	Y‡	N	11
[68]	Y†	Y	Y	Y†	Y	N	N	N	N	Y‡	Y	Y‡	N	Y	Y	Y	N	7
[69]	Y†	N	Y	Y‡	Y	N	N	Y	N	N	Y	Y‡	N	Y‡	Y	Y‡	N	5
[70]	Y	Y	Y	Y†	Y	N	Y‡	Y‡	N	N	Y	Y	Y	Y‡	Y	Y	N	9
[71]	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y‡	Y	Y	15
[72]	Y	Y	Y	Y†	N	Y	N	Y‡	N	N	Y	N	N	N	Y	Y	N	7
[73]	Y†	Y	Y	Y†	Y	N	Y‡	N	N	N	Y	Y	N	Y‡	Y	Y	N	7
[74]	Y†	Y	Y	Y‡	Y	N	N	N	N	N	N	Y‡	N	N	Y	Y	N	5
[75]	Y†	N	Y	Y†	Y	Y	Y‡	N	N	N	Y	Y‡	N	N	Y	N	N	5
[82]	Y†	Y	N	Y	Y	Y†	N	N	N	N	Y	Y	Y	Y‡	Y	Y‡	N	7
[84]	Y	Y	Y	Y	Y	N	N	Y	N	N	Y	N	N	Y‡	Y‡	Y	N	8
[87]	Y	Y	Y	Y	Y	N	N	Y‡	N	N	Y	Y‡	N	Y	Y	Y‡	N	8
[89]	Y†	Y	Y	Y	N	N	Y‡	Y†‡	N	N	Y	Y	Y	N	Y	Y‡	N	7
[90]	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	Y‡	Y‡	Y‡	Y	Y	Y	10
[92]	Y†	Y	Y	Y	Y	Y	Y‡	Y‡	N	Y	Y	Y	Y‡	Y‡	Y	Y‡	N	9
[93]	Y	N	Y	Y	Y	N	Y	N	N	N	Y	Y	Y‡	Y‡	Y	Y	Y	10
[95]	Y	N	Y	Y	Y	N	Y	N	N	N	Y	Y	N	Y‡	N	N	N	7
[100]	Y	Y	Y	Y	Y	N	N	Y†‡	Y	Y†	N	N	N	Y‡	Y	Y	Y	9
[104]	Y†	Y	Y	Y	Y	N	Y	Y‡	Y	N	Y	Y	Y‡	Y‡	Y	Y‡	N	9
[105]	Y	Y	Y	Y‡	N	N	Y	N	N	N	Y	Y	Y	N	Y	Y‡	N	8
[108]	Y†	N	Y	Y	Y	N	N	Y‡	N	N	Y	N	Y‡	Y‡	Y	Y‡	N	5
[109]	Y	Y	Y	Y	Y	Y	N	Y‡	N	N	Y	Y	Y‡	Y	Y	Y‡	N	10
[110]	Y	N	Y	Y‡	Y	N	Y	N	N	N	Y	Y‡	Y	Y‡	N	Y	N	7
[111]	Y†	Y	Y	Y	Y	Y	Y	Y‡	N	N	Y	Y	Y	N	Y	Y‡	N	10
[113]	Y†	Y	Y	Y	Y	N	Y	Y‡	N	N	Y	Y‡	Y‡	N	Y	Y	N	8
[114]	Y	N	Y	Y	Y	N	Y	N	N	N	Y	N	Y	Y‡	Y	Y‡	N	8
[115]	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	N	Y‡	Y‡	Y	Y‡	N	8
[116]	Y	Y	Y	Y†	Y	N	Y	N	N	N	Y	Y	Y	Y‡	Y‡	N	N	8
[118]	Y	Y	Y	Y	Y	Y	N	Y‡	N	N	N	N	Y‡	N	Y‡	Y‡	N	6
[120]	Y	Y	Y	Y‡	Y	N	N	Y	N	N	Y	Y‡	Y	Y‡	Y	Y‡	N	8
[122]	Y	Y	Y	Y†	Y	Y	Y	N	N	N	N	N	N	Y	Y‡	Y‡	Y	8
[126]	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	Y‡	Y‡	Y	Y	N	7
[127]	Y	Y	Y	Y	Y	N	Y	Y	N	N	Y	N	Y	Y‡	Y	Y‡	N	10
[128]	Y	N	Y	Y	N	N	Y	Y	N	Y	N	N	N	Y‡	Y†	Y	N	7
[132]	Y†	N	Y	Y†	Y	N	Y‡	Y‡	N	N	Y	Y‡	Y‡	Y‡	N	Y‡	N	3
[133]	Y	Y	Y	Y	Y	N	Y‡	Y‡	N	N	Y	Y	Y	Y‡	Y	Y‡	N	9
[137]	Y	N	Y	Y	N	Y	N	Y‡	N	N	N	N	N	Y‡	Y	N	N	5
[138]	Y†	Y	Y	Y	Y	N	Y	Y‡	N	Y	Y	Y‡	Y	Y‡	Y	Y‡	N	9
[143]	Y	Y	Y	Y†	Y	N	N	Y	N	N	N	N	N	N	N	Y	N	6
[144]	Y	Y	N	N	N	N	N	Y	N	N	Y	N	Y	N	N	N	N	5
[146]	Y	Y	Y	Y	Y	N	Y‡	Y	N	N	Y	Y‡	Y	Y	Y	Y‡	N	10
[147]	Y	Y	Y	Y	Y	N	Y	Y	N	N	Y	Y‡	N	Y‡	Y	Y‡	N	9
[148]	Y	Y	Y	Y†	Y	N	N	Y‡	N	N	N	N	N	Y‡	Y	N	N	5
[149]	Y	Y	Y	Y	Y	N	Y	Y‡	N	N	Y	Y	Y	Y‡	Y	Y‡	N	10
[151]	Y†	Y	Y	Y‡	Y	N	N	N	N	N	N	N	N	N	N	N	N	3
[152]	Y	N	Y	Y	Y	N	Y	Y	N	N	Y	Y	Y‡	Y‡	Y	Y‡	N	9
[156]	Y†	Y	Y	Y	Y	N	N	N	N	N	Y	Y	Y	N	Y	N	N	8
[159]	Y	Y	Y	N	N	N	Y	N	N	N	Y	Y‡	Y‡	Y‡	Y‡	Y	N	6
[160]	Y	Y	Y	Y†‡	N	N	Y	N	N	N	Y	Y	Y‡	Y‡	Y‡	Y	N	7
[164]	Y†	Y	Y	Y†	Y	N	Y	N	N	N	Y	Y	Y‡	N	Y	Y	N	8
[165]	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	Y‡	Y	N	N	Y‡	N	8
[166]	Y†	N	N	Y	Y	N	Y	N	Y	Y‡	Y‡	Y‡	Y‡	Y‡	Y	Y‡	N	5
[167]	Y	Y	Y	Y	Y	N	N	N	N	N	Y	Y	Y	Y‡	Y	Y‡	Y	10
[168]	Y	Y	Y	Y	Y	N	N	N	N	N	Y	Y‡	Y‡	Y‡	Y	Y‡	N	7
[169]	Y†	Y	Y	Y	Y	N	Y	Y	Y	Y†	Y	Y‡	Y	Y‡	Y	Y‡	N	12
[170]	Y	Y	Y	Y	Y	Y	Y	N	N	Y‡	Y	Y	Y	Y	Y	Y‡	N	12

