Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Novel deep learning approach to model and predict the spread of COVID-19

Devante Ayris, Maleeha Imtiaz, Kye Horbury, Blake Williams, Mitchell Blackney, Celine Shi Hui See, Syed Afq Ali Shah

*Discipline of Information Technology, Murdoch University, Australia
bJoondalup Health Campus, Australia
CSehat Kahnmi, Pakistan
dCentre for AI and Machine Learning, Edith Cowan University, Australia

**Corresponding author at: Centre for AI and Machine Learning, Edith Cowan University, Australia.
E-mail address: afq.shah@ecu.edu.au (S.A.A. Shah).

ARTICLE INFO

Article history:
Received 18 December 2021
Revised 3 February 2022
Accepted 1 March 2022
Available online 16 March 2022

Keywords:
COVID-19 prediction
Machine learning
Regression
MAE

ABSTRACT

SARS-CoV2, which causes coronavirus disease (COVID-19) is continuing to spread globally, producing new variants and has become a pandemic. People have lost their lives not only due to the virus but also because of the lack of counter measures in place. Given the increasing caseload and uncertainty of spread, there is an urgent need to develop robust artificial intelligence techniques to predict the spread of COVID-19. In this paper, we propose a deep learning technique, called Deep Sequential Prediction Model (DSPM) and machine learning based Non-parametric Regression Model (NRM) to predict the spread of COVID-19. Our proposed models are trained and tested on publicly available novel coronavirus dataset. The proposed models are evaluated by using Mean Absolute Error and compared with the existing methods for the prediction of the spread of COVID-19. Our experimental results demonstrate the superior prediction performance of the proposed models. The proposed DSPM and NRM achieve MAEs of 388.43 (error rate 1.6%) and 142.23 (0.6%), respectively compared to 6508.22 (27%) achieved by baseline SVM, 891.13 (9.2%) by Time-Series Model (TSM), 615.25 (7.4%) by LSTM-based Data-Driven Estimation Method (DDEM) and 929.72 (8.1%) by Maximum-Hasting Estimation Method (MHEM).

© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

COVID-19 is a pandemic that has spread and devastated countries around the world. Even months on from the original outbreak of the virus, it still poses a large threat to everyone around the globe, as with each passing day, the death toll still increases, and more and more cases are identified (Li et al., 2022; Rahimi, Chen, & Gandomi, 2021). Countries have been brought to a standstill as citizens are forced to self-isolate and worldwide economies have come to a halt as a result of the negative impacts on trade and industry (Alamoodi et al., 2020; Chandra, Verma, Singh, Jain, & Nemat, 2020; Loey, Manogaran, & Khalifa, 2020).

First discovered in Wuhan City, Hubei Province of China, on the 31st of December 2019, COVID-19 is a respiratory illness with pneumonia-like qualities and was initially thought to be caused by human contact with exotic fauna, eventually resulting in a person-to-person spread. This virus has caused a massive negative international impact and has affected the day-to-day lives of millions of people.

It is still difficult to predict where and when new cases will appear, and many governments have failed to understand the scale and impact of the virus. The exponential spread of the virus (including its variants) means that until there is a fully vaccinated population, or it has been completely removed from the population, it will always pose a threat even in locations with the best circumstances (D’Angelo & Palmieri, 2021). A few techniques have been proposed for the prediction of COVID-19, however, most of these techniques are based on traditional machine learning methods and mathematical modelling (Atlam, Ewis, Abd El-Raouf, Ghoneim, & Gad, 2021; Malki et al., 2021). In addition, these techniques focus only on specific region e.g., India, China and Africa (Fanelli & Piazzo, 2020; Tomar & Gupta, 2020; Yang et al., 2020). There is, therefore, a strong need to develop automatic techniques...
for the prediction of the spread of this virus through the world’s population.

Deep learning has been a growing trend in data analysis and predictive modeling in recent years, and has been termed one of the ten breakthrough technologies (Garain, Basu, Giampaloo, Velasquez, & Sarkar, 2021; Shah, 2019; Xue et al., 2021). It is emerging as the leading machine learning tool in computer vision. This data-driven approach has shown unprecedented performance for several computer vision tasks. It learns the most predictive features (learned features) directly from data given a large dataset of labeled examples. In recent years, deep learning techniques have emerged as highly effective methods for prediction and decision-making in a multitude of disciplines including health (hearing aids) and aged care.

