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.
Correlating Dynamic Climate Conditions and Socioeconomic-Governmental Factors to Spatiotemporal Spread of COVID-19 via Semantic Segmentation Deep Learning Analysis

Alvin Wei Ze Chew¹, Ying Wang², Limao Zhang³*

¹Research Engineer, Ph.D., Bentley Systems Research Office, Singapore, 1 Harbourfront Pl, HarbourFront Tower One, Singapore 098633
²Ph.D. candidate, School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
³Assistant Professor, Ph.D., School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798.

*Corresponding author: L. Zhang, Assistant Professor, Ph.D., School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. (limao.zhang@ntu.edu.sg)

Abstract

In this study, we develop a deep learning model to forecast the transmission rate of COVID-19 globally, via a proposed G parameter, as a function of fused data features which encompass selected climate conditions, socioeconomic and restrictive governmental factors. A 2-step optimization process is adopted for the model’s data fusion component which systematically performs the following: (Step I) determining the optimal climate feature which can achieve good precision score (> 70%) when predicting the spatial classes distribution of the G parameter on a global scale consisting of 251 countries, followed by (Step II) fusing the optimal climate feature with 11 selected socioeconomic-governmental factors to further improve the model’s predictive capability. By far, the obtained results from the model’s testing step indicate that land surface day temperature (LSDT) has the strongest correlation with the global G parameter over time by achieving an average precision score of 72%. When coupled with relevant socioeconomic-governmental factors, the model’s average precision score improves to 77%. At the local scale analysis for selected countries, our proposed model can provide insights into the relationship between the fused data features and the respective local G parameter by achieving an average accuracy score of 79%.

Highlights

- Data fusion between optimal climate feature and socioeconomic-governmental factors
- 2-step optimization workflow to fuse data features for training deep learning model
- Land surface temperature day correlates with COVID-19 spread with precision of 72%
- Fused features achieve higher precision and accuracy score of 77% and 79%
- Trained model can directly predict