Figure 7: Summary of scores that papers received in each criteria: the column titles corresponds to the item numbers used in Section III-B. In column 16, † refers to the miss-use of classification metrics together with prediction metrics by the authors, and the ‡ mark highlights when only classification metrics are employed.

calculation [168]. But when having more samples, RMSE can reconstruct error distribution, with a standard deviation lower than 5%. Inconsistency in comparing RMSEs from different studies is not due to error-scale variance alone [167].

Choosing one single metric removes a lot of information, so an error distribution should always be provided. MAE is suitable for uniformly distributed errors, while RMSE is better when errors follow a normal distribution, which is the most common case. For other kinds of distributions, more statistics, such as mean, variance, skewness, and flatness, should be provided [168]. So to better depict the model behaviour, the best recommendation might be to provide the full probability distribution of the error, or at least several standard metrics which facilitate comparisons.

When reporting results, including a statistical significance test with the p-value obtained (rather than just simply passing or not the famous 0.05 threshold) and/or confidence intervals to reflect the uncertainty in the forecast is strongly recommended. But also, in order to test if a proposal makes sense or not, it is essential to use simple reference models as baselines, such as naïve or persistence forecasting models. It is very common to see how the interest that has been put in developing the proposed model is inversely proportional to the effort invested in the benchmarking models. This may lead to overoptimistic interpretations of the results, as well as an unrealistic idea of the real capabilities of the developed model.

MonteCarlo stochastic simulations seem to be a suitable practice for modeling infectious outbreaks that change across geographical areas and through time [37]. Also, hyperparameter search and sensitivity tests are strongly recommended.

According to the American Statistical Association (ASA) [169], a study is reproducible if one can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study. On the other hand, replicability is the possibility of repeating an entire experiment, independently of the original investigator and without the use of original data (and generally using the same methods).

Although it might be argued that full replicability is theoretically not achievable, a clear description of the methods, models, materials, procedures, metrics, and other variables involved in the study would facilitate it. A clear description of the dataset, data pre-processing, and missing data handling is essential. A description of the statistical inference decisions made and whether the study is exploratory or confirmatory, as well as discussion of the expected constraints for generality, uncertainty of the measurements, results, and inferences are definitely helpful.

Furthermore, while the easiest way to replicate an experiment in DL is to count with the full source code and the original dataset employed, a potential opacity might occur when publicly available datasets or code are being updated. Therefore, it is also advisable to keep track of specific cached versions of datasets and code, so those can be correctly referenced. Many public repository sites provide tools to make this task much easier. These practices are also enabling scientific reproducibility, speeding up future discoveries in any discipline.

VI. Conclusions

In this systematic review, current deep learning literature for COVID-19 forecasting has been considered. We focused on evaluating a set of papers, underlining the quality flaws of the methods employed and the reproducibility and replicability issues.

After establishing a set of minimum quality indicators, it has been observed that no papers in the reviewed literature currently have documented satisfactorily the methodologies employed for the entire process, failing to follow good practices for developing a reproducible deep learning model. A common pitfall is the lack of a robust cross-validation methodology. There is a lot of room for improvement in model comparison against naïve or persistence baselines, as well as the extended use of any kind of statistical inference, to minimally discard any possibility of changes in the results. The different kinds of error metrics presented in the analyzed papers, the variety of forecast periods, and the different kinds of variables to predict, render comparisons difficult.

We agree with [19] that it is vital to develop a standardized reporting protocol and checklists to reduce the poorly conducted COVID-19 studies in favor of more properly conducted studies, and to improve replicability. Finally, some specific recommendations to the researchers for better practices regarding all the analyzed criteria have been provided.

References

[1] M. Castro *et al.*, "The turning point and end of an expanding epidemic cannot be precisely forecast," en, *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 26 190–26 196, Oct. 20, 2020, publisher: National Academy of Sciences section: Biological Sciences PMID: 33004629.