Inspired by the recent advancement in machine/deep learning, this research hypothesizes that deep learning can be used to predict the spread of the virus and potentially be used to help allocate resources and prepare procedures ahead of time to mitigate the impacts of COVID-19, potentially saving lives. In this paper, we propose two different techniques to predict the spread of COVID-19. The paper proposes Deep Sequential Prediction Model (DSPM), which benefits from the sequential nature of the data to make accurate prediction about the spread of this disease. The paper also presents an efficient Non-parametric Regression Model (NRM), which avoids computationally expensive parameter learning process to efficiently predict the spread of COVID-19. We extensively evaluate the proposed models and analyse their viability to predict the spread of COVID-19. The contributions of this paper can be summarized as follows:

- The paper proposes a deep sequential prediction model (DSPM) to learn distinctive features from the input time series data for accurate prediction of COVID-19 spread.
- The paper also proposes a non-parametric regression model (NRM) to accurately and efficiently predict the spread of this contagious disease.
- Extensive evaluation of the proposed models has been performed on publicly available large coronavirus dataset. Our experimental results demonstrate the superior performance of the proposed models.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 presents our proposed techniques to predict the spread of COVID-19. Experimental results are provided in Section 4, which also provides details of the novel Coronavirus dataset. Section 5 provides discussion and analysis about the proposed techniques. The paper is concluded in Section 6.

2. Literature review

With the rising issue of the Coronavirus infectious disease (and other similar diseases such as SARS and MERS), there have been few studies involving machine learning to predict the recovery of infected patients and study the similarity of SARS virus protein with other viruses. John & Shaiba (2019) proposed machine learning techniques to track and analyze different factors that are involved in the recovery from MERS. SVM, conditional inference tree, Naive Bayes and J48 models were used to determine and predict whether the categories, including gender and age, the patient is a healthcare worker, status at time of identification of disease, the patient had symptoms and whether the patient had any pre-existing diseases or conditions, were important factors in determining the recovery of a patient from MERS. Their models determined that age, being a healthcare worker, the status at the time of identification and whether they had pre-existing disease are good indicators for predicting the recovery from MERS, with a p-value of 0.001278, 0.001260, 2e−16 and 0.001067, respectively.

Cai, Han, Chen, Cao, & Chen (2005) proposed a method to compare the SARS virus proteins to those of other viruses, to predict how many of those proteins are similar with each other. They used an SVM model in conjunction with the sequence comparison method BLAST to predict the functional class of a given protein i.e., it is a part of the 46 enzyme families, the 21 channel/transporter families or the 5 RNA-binding protein families to name a few. Their evaluation showed that an SVM can accurately predict the functional class with 73% accuracy.

Tang et al. (2015) proposed a machine learning technique to predict the potential animal hosts of the SARS and MERS viruses. Two machine learning models were used, a non-linear SVM using a radial kernel and a Mahalanobis distance (MD) discriminant model, with both using leave-one-out cross-validation of the training data, to determine host candidates. Both models were successful, with the SVM model having a 99.86% prediction rate in inferring potential hosts, while the MD model having a 98.08% prediction rate.

Ismael & Şengur (2020) proposed a deep learning based technique to classify COVID-19 and normal (healthy) chest X-ray images. They used pre-trained CNN models to extract features and used SVM for classification. The ResNet50 model in their approach achieved 94.7% classification accuracy.

In another work, Chakraborty & Mali (2020) proposed an unsupervised image segmentation approach based on super-pixel and fuzzy clustering system to explicate COVID-19 radiology images. Their reported results were promising and better than the other existing techniques.

Atlam et al. (2021) proposed a machine learning approach to study the impact of COVID-19 pandemic on educations systems especially on university students’ psychological health. Participants’ responses were collected using a questionnaire. The collected data was then analysed using ensemble machine learning technique. Their results showed promising performance.

Several approaches to predict the spread of COVID-19 have recently emerged. In the following, we discuss few relevant techniques. For a detailed review, the readers are referred to this survey paper (Shah, Mulahuwaish, Chafoor, & Maghdid, 2020). Elmousalami & Hassanien (2020) proposed time series models and mathematical formulation to predict the spread of COVID-19 using publicly available data. In their proposed approach, they used different models to forecast and validate assumptions related to COVID-19 spread. Tomar & Gupta (2020) used an LSTM (Long Short-Time Memory) model to predict the spread of COVID-19 in India. Their model was shown to achieve good prediction performance, however, their model has few limitations. It has been tested only on COVID-19 data for India and cannot be generalised to global impact and spread of the disease. Their reported results are based on limited data and this impacts the significance of the model.