[2] A. Adiga *et al.*, "Mathematical models for covid-19 pandemic: A comparative analysis," en, *Journal of the Indian Institute of Science*, vol. 100, no. 4, pp. 793–807, Oct. 1, 2020, Company: Springer Distributor: Springer Institution: Springer Label: Springer number: 4 publisher: Springer India.

[3] D. Lazer *et al.*, "The parable of google flu: Traps in big data analysis," en, *Science*, vol. 343, no. 6176, pp. 1203–1205, Mar. 14, 2014.

[4] J. W. Schooler, "Metascience could rescue the "replication crisis"," *Nature*, vol. 515, no. 7525, pp. 9–9, Nov. 2014.

[5] W. Naudé, "Artificial intelligence against covid-19: An early review," en, Social Science Research Network, Rochester, NY, Tech. Rep., Apr. 6, 2020, [Online; accessed 2021-01-24].

[6] D. C. Nguyen *et al.*, "Blockchain and ai-based solutions to combat coronavirus (covid-19)-like epidemics: A survey," en, *Preprint*, Apr. 19, 2020, publisher: Preprints.

[7] Q. Pham *et al.*, "Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 130 820–130 839, Apr. 21, 2020, event: IEEE Access.

[8] N. L. Bragazzi *et al.*, "How big data and artificial intelligence can help better manage the covid-19 pandemic," eng, *International Journal of Environmental Research and Public Health*, vol. 17, no. 9, May 2, 2020, PMID: 32370204 PMCID: PMC7246824.

[9] P. N. Mahalle *et al.*, "Data analytics: Covid-19 prediction using multimodal data," en, *Preprint*, May 14, 2020, publisher: Preprints.

[10] T. Alamo *et al.*, "Data-driven methods to monitor, model, forecast and control covid-19 pandemic: Leveraging data science, epidemiology and control theory," *arXiv:2006.01731 [physics, q-bio]*, Jun. 10, 2020, arXiv: 2006.01731.

[11] G. R. Shinde *et al.*, "Forecasting models for coronavirus disease (covid-19): A survey of the state-of-the-art," en, *SN Computer Science*, vol. 1, no. 4, p. 197, Jun. 11, 2020.

[12] R. Vaishya *et al.*, "Artificial intelligence (ai) applications for covid-19 pandemic," en, *Diabetes & Metabolic Syndrome: Clinical ResearchReviews*, vol. 14, no. 4, pp. 337–339, Jul. 1, 2020.

[13] J. Chen *et al.*, "A survey on applications of artificial intelligence in fighting against covid-19," *arXiv:2007.02202 [cs, q-bio]*, Jul. 4, 2020, arXiv: 2007.02202.

[14] A. Sufian *et al.*, "A survey on deep transfer learning to edge computing for mitigating the covid-19 pandemic," *Journal of Systems Architecture*, vol. 108, p. 101 830, Sep. 2020, PMID: null PMCID: PMC7326453.

[15] W. Naudé, "Artificial intelligence vs covid-19: Limitations, constraints and pitfalls," en, *AISOCIETY*, vol. 35, no. 3, pp. 761–765, Sep. 1, 2020.

[16] S. Lalmuanawma *et al.*, "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review," en, *Chaos, Solitons & Fractals*, vol. 139, p. 110 059, Oct. 1, 2020.

[17] P. Sarosh *et al.*, "Artificial intelligence for covid-19 detection – a state-of-the-art review," *arXiv:2012.06310 [cs]*, Nov. 25, 2020, arXiv: 2012.06310.

[18] J. Rasheed *et al.*, "A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the covid-19 pandemic," en, *Chaos, Solitons & Fractals*, vol. 141, p. 110 337, Dec. 1, 2020.

[19] A. Abd-Alrazaq *et al.*, "Artificial intelligence in the fight against covid-19: Scoping review," eng, *Journal of Medical Internet Research*, vol. 22, no. 12, e20756, Dec. 15, 2020, PMID: 33284779 PMCID: PMC7744141.

[20] M.-H. Tayarani N., "Applications of artificial intelligence in battling against covid-19: A literature review," en, *Chaos, Solitons & Fractals*, vol. 142, p. 110 338, Jan. 1, 2021.

[21] J. Bullock *et al.*, "Mapping the landscape of artificial intelligence applications against covid-19," *arXiv:2003.11336 [cs]*, Jan. 11, 2021, arXiv: 2003.11336.

[22] J. Devaraj *et al.*, "Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant?" en, *Results in Physics*, p. 103 817, Jan. 14, 2021.

[23] A. Alimadadi *et al.*, "Artificial intelligence and machine learning to fight covid-19," *Physiological Genomics*, vol. 52, no. 4, pp. 200–202, Mar. 27, 2020, publisher: American Physiological Society.

[24] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," English, *IEEE Reviews in Biomedical Engineering*, 2020, publisher: Institute of Electrical and Electronics Engineers Inc.

[25] Y. Mohamadou *et al.*, "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19," en, *Appl Intell*, 2020, [Online; accessed 2021-01-30].

[26] A. Kumar *et al.*, "A review of modern technologies for tackling covid-19 pandemic," eng, *DiabetesMetabolic Syndrome*, vol. 14, no. 4, pp. 569–573, Aug. 2020, PMID: 32413821 PMCID: PMC7204706.

[27] M. Tsikala Vafea *et al.*, "Emerging technologies for use in the study, diagnosis, and treatment of patients with covid-19," en, *Cellular and Molecular Bioengineering*, vol. 13, no. 4, pp. 249–257, Aug. 1, 2020.

[28] A. S. Ahuja *et al.*, "Artificial intelligence and covid-19: A multidisciplinary approach," en, *Integrative Medicine Research*, Integrative Medicine for COVID-19: Researches and Evidence, vol. 9, no. 3, p. 100 434, Sep. 1, 2020.

[29] A. Shoeibi *et al.*, "Automated detection and forecasting of covid-19 using deep learning techniques: A review," *arXiv:2007.10785 [cs, eess]*, Jul. 27, 2020, arXiv: 2007.10785.

[30] R. Madurai Elavarasan and R. Pugazhendhi, "Restructured society and environment: A review on potential technological strategies to control the covid-19 pandemic," en, *Science of The Total Environment*, vol. 725, p. 138 858, Jul. 10, 2020.

[31] A. Waleed Salehi *et al.*, "Review on machine and deep learning models for the detection and prediction of coronavirus," eng, *Materials Today. Proceedings*, vol. 33, pp. 3896–3901, 2020, PMID: 32837918 PMCID: PMC7309744.

[32] S. Latif *et al.*, "Leveraging data science to combat covid-19: A comprehensive review," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 1, pp. 85–103, Aug. 2020, event: IEEE Transactions on Artificial Intelligence.

[33] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal," *BMJ*, vol. 369, 2020, [Online; accessed 2021-01-30].

[34] Z. Hu *et al.*, "Artificial intelligence forecasting of covid-19 in china," en, *Preprint*, Feb. 17, 2020, [Online; accessed 2020-12-15].

[35] Z. Hu *et al.*, "Forecasting and evaluating intervention of covid-19 in the world," *arXiv:2003.09800 [q-bio]*, Mar. 21, 2020, arXiv: 2003.09800.

[36] Z. Yang *et al.*, "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," eng, *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, Mar. 2020, PMID: 32274081 PMCID: PMC7139011.

[37] S. J. Fong *et al.*, "Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction," en, *Applied Soft Computing*, vol. 93, p. 106 282, Aug. 1, 2020.

[38] S. J. Fong *et al.*, "Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak," en, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. Special Issue on Soft Computing, 2020, [Online; accessed 2020-12-15].

[39] R. M. Rizk-Allah and A. E. Hassanien, "Covid-19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network," *arXiv:2004.05960 [cs, eess]*, Apr. 6, 2020, arXiv: 2004.05960.

[40] O. Torrealba-Rodriguez *et al.*, "Modeling and prediction of covid-19 in mexico applying mathematical and computational models," en, *Chaos, Solitons & Fractals*, vol. 138, p. 109 946, Sep. 1, 2020.

[41] V. K. R. Chimmula and L. Zhang, "Time series forecasting of covid-19 transmission in canada using lstm networks," en, *Chaos, Solitons & Fractals*, vol. 135, p. 109 864, Jun. 1, 2020.

[42] A. Tomar and N. Gupta, "Prediction for the spread of covid-19 in india and effectiveness of preventive measures," en, *Science of The Total Environment*, vol. 728, p. 138 762, Aug. 1, 2020.

[43] N. Zheng *et al.*, "Predicting covid-19 in china using hybrid ai model," *IEEE Transactions on Cybernetics*, 2020.

[44] C.-J. Huang *et al.*, "Multiple-input deep convolutional neural network model for covid-19 forecasting in china," en, *medRxiv*, p. 2020.03.23.20041608, Mar. 27, 2020, publisher: Cold Spring Harbor Laboratory Press.

[45] M. A. A. Al-qaness *et al.*, "Optimization method for forecasting confirmed cases of covid-19 in china," en, *Journal of Clinical Medicine*, vol. 9, no. 3, p. 674, Mar. 2020, number: 3 publisher: Multidisciplinary Digital Publishing Institute.

[46] S. Ayyoubzadeh *et al.*, "Predicting covid-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study," *JMIR Public Health and Surveillance*, vol. 6, Mar. 21, 2020.

[47] N. S. Punn *et al.*, "Covid-19 epidemic analysis using machine learning and deep learning algorithms," en, *medRxiv*, p. 2020.04.08.20057679, Jun. 1, 2020, publisher: Cold Spring Harbor Laboratory Press.

[48] S. k. Paul *et al.*, "A multivariate spatiotemporal spread model of covid-19 using ensemble of convlstm networks," *medRxiv*, 2020.

[49] A. Zeroual *et al.*, "Deep learning methods for forecasting covid-19 time-series data: A comparative study," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 121, Nov. 1, 2020.

[50] R. Dandekar and G. Barbastathis, "Neural network aided quarantine control model estimation of global covid-19 spread," *arXiv:2004.02752 [physics, q-bio]*, Apr. 2, 2020, arXiv: 2004.02752.

[51] M. R. Ibrahim *et al.*, "Variational-lstm autoencoder to forecast the spread of coronavirus across the globe," en, *medRxiv*, p. 2020.04.20.20070938, Apr. 24, 2020, publisher: Cold Spring Harbor Laboratory Press.

[52] R. Kafieh *et al.*, "Covid-19 in iran: A deeper look into the future," en, *medRxiv*, p. 2020.04.24.20078477, Apr. 27, 2020, publisher: Cold Spring Harbor Laboratory Press.

[53] L. R. Kolozsvari *et al.*, "Predicting the epidemic curve of the coronavirus (sars-cov-2) disease (covid-19) using artificial intelligence," en, *medRxiv*, p. 2020.04.17.20069666, Jan. 27, 2021, publisher: Cold Spring Harbor Laboratory Press.

[54] S. Dutta and S. K. Bandyopadhyay, "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," en, *medRxiv*, p. 2020.03.25.20043505, Mar. 30, 2020, publisher: Cold Spring Harbor Laboratory Press.