Zhao et al. (2020) proposed Maximum-Hasting parameter estimation method and the modified version of Susceptible Exposed Infectious Recovered (SEIR) model to analyse the spread of COVID-19 in six African countries/nations. They classify these countries into three categories including mitigation, suppression or mildness. One of the drawbacks of their approach is that they assume intervention intensity of studied nations at a fraction of comparison model (i.e., China in their case) and the prediction accuracy of their model drops if suggested interventions are not carried out. In addition, their analysis is only restricted to six African countries and hence does not reflect the impact of the technique on global level.

Yang et al. (2020) proposed Susceptible-Exposed-Infectious-Removed (SEIR) and LSTM models to predict the probability of epidemic including its peak and the impact of intervention measures in China. Their model is able to predict the spread of COVID-19 in
China with a reasonable confidence. The limitations of this technique are that the accuracy of the model depends on the implementations of pre-defined control measures and the prediction is limited to China only.

Malki et al. (2021) proposed a decision tree based technique for the prediction of COVID-19. The core idea of their approach is to utilize supervised machine learning algorithms for time-series forecasting. Their model predicted that COVID-19 infections will greatly decline during the first week of September 2021 when it will be going to an end shortly afterward. However, the current state of the pandemic, particularly due to the recent spread of the Omicron variant, does not support the reported results.

Fanelli & Piazza (2020) proposed Mean-field approximation in modified Susceptible Infectious-Recovered Deceased (SIRD) model to predict the maximum number of infected individuals in China and the peak of pandemic. Their technique is shown to provide estimates for the magnitude and time of the epidemic peak. The major limitation of their proposed technique is that it relies on predefined conditions and overestimates the number of deaths.

In view of the above, it can be noted that the most recent approaches are geared towards the predictive modelling using mathematical formulation and statistical techniques. Most of the techniques are limited to a specific region e.g., India Tomar & Gupta (2020), or China Fanelli & Piazza (2020); Yang et al. (2020). There are only a few deep learning-based techniques in the literature for COVID-19 prediction, however, those approaches have not been evaluated to predict the global spread of COVID-19 i.e., in all or most of the countries of the world. In contrast to the existing techniques, this paper proposes deep learning techniques to predict the spread of novel coronavirus COVID-19. The proposed models have been evaluated on 6.4 million confirmed COVID-19 cases reported in different countries (around 90 in our case) and their provinces/states.

3. Proposed models

In this section, we present our proposed prediction models including Deep Sequential Prediction Model (DSPM) and Non-parametric Regression Model (NRM).

3.1. Deep sequential prediction model (DSPM)

Fig. 1 shows the proposed DSPM to predict the spread of COVID-19. As can be noted, our proposed DSPM is a stacked long short-term memory (LSTM) deep neural network. DSPM consists of four stacked LSTMs that feed into each other. These LSTMs contain four hidden layers each (for each stack) that process the data to yield a highly accurate model. We chose stacked LSTMs in our proposed models because the COVID-19 dataset has unknown duration of infection between the countries. This makes training a traditional recurrent neural network (RNN) difficult. This unknown duration period can cause RNN to encounter the vanishing gradient problem, which can completely halt an RNN from further training (Pascaru, Mikolov, & Bengio, 2013). On the other hand, an LSTM model is designed to handle this error. In the following, we discuss the different stages of our proposed DSPM.

3.1.1. Stage 1

Given an input data $X_t$, this stage (also known as the forget layer) decides whether the cell will throw away the previous data or keep it for modification. It makes this decision through a sigmoid calculation that returns a binary (either one or zero) value.

The sigmoid calculation is based on the input vector and the output of the previous block and the memory from the previous block. Therefore, if a new subject is seen, the cell will want to forget the old subject (Yan, 2015):

$$f_t = \sigma(W_f \cdot [H_{t-1}, X_t] + b_f)$$  \hspace{1cm} (1)

where $X_t$ is the input vector, $H_{t-1}$ is output of the previous block, $b_f$ is a bias term and $\sigma$ is a nonlinear function.