[55] Z. Car *et al.*, "Modeling the spread of covid-19 infection using a multilayer perceptron," Engels, *Computational and Mathematical Methods in Medicine*, vol. 2020, Jan. 1, 2020, [Online; accessed 2021-02-01].

[56] P. Hartono, "Similarity maps and pairwise predictions for transmission dynamics of covid-19 with neural networks," en, *Informatics in Medicine Unlocked*, vol. 20, p. 100 386, Jan. 1, 2020.

[57] S. Uhlig *et al.*, "Modeling projections for covid-19 pandemic by combining epidemiological, statistical, and neural network approaches," en, *medRxiv*, p. 2020.04.17.20059535, Apr. 22, 2020, publisher: Cold Spring Harbor Laboratory Press.

[58] H. Dutta, "Neural network model for prediction of covid-19 confirmed cases and fatalities," *Preprint*, May 1, 2020.

[59] B. Yan *et al.*, "An improved method for the fitting and prediction of the number of covid-19 confirmed cases based on lstm," en, *Computers, MaterialsContinua*, vol. 64, no. 3, pp. 1473–1490, Jun. 30, 2020.

[60] M. A. A. Al-qaness *et al.*, "Marine predators algorithm for forecasting confirmed cases of covid-19 in italy, usa, iran and korea," eng, *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, May 18, 2020, PMID: 32443476 PMCID: PMC7277148.

[61] M. Karimuzzaman *et al.*, "Forecasting the covid-19 pandemic with climate variables for top five burdening and three south asian countries," en, *medRxiv*, p. 2020.05.12.20099044, May 19, 2020, publisher: Cold Spring Harbor Laboratory Press.

[62] S. Cabras, "A bayesian - deep learning model for estimating covid-19 evolution in spain," *arXiv:2005.10335 [stat]*, May 20, 2020, arXiv: 2005.10335.

[63] A. M. Javid *et al.*, "Predictive analysis of covid-19 time-series data from johns hopkins university," *arXiv:2005.05060 [cs, eess]*, May 22, 2020, arXiv: 2005.05060.

[64] M. Azarafza *et al.*, "Covid-19 infection forecasting based on deep learning in iran," en, *medRxiv*, p. 2020.05.16.20104182, May 24, 2020, publisher: Cold Spring Harbor Laboratory Press.

[65] S. M. Zandavi *et al.*, "Forecasting the spread of covid-19 under control scenarios using lstm and dynamic behavioral models," *arXiv:2005.12270 [physics]*, May 24, 2020, arXiv: 2005.12270.

[66] C. Direkoglu and M. Sah, "Worldwide and regional forecasting of coronavirus (covid-19) spread using a deep learning model," en, *medRxiv*, p. 2020.05.23.20111039, May 26, 2020, publisher: Cold Spring Harbor Laboratory Press.

[67] L.-P. Chen, "Analysis and prediction of covid-19 data in taiwan," en, Social Science Research Network, Rochester, NY, Tech. Rep., May 27, 2020, DOI: 10.2139/ssrn.3611761.

[68] A. Chatterjee *et al.*, "Statistical explorations and univariate timeseries analysis on covid-19 datasets to understand the trend of disease spreading and death," *Sensors (Basel, Switzerland)*, vol. 20, no. 11, May 29, 2020, PMID: 32486055 PMCID: PMC7308840.

[69] G. Pinter *et al.*, "Covid-19 pandemic prediction for hungary; a hybrid machine learning approach," *Mathematics*, vol. 8, no. 6, 2020.

[70] T. H. H. Aldhyani *et al.*, "Deep learning and holt-trend algorithms for predicting covid-19 pandemic," en, *medRxiv*, p. 2020.06.03.20121590, Jun. 5, 2020, publisher: Cold Spring Harbor Laboratory Press.

[71] R. S. Pontoh *et al.*, "Effectiveness of the public health measures to prevent the spread of covid-19," en, *Commun. Math. Biol. Neurosci.*, vol. 2020, no. 0, Article ID 31, Jun. 18, 2020, number: 0.

[72] P. Melin *et al.*, "Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: The case of mexico," eng, *Healthcare (Basel, Switzerland)*, vol. 8, no. 2, Jun. 19, 2020, PMID: 32575622 PMCID: PMC7349072.

[73] S. R. Vadyala *et al.*, "Prediction of the number of covid-19 confirmed cases based on k-means-lstm," *arXiv:2006.14752 [physics, q-bio]*, Jun. 25, 2020, arXiv: 2006.14752.

[74] Y. Tian *et al.*, "Forecasting covid-19 cases using machine learning models," en, *medRxiv*, p. 2020.07.02.20145474, Jul. 4, 2020, publisher: Cold Spring Harbor Laboratory Press.

[75] A. Kapoor *et al.*, "Examining covid-19 forecasting using spatio-temporal graph neural networks," *arXiv:2007.03113 [cs]*, Jul. 6, 2020, arXiv: 2007.03113.

[76] L. Moftakhar *et al.*, "Exponentially increasing trend of infected patients with covid-19 in iran: A comparison of neural network and arima forecasting models," *Iranian Journal of Public Health*, vol. 49, Jul. 11, 2020.

[77] S. K. Tamang *et al.*, "Forecasting of covid-19 cases based on prediction using artificial neural network curve fitting technique," *Global Journal of Environmental Science and Management*, vol. 6, no. Special Issue (Covid-19), pp. 53–64, Aug. 1, 2020.

[78] N. Hasan, "A methodological approach for predicting covid-19 epidemic using eemd-ann hybrid model," en, *Internet of Things*, vol. 11, p. 100 228, Sep. 1, 2020.