3.1.2. Stage 2

The second stage, also known as the input gate layer or new memory valve, processes the data from the previous stage and decides what will be stored in the second memory gate. It is based on a sigmoid layer and a tanh layer. The sigmoid layer works the same way as in Stage 1, while the tanh layer only takes input from the output of the previous block and the input vector. The $\text{tanh}$ layer then outputs to the memory gate forming new data (Yan,
2015):

\[ i_t = \sigma (W_i \cdot [H_{t-1}, X_t] + b_i) \]  
\[ \tilde{C}_t = \tanh (W_c \cdot [H_{t-1}, X_t] + b_c) \]  

3.1.3. Stage 3

In Stage 1, the model decides what data it needs to forget, and in Stage 2 it decides what data it is going to store. With the previous stages deciding what to do with the old data, the model now combines the data to form a new data by combining everything together. To achieve this, it uses the 2 element wise multiplication gates to one summation gate on the memory pipe, as follows:

\[ C_t = f_t \ast C_{t-1} + i_t \ast \tilde{C}_t \]  

3.1.4. Stage 4

In the final stage, the model finally outputs the data through two channels i.e., the memory channel and the actual output of the cell. First a sigmoid operation is performed that decides about the output. Then the processed memory is put through a \( \tanh \) non-linearity. These two operations push through to an element wise multiplication gate. This action is the final output of the cell data. The processed memory then continues onto its own output untouched by this final calculation, while the data output continues after processing (Yan, 2015):

\[ o_t = \sigma (W_o[H_{t-1}, X_t] + b_o) \]  
\[ H_t = o_t \ast \tanh (C_t) \]  

DSPM training and testing To train the proposed DSPM, the publicly available time series data that is fed to the model is first pre-processed. The data is split between country and provinces, and the time series data is then converted to a data frame that includes a date of the confirmed cases. Using empirically selected scalar threshold, this data frame is then converted to 0s and 1s and inputted into the DSPM for its training. DSPM training was found
3.2. Proposed non-parametric regression model (NRM)

In this section, we discuss our proposed non-parametric regression model (NRM). The NRM is based on an additive regression time-series algorithm and uses a decomposed time series model with three major components, i.e.,

\[ y(t) = g(t) + s(t) + h(t) + \epsilon_t \]  

where \( g(t) \) is either linear or a logistic growth curve trend, \( s(t) \) are periodic changes, \( h(t) \) captures irregular effects, and \( \epsilon_t \) represents errors created by unusual changes that are not supported by the model.

There are two trend models for \( g(t) \). These include a saturating growth model and a piece-wise linear model. A saturating growth model typically handles non-linear prediction, which meets our requirement. In the proposed NRM, we therefore use the saturating growth model for predicting the spread of the virus. The saturating growth model is represented as follows:

\[ g(t) = \frac{C}{1 + \exp(-k(t - m))} \]  

where \( C \) is the carrying capacity, \( k \) is the growth rate and \( m \) is the offset parameter. However, the growth rate is not constant, and therefore NRM incorporate trend changes in the growth model by defining change points where the growth rate can change. This is done by defining a vector of rate adjustments as follows:

\[ \delta \in \mathbb{R}^S \]  

where \( S \) represents change points at times and can be seen as \( S_j, j = 1, \ldots, S \); \( \delta_j \) is the change in rate that occurs at \( S_j \).

When the rate at time \( t \) is equal to \( k + a(t)^T \delta, k \) is adjusted, the offset parameter \( m \) must also be adjusted to connect endpoints of segments. When there is a correct adjustment \( \gamma_j \) at change point
Fig. 4. Prediction results for the proposed Non-Parametric Regression Model (Model1). Countries/Regions have been randomly selected from the overall results to demonstrate the prediction performance of the proposed NRM. Additional prediction results are shown in Fig. 7.

\[ \gamma_j = \left( s_j - m - \sum_{l<j} \gamma_l \right) \left( 1 - \frac{1}{k + \sum_{l<j} \delta_l} \right) \]  

(10)

Finally, the model for logistic growth is given by the following equation:

\[ g(t) = \frac{C(t)}{1 + \exp \left( - \left( k + \mathbf{a}(t)^\top \mathbf{\delta} \right) \left( t - \left( m + \mathbf{a}(t)^\top \mathbf{y} \right) \right) \right)} \]  

(11)

The proposed NRM was trained and tested in the same way as the DSPM, however, without using scalars for data input vectors.