[79] R. G. da Silva *et al.*, "Forecasting brazilian and american covid-19 cases based on artificial intelligence coupled with climatic exogenous variables," *Chaos, Solitons, and Fractals*, vol. 139, p. 110 027, Oct. 2020, PMID: 32834591 PMCID: PMC7324930.

[80] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 120, Nov. 1, 2020.

[81] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," en, *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, Mar. 2021, number: 3 publisher: Nature Publishing Group.

[82] *Strategies for the surveillance of covid-19*, en, Available at https://www.ecdc.europa.eu/en/publications-data/strategies-surveillance-covid-19 [Online; accessed 2021-02-28], Apr. 9, 2020.

[83] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," en, *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1, 1943, Company: Springer Distributor: Springer Institution: Springer Label: Springer number: 4 publisher: Kluwer Academic Publishers.

[84] J. Patterson and A. Gibson, *Deep Learning*, en. O'Reilly Media, Inc., Aug. 2017.

[85] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," eng, *Psychological Review*, vol. 65, no. 6, pp. 386–408, Nov. 1958, PMID: 13602029.

[86] ——, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," *The American Journal of Psychology*, vol. 76, no. 4, pp. 705–707, 1963.

[87] B. Curry and P. H. Morgan, "Neural networks, linear functions and neglected non-linearity," en, *Computational Management Science*, vol. 1, no. 1, pp. 15–29, Dec. 1, 2003.

[88] Y. Lecun, *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*, En-

glish (US). Universite P. et M. Curie (Paris 6), Jun. 1987.

[89] D. Ballard, "Modular learning in neural networks," *AAAI 87*, pp. 279–284, 1987.

[90] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," PhD thesis, Harvard University, 1974.

[91] D. E. Rumelhart *et al.*, "Learning representations by back-propagating errors," en, *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, number: 6088 publisher: Nature Publishing Group.

[92] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1, 1997.

[93] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv:1406.1078 [cs, stat]*, Sep. 2, 2014, arXiv: 1406.1078.

[94] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, event: IEEE Transactions on Signal Processing.

[95] A. Vaswani *et al.*, "Attention is all you need," *arXiv:1706.03762 [cs]*, Dec. 5, 2017, arXiv: 1706.03762 version: 5.

[96] Y. Lecun, "Generalization and network design strategies," English (US), *Connectionism in perspective*, 1989, publisher: Elsevier.

[97] G.-B. Huang *et al.*, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, vol. 2, 2004, 985–990 vol.2.

[98] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," en, *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 28, 2006, publisher: American Association for the Advancement of Science section: Report PMID: 16873662.

[99] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, vol. 286, no. 3, pp. 800–809, Jan. 8, 2018, publisher: Radiological Society of North America.

[100] S. H. Park and H. Y. Kressel, "Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: What peer-reviewed medical journals could do," eng, *Journal of Korean Medical Science*, vol. 33, no. 22, e152, May 28, 2018, PMID: 29805337 PMCID: PMC5966371.

[101] G. S. Handelman *et al.*, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *American Journal of Roentgenology*, vol. 212, no. 1, pp. 38–43, Oct. 17, 2018, publisher: American Roentgen Ray Society.

[102] D. A. Bluemke *et al.*, "Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers—from the radiology editorial board," *Radiology*, vol. 294, no. 3, pp. 487–489, Dec. 31, 2019, publisher: Radiological Society of North America.

[103] D. G. Altman *et al.*, "Equator: Reporting guidelines for health research," English, *The Lancet*, vol. 371, no. 9619, pp. 1149–1150, Apr. 5, 2008, publisher: Elsevier PMID: 18395566.

[104] J. M. Provenzale and R. J. Stanley, "A systematic guide to reviewing a manuscript," eng, *AJR. American journal of roentgenology*, vol. 185, no. 4, pp. 848–854, Oct. 2005, PMID: 16177399.

[105] W. Luo *et al.*, "Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view," EN, *Journal of Medical Internet Research*, vol. 18, no. 12, e5870, Dec. 16, 2016, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research publisher: JMIR Publications Inc., Toronto, Canada.

[106] J. Mongan *et al.*, "Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers," *Radiology: Artificial Intelligence*, vol. 2, no. 2, e200029, Mar. 1, 2020, publisher: Radiological Society of North America.

[107] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice (3rd ed)*, 3rd ed. Melbourne, Australia: OTexts, 2021, [Online; accessed 2021-04-14].

[108] S. K. Smith and T. Sincich, "An empirical analysis of the effect of length of forecast horizon on population forecast errors," *Demography*, vol. 28, no. 2, pp. 261–274, 1991, publisher: Springer.

[109] Y. Wang *et al.*, "The influence of the activation function in a convolution neural network model of facial expression recognition," en, *Applied Sciences*, vol. 10, no. 5, p. 1897, Jan. 2020, number: 5 publisher: Multidisciplinary Digital Publishing Institute.

[110] I. Goodfellow *et al.*, *Deep Learning*. The MIT Press, 2016.

[111] G. Perin and S. Picek, "On the influence of optimizers in deep learning-based side-channel analysis," *IACR Cryptol. ePrint Arch.*, 2020.

[112] C. Bergmeir *et al.*, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018.

[113] E. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, en, 2nd ed., ser. Statistics for Biology and Health. Springer International Publishing, 2019, DOI: 10.1007/978-3-030-16399-0.

[114] V. Amrhein *et al.*, "Scientists rise up against statistical significance," en, *Nature*, vol. 567, no. 7748, pp. 305–307, Mar. 2019, number: 7748 publisher: Nature Publishing Group.

[115] D. W. Hosmer *et al.*, "A comparison of goodness-of-fit tests for the logistic regression model," en, *Statistics in Medicine*, vol. 16, no. 9, pp. 965–980, 1997.