4. Experimental results

We extensively evaluated the performance of the proposed models on the publicly available novel coronavirus (COVID-19) dataset. In this section, we first provide the details of the dataset and then present our experimental results.

4.1. Novel coronavirus dataset

We used publicly available novel Coronavirus dataset collected/compiled by John Hopkins University (2020). The dataset is available via Kaggle and Github (SRK, 2020). The dataset contains globally reported COVID-19 cases in the following format:

- **ObservationDate** - Date of the observation in MM/DD/YYYY
- **Province/State** - Province or state of the observation
- **Country/Region** - Country of observation
- **Last Update** - Time in UTC at which the row is updated for the given province or country.
- **Confirmed** - Cumulative number of confirmed cases till that date
- **Deaths** - Cumulative number of deaths till that date
- **Recovered** - Cumulative number of recovered cases till that date

In the dataset, there are 133 dates that are represented as time series points, and each time series point includes the number of
confirmed COVID-19 cases on that date. There are 266 rows for countries that are split up into provinces that have data for those 133 dates. There is also other data that includes recovery cases, and death cases that follow the same format as the confirmed cases. Our proposed models have been evaluated on 6.4 million COVID-19 cases, which have been reported from 22nd January to 6th June 2020.

4.2. Experimental setup

4.2.1. Data pre-processing

The data fed to each model is divided into country and state/province level and stored in objects to allow easy access to country predictions and error rates. Some of the predictions are in decimal value. All these prediction values are rounded to the...
4.2.2. Metric for evaluation

Prediction values are compared to real cases i.e., ground truth by using Mean Absolute Error (MAE), which is a loss function mostly used for regression models. MAE is a metric that is used to compare both predicted and the actual values. MAE is measured for each prediction, before the prediction values are rounded for computing an accurate error rate.

4.3. Prediction results

In the following, we present the prediction results for the baseline, our proposed models and comparison with the existing approaches.
4.3.1. Prediction results for baseline method

We use the popular Support Vector Machine (SVM) as our baseline method (called Model 1 in our experiments) to predict and analyze the spread of coronavirus. There are a few reasons for choosing SVM: (1) Its ease of implementation, (2) There’s no publicly available machine learning approach to predict the spread of COVID-19, (3) SVM can be used for modelling the linear and non-linear (exponential) regression, meaning that it is able to model output variables that are real and/or continuous values, for example predicting the average age of a person, or in the case of this paper, predicting the spread of coronavirus in a certain location. (4) Lastly it is computationally and memory efficient, as it uses a subset of the data given as training data and that makes it suitable for training on smaller datasets.

Table 1 (Column 3) reports the predictions of our baseline model and comparison with ground truth values. Fig. 2 shows prediction results for the baseline model. Fig. 2 (first column and row) shows the country (Bangladesh) that has the highest MAE out of all the countries that were analyzed by this model. It can be noted that this model was not able to accurately predict COVID-19 cases for this country. A similar trend was observed for other countries that have a large number of confirmed coronavirus cases.

4.3.2. Prediction results for DSPM

Table 1 (Column 4) and Table 2 report the prediction results for our proposed DSPM (called Model 2 in our experiments). The average MAE for this model is 388.43 (Table 2), which is very low compared to other methods.

| Country            | Ground truth (confirmed cases) | Baseline prediction | DSPM prediction | NRM prediction |
|--------------------|--------------------------------|---------------------|-----------------|----------------|
| Angola             | 86                             | 115                 | 90              | 81             |
| Argentina          | 18,319                         | 75                  | 18,290          | 16,423         |
| Austria            | 16,759                         | 295                 | 16,220          | 19,713         |
| Bahamas            | 102                            | 233                 | 97              | 103            |
| Bahrain            | 12,311                         | 77                  | 12,245          | 11,732         |
| Belgium            | 58,615                         | 246                 | 56,654          | 59,266         |
| Benin              | 244                            | 24                  | 217             | 240            |
| Bosnia and Herzegovina | 2535                      | 211                 | 2538            | 2569           |
| Brazil             | 555,383                        | 14                  | 578,432         | 509,319        |
| Bulgaria           | 2538                           | 181                 | 2375            | 2635           |
| Cambodia           | 125                            | 64                  | 125             | 151            |
| Chile              | 108,686                        | 49                  | 109,760         | 98,245         |
| Congo (Brazzaville)| 611                            | 112                 | 570             | 608            |
| Costa Rica         | 1105                           | 206                 | 1128            | 1058           |
| Cuba               | 2092                           | 211                 | 2019            | 2068           |
| Cyprus             | 952                            | 268                 | 880             | 953            |
| Czechia            | 9364                           | 256                 | 8484            | 9375           |
| Diamond Princess   | 712                            | 131                 | 6781            | 710            |
| Denmark            | 11,734                         | 240                 | 11,976          | 11,882         |
| Dominican Republic | 17,752                         | 126                 | 17,700          | 17,759         |
| Egypt              | 27,536                         | 63                  | 26,582          | 23,998         |
| France             | 184,980                        | 265                 | 177,107         | 186,533        |
| Gabon              | 2803                           | 2                   | 2998            | 2813           |
| Gambia             | 25                             | 179                 | 24              | 28             |
| Georgia            | 796                            | 202                 | 749             | 788            |
| Germany            | 183,879                        | 276                 | 157,952         | 184,833        |
| Greece             | 2937                           | 281                 | 2811            | 2964           |
| Guinea             | 3886                           | 16                  | 3719            | 3891           |
| Guyana             | 153                            | 163                 | 150             | 154            |
| Haiti              | 2226                           | 12                  | 2758            | 1493           |
| Holy See           | 12                             | 272                 | 11              | 12             |
| Honduras           | 5527                           | 48                  | 5728            | 5283           |
| Hungary            | 3921                           | 210                 | 3811            | 3972           |
| Iceland            | 1806                           | 326                 | 1743            | 2202           |
| India              | 207,191                        | 14                  | 219,792         | 191,044        |
| Indonesia          | 27,549                         | 116                 | 27,994          | 27,137.83      |
| Iraq               | 7387                           | 91                  | 7028            | 6076           |
| Ireland            | 25,066                         | 233                 | 23,658          | 25,437         |
| Israel             | 17,285                         | 282                 | 16,998          | 17,042         |
| Italy              | 233,515                        | 264                 | 227,832         | 235,225        |
| Jamaica            | 590                            | 191                 | 550             | 587            |
| Japan              | 16,837                         | 239                 | 15,845          | 16,954         |
| Jordan             | 755                            | 193                 | 682             | 780            |
| Kazakhstan         | 11,571                         | 87                  | 11,734          | 10,796         |

### Table 2

MAE and error rates of our proposed models, the baseline approach and comparison with state-of-the-art methods.

| Model                                | Average MAE | Error rate |
|--------------------------------------|-------------|------------|
| Baseline (Model 1)                   | 6508.22     | 27%        |
| TSM (Elmousalam and Hassani, 2020)   | 891.13      | 9.2%       |
| DSPM (Tomar and Gupta, 2020)         | 615.25      | 7.4%       |
| MHEM (Zhao et al., 2020)             | 929.72      | 8.1%       |
| Proposed DSPM                        | 388.43      | 1.6%       |
| Proposed NRM                         | 142.23      | 0.6%       |

Fig. 2 also shows countries with better prediction results. Table 2 reports the average MAE and error rate that can be expected as error estimate when the model predicts COVID-19 cases for a given country/region. As can be noted, the average MAE is really high compared to the total cases analyzed. Additional prediction results for this model have been provided in Fig. 5.
compared to the baseline model. The error rate for this model is 1.62%. As can be noted, the prediction results are very similar to the ground truth curve. Fig. 3 shows the prediction results for the proposed DSPM. For this model, most countries and provinces with the lowest MAEs include countries and provinces that generally have lower cases of the virus (Fig. 3). Additional prediction results for this model have been provided in Fig. 6.

4.3.3. Prediction results for NRM

Table 1 (Column 5) reports the prediction results for our proposed NRM (called Model 3 in our experiments). The average MAE for this model is 142.23 (Table 2), which is low compared to the baseline method and DSPM. The error rate for the proposed NRM is only 0.6%. Fig. 4 shows the prediction results (randomly selected for demonstration) for this model. As can be noted this model
achieves the best prediction results. The last row of Fig. 4 shows the countries and provinces that have the lowest error rate in their continent. Our NRM model outperforms the baseline and DSPM. Additional prediction results for this model have been provided in Fig. 7.

Fig. 8 (left column) shows the country that has the lowest MAE out of all the countries that were analyzed. Low MAEs are usually found within countries that have the lowest number of confirmed cases. This can be seen in Fig. 8 for two different models, which have the lowest MAE for this country. It can be generalized that the baseline model has a high failure rate when a country has a large number of cases to analyze.