[116] R. Nuzzo, "Scientific method: Statistical errors," en, *Nature News*, vol. 506, no. 7487, p. 150, Feb. 13, 2014, section: News Feature.

[117] A. Davydenko and R. Fildes, *Forecast error measures: Critical review and practical recommendations*, Jan. 2016.

[118] *Retracted coronavirus (covid-19) papers*, en-US, Available at https://retractionwatch.com/retracted-coronavirus-covid-19-papers [Online; accessed 2021-04-27], Apr. 29, 2020.

[119] N. B. Yahia *et al.*, "Deep ensemble learning method to forecast covid-19 outbreak," In Review, Tech. Rep., May 21, 2020, DOI: 10.21203/rs.3.rs-27216/v1.

[120] N. Yudistira, "Covid-19 growth prediction using multivariate long short term memory," en, *Preprint*, May 10, 2020, [Online; accessed 2020-12-15].

[121] A. Chatterjee and S. Roy, "An analytics overview & lstm-based predictive modeling of covid-19: A hardheaded look across india," in *Machine Intelligence and Soft Computing*, D. Bhattacharyya and N. Thirupathi Rao, Eds., Singapore: Springer Singapore, 2021, pp. 289–307.

[122] F. Shahid *et al.*, "Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 212, Nov. 1, 2020.

[123] L. Mohimont *et al.*, "Convolutional neural networks and temporal cnns for covid-19 forecasting in france," en, *Applied Intelligence*, 2021, [Online; accessed 2021-01-08].

[124] H. Abbasimehr and R. Paki, "Prediction of covid-19 confirmed cases combining deep learning methods and bayesian optimization," en, *Chaos, Solitons & Fractals*, vol. 142, p. 110 511, Jan. 1, 2021.

[125] V. Bharadi, "Random net implementation of mlp and lstms using averaging ensembles of deep learning models," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 1197–1204, 2020.

[126] A. Prakash *et al.*, "Spread peak prediction of covid-19 using ann and regression (workshop paper)," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 356–365.

[127] A. Mollalo *et al.*, "Artificial neural network modeling of novel coronavirus (covid-19) incidence rates across the continental united states," en, *International Journal of Environmental Research and Public Health*, vol. 17, no. 12, p. 4204, Jan. 2020, number: 12 publisher: Multidisciplinary Digital Publishing Institute.

[128] M. Saqib, "Forecasting covid-19 outbreak progression using hybrid polynomial-bayesian ridge regression model," en, *Applied Intelligence*, Oct. 23, 2020, [Online; accessed 2021-01-22].

[129] R. Moulay Taj *et al.*, "Towards using recurrent neural networks for predicting influenza-like illness: Case study of covid-19 in morocco," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, pp. 7945–7950, Oct. 22, 2020.

[130] K. Shyam Sunder Reddy *et al.*, "Recurrent neural network based prediction of number of covid-19 cases in india," en, *Materials Today: Proceedings*, Nov. 17, 2020, [Online; accessed 2021-01-12].

[131] M. A. M. Arceda *et al.*, "Forecasting time series with multiplicative trend exponential smoothing and lstm: Covid-19 case study," in *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2*, K. Arai *et al.*, Eds., Cham: Springer International Publishing, 2021, pp. 568–582.

[132] S. Chander *et al.*, "Jaya spider monkey optimization-driven deep convolutional lstm for the prediction of covid'19," *Bio-Algorithms and Med-Systems*, vol. 16, Nov. 13, 2020.

[133] A. S. Fokas *et al.*, "Mathematical models and deep learning for predicting the number of individuals reported to be infected with sars-cov-2," *Journal of The Royal Society Interface*, vol. 17, no. 169, p. 20 200 494, 2020.

[134] S. Chakraborty *et al.*, "Reaction order and neural network approaches for the simulation of covid-19 spreading kinetic in india," en, *Infectious Disease Modelling*, vol. 5, pp. 737–747, Jan. 1, 2020.

[135] Z. Li *et al.*, "A recurrent neural network and differential equation based spatiotemporal infectious disease model with application to covid-19," *arXiv:2007.10929 [cs, q-bio, stat]*, Sep. 17, 2020, arXiv: 2007.10929.

[136] M. Wieczorek *et al.*, "Neural network powered covid-19 spread forecasting model," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 203, Nov. 1, 2020.

[137] I. Pereira *et al.*, "Forecasting covid-19 dynamics in brazil: A data driven approach," *International Journal of Environmental Research and Public Health*, vol. 17, p. 5115, Jul. 15, 2020.

[138] İ. Kırbaş *et al.*, "Comparative analysis and forecasting of covid-19 cases in various european countries with arima, narnn and lstm approaches," en, *Chaos, Solitons & Fractals*, vol. 138, p. 110 015, Sep. 1, 2020.

[139] M. Wieczorek *et al.*, "Real-time neural network based predictor for cov19 virus spread," en, *PLOS ONE*, vol. 15, no. 12, e0243189, 2020, publisher: Public Library of Science.

[140] A. Rodriguez *et al.*, "Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting," en, *medRxiv*, p. 2020.09.28.20203109, Sep. 29, 2020, publisher: Cold Spring Harbor Laboratory Press.

[141] A. Ramchandani *et al.*, "Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions," *IEEE Access*, vol. 8, pp. 159 915–159 930, 2020, event: IEEE Access.

[142] R. Sujath *et al.*, "A machine learning forecasting model for covid-19 pandemic in india," en, *Stochastic Environmental Research and Risk Assessment*,

vol. 34, no. 7, pp. 959–972, Jul. 1, 2020, Company: Springer Distributor: Springer Institution: Springer Label: Springer number: 7 publisher: Springer Berlin Heidelberg.