4.4. Comparison with existing techniques:

To demonstrate the effectiveness of our proposed models, we compare them with the existing techniques for the prediction of the spread of COVID-19. These techniques include Time-Series Model (TSM) (Elmousalami & Hassanien, 2020), LSTM-based DDEM (Tomar & Gupta, 2020) and Maximum-Hasting Estimation Method (MHEM) (Zhao et al., 2020). These methods have been carefully implemented using the implementation details in the original papers and evaluated on Coronavirus datasets in our experiments. Our experimental results are reported in Table 2. These results demonstrate the superior performance of our proposed techniques compared to the existing methods. MAE and error rates for these approaches are high compared to our techniques. This is obvious because these approaches have been developed to estimate the prediction of COVID-19’s spread on small datasets and for specific regions/countries e.g., India or China. The performance of these techniques has deteriorated due to exposure to a larger and diverse data.

5. Discussion and analysis

In this paper, we have analysed the publicly available data for around 90 different countries (and their provinces). Variation in data is large as there were no cases of COVID-19 in most of the countries in early 2020, and a sudden surge was seen from March 2020. In other cases, e.g., mainland China, the data pattern is slightly different and uprising trend of the spread can be seen from January 2020. The distribution of COVID-19 data makes the dataset challenging.

Table 2 reports average MAE for our proposed techniques, DSPM and NRM, on COVID-19 dataset. High MAEs generally do not always mean bad predictions. For instance in Fig. 4 (Brazil, first row and first column), there were 555,383 confirmed cases analyzed in Brazil and having only a MAE error of 5472 basically means out of all the confirmed cases, 5472 individuals were predicted incorrectly. This means that there was only a 0.98% error for the entire data for Brazil and overall this is a good prediction. High MAEs can be classified as a bad error rate for the model predictions when the error rate is over 10% out of all confirmed cases for a country and province as seen in Fig. 2 (Bangladesh, first row and first column) for baseline methods (Model 1). The MAE for this case is 522297.28 out of 1.83 million confirmed cases. The error in this case is 28.51%. We observed that countries that have a small number of confirmed cases, generally have lower MAEs because there are not enough confirmed cases, thus models will have a limited range of cases that it can predict. This can be seen in Fig. 8 (for Lesotho), which shows different predictions for each model and both have low MAEs. Similar results are prevalent in other countries with small numbers of confirmed cases. Note that baseline model (Model 1) has an error rate of 27%, the proposed DSPM has an error rate of 1.62% and the proposed NRM has an error rate of 0.6%. Baseline model was not accurate enough compared to DSPM and NRM. In addition, proposed NRM performed better than the proposed DSPM, however, the difference in performance is not large. Both models can be used to model prediction for COVID-19 i.e., predict the number of people that can get infected by this disease. It is worth mentioning that our proposed models were only tested on the number of people being infected by Coronavirus and confirmed cases. These models do not consider other factors such as recoveries, deaths, and restrictions being implemented that reduce the chances for a person contracting COVID-19. However, this does not limit the predictions that the models will make as they will follow trends that are continuously being updated within the provided COVID-19 dataset.

6. Conclusion and future work

COVID-19 is a virus that the world was poorly prepared for. The use of machine learning techniques as tools to predict the spread of the virus would allow for greater levels of preparedness through better resource management and distribution based on the prediction made by the models. These models can help prevent more waves of COVID-19 from occurring or even provide groundwork for the creation of similar predictive models for future strains of viruses.

In this paper, we propose a deep learning model DSPM and a non-parametric machine learning model NRM to automatically predict the spread of COVID-19. The proposed models have been trained and tested on a publicly available dataset. As reported in the paper, our proposed models successfully predict the spread of COVID-19 with low error rates. NRM was deemed the most accurate model to be used to predict the spread of the virus due to its low MAE and error rate (0.6%). The performance of our DSPM model was on par with NRM, as DSPM had lower overall error rates compared to cases per specific country and province. It can
be concluded that the proposed DSPM and the NRM models have the potential to predict the spread of COVID-19 in the future. Our experimental results also demonstrate the superior performance of the proposed techniques compared to existing approaches and our baseline.

In our future work, we intend to fuse DSPM and NRM features to refine the prediction of the proposed models. We would also train our model on additional data (as the publicly available dataset is being regularly updated) to further improve the prediction performance of our proposed techniques.