[143] S. D. Khan *et al.*, "Toward smart lockdown: A novel approach for covid-19 hotspots prediction using a deep hybrid neural network," en, *Computers*, vol. 9, no. 4, p. 99, Dec. 11, 2020.

[144] C. Distante *et al.*, "Forecasting covid-19 outbreak progression in italian regions: A model based on neural network training from chinese data," en, *medRxiv*, p. 2020.04.09.20059055, Apr. 14, 2020, publisher: Cold Spring Harbor Laboratory Press.

[145] A. H. Elsheikh *et al.*, "Deep learning-based forecasting model for covid-19 outbreak in saudi arabia," en, *Process Safety and Environmental Protection*, vol. 149, pp. 223–233, May 1, 2021.

[146] S. Dhamodharavadhani *et al.*, "Covid-19 mortality rate prediction for india using statistical neural network models," English, *Frontiers in Public Health*, vol. 8, 2020, publisher: Frontiers.

[147] S. Prasanth *et al.*, "Forecasting spread of covid-19 using google trends: A hybrid gwo-deep learning approach," en, *Chaos, Solitons & Fractals*, vol. 142, p. 110 336, Jan. 1, 2021.

[148] H. T. Rauf *et al.*, "Time series forecasting of covid-19 transmission in asia pacific countries using deep neural networks," en, *Personal and Ubiquitous Computing*, Jan. 10, 2021, [Online; accessed 2021-01-22].

[149] P. Arora *et al.*, "Prediction and analysis of covid-19 positive cases using deep learning models: A descriptive case study of india," en, *Chaos, Solitons & Fractals*, vol. 139, p. 110 017, Oct. 1, 2020.

[150] O. Istaiteh *et al.*, "Machine learning approaches for covid-19 forecasting," in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2020, pp. 50–57.

[151] Y. Karadayi *et al.*, "Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: Early detection of covid-19 outbreak in italy," *IEEE Access*, vol. 8, pp. 164 155–164 177, 2020, event: IEEE Access.

[152] A. I. Saba and A. H. Elsheikh, "Forecasting the prevalence of covid-19 outbreak in egypt using nonlinear autoregressive artificial neural networks," en, *Process Safety and Environmental Protection*, vol. 141, pp. 1–8, Sep. 1, 2020.

[153] S. Ballı, "Data analysis of covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," en, *Chaos, Solitons & Fractals*, vol. 142, p. 110 512, Jan. 1, 2021.

[154] K. T. Ly, "A covid-19 forecasting system using adaptive neuro-fuzzy inference," en, *Finance Research Letters*, p. 101 844, Nov. 12, 2020.

[155] V. Papastefanopoulos *et al.*, "Covid-19: A comparison of time series methods to forecast percentage of active cases per population," en, *Applied Sciences*, vol. 10, no. 11, p. 3880, Jan. 2020, number: 11 publisher: Multidisciplinary Digital Publishing Institute.

[156] M. Hawas, "Generated time-series prediction data of covid-19's daily infections in brazil by using recurrent neural networks," en, *Data in Brief*, vol. 32, p. 106 175, Oct. 1, 2020.

[157] M. A. Achterberg *et al.*, "Comparing the accuracy of several network-based covid-19 prediction algorithms," en, *International Journal of Forecasting*, Oct. 9, 2020, [Online; accessed 2021-01-20].

[158] P. Wang *et al.*, "Time series prediction for the epidemic trends of covid-19 using the improved lstm deep learning method: Case studies in russia, peru and iran," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 214, Nov. 1, 2020.

[159] S. Thakur *et al.*, "Prediction for the second wave of covid-19 in india," in *Big Data Analytics*, L. Bellatreche *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 134–150.

[160] S. Shastri *et al.*, "Time series forecasting of covid-19 using deep learning models: India-usa comparative case study," en, *Chaos, Solitons & Fractals*, vol. 140, p. 110 227, Nov. 1, 2020.

[161] S. Saif *et al.*, "A hybrid model based on mba-anfis for covid-19 confirmed cases prediction and forecast," en, *Journal of The Institution of Engineers (India): Series B*, Jan. 19, 2021, [Online; accessed 2021-01-22].

[162] Z. Zhao *et al.*, "How well can we forecast the covid-19 pandemic with curve fitting and recurrent neural networks?" *Preprint*, May 18, 2020, DOI: 10.1101/2020.05.14.20102541.

[163] N. M. Ghazaly *et al.*, "Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network," en, *International Journal of Advanced Science and Technology*, vol. 29, no. 5s, pp. 1831–1849, Apr. 9, 2020, number: 5s.

[164] S. Bahri *et al.*, "Deep learning for covid-19 prediction," in *2020 4th International Conference on Advanced Systems and Emergent Technologies*, 2020, pp. 406–411.

[165] Y. Gautam, "Transfer learning for covid-19 cases and deaths forecast using lstm network," en, *ISA Transactions*, Jan. 4, 2021, [Online; accessed 2021-01-20].

[166] J. A. L. Marques *et al.*, "Artificial intelligence prediction for the covid-19 data based on lstm neural networks and h2o automl," en, in *Predictive Models for Decision Support in the COVID-19 Crisis*, ser. SpringerBriefs in Applied Sciences and Technology, J. A. L. Marques *et al.*, Eds., Cham: Springer International Publishing, 2021, pp. 69–87.

[167] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005, publisher: Inter-Research Science Center.

[168] T. Chai and R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?" *Geosci. Model Dev.*, vol. 7, Jan. 31, 2014.

[169] K. Broman *et al.*, *Recommendations to funding agencies for supporting reproducible research*, en.