Credit author statement

All the authors have equal contribution.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgments

This research is supported by Edith Cowan University and Murdoch University, Australia.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.iswa.2022.200068.

References

Alamoudi, A., Zaidan, B., Zaidan, A., Albahri, O., Mohammed, K., Malik, R., et al. (2020). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert Systems with Applications, 167, 114155.

Atlam, E.-S., Ewi, A., Abd El-Raouf, M., Gheoneim, O., & Gad, I. (2021). A new approach in identifying the psychological impact of COVID-19 on university student’s academic performance. Alexandria Engineering Journal, 61, 5223–5233.

Cai, C., Han, L., Chen, X., Cao, Z., & Chen, Y. (2005). Prediction of functional class of the SARS-CoV-2 proteins by a statistical learning method. Journal of Proteome Research, 4(5), 1855–1862.

Chakraborty, S., & Malik, K. (2020). SufnoTSA: A superpixel and meta-heuristic based fuzzy image segmentation approach to explicate COVID-19 radiological images. Expert Systems with Applications, 167, 114142.

Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2020). Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble. Expert Systems with Applications, 165, 113909.

D’Angelo, G., & Palmieri, F. (2021). Enhancing COVID-19 tracking apps with human activity recognition using a deep convolutional neural network and HAR-images. Neural Computing and Applications, 33, 1–17.

Elmousalami, H. H., & Hassanien, A. E. (2020). Day level forecasting for coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations. arXiv preprint arXiv:2003.07778.

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons and Fractals, 134, 109761.

Garain, A., Basu, A., Giampaolo, F., Velazquez, J. D., & Sarkar, R. (2021). Detection of COVID-19 from CT scan images: A spiking neural network-based approach. Neural Computing and Applications, 33, 1–14.

Ismael, A. M., & Sengar, A. (2020). Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Systems with Applications, 164, 114054.

John, M., & Shaiba, H. (2019). Main factors influencing recovery in MERS CoV-patients using machine learning. Journal of Infection and Public Health, 12(5), 700–704.

John Hopkins University (2020). COVID-19 novel coronavirus EDA & forecasting cases. https://github.com/CSSEGISandData/COVID-19.

Li, H., Zheng, F., Zhong, Z., Xu, C., Roma, N., Lankin, S., et al. (2022). Stress prediction using micro-EMA and machine learning during COVID-19 social isolation. Smart Health, 23, 100242.

Loey, M., Manogaran, G., & Khalifa, N. E. M. (2020). A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest ct radiography digital images. Neural Computing and Applications, 22, 1–13.

Maliki, Z., Atlam, E.-S., Ewi, A., Dagnew, C., Gheoneim, O. A., Mohamed, A. A., et al. (2021). The COVID-19 pandemic: Prediction study based on machine learning models. Environmental Science and Pollution Research, 28, 1–11.

Pascana, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In International conference on machine learning (pp. 1310–1318).

Rahimi, L., Chen, F., & Gandomi, A. H. (2021). A review on COVID-19 forecasting models. Neural Computing and Applications, 33, 1–11.

Shah, S., Mulahuwaisa, A., Ghafor, K., & Maghhoodi, H. S. (2020). Prediction of global spread of COVID-19 pandemic: A review and research challenges.

Shah, S. A. A. (2019). Spatial hierarchical analysis deep neural network for RG-3-D object recognition. In Pacific-rim symposium on image and video technology (pp. 183–193). Springer.

SRK (2020). COVID-19 novel coronavirus EDA & forecasting cases. Kaggle, Available: https://www.kaggle.com/khongweihao/covid-19-novel-coronavirus-eda-forecasting-cases.

Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., & Xia, X.-Q. (2015). Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Scientific Reports, 5, 17155.

Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of the Total Environment, 728, 138762.

Xue, Z., Li, P., Zhang, L., Lu, X., Zhu, G., Shen, P., et al. (2021). Multi-modal co-learning for liver lesion segmentation on PET-CT images. IEEE Transactions on Medical Imaging, 40(12), 3531–3542.

Yan, S. (2015). Understanding LSTM networks. Online. Accessed on August 11.

Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zainn, M., et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease, 12(3), 165.

Zhao, Z., Li, X., Liu, F., Zhu, G., Ma, C., & Wang, L. (2020). Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Science of the Total Environment, 729, 138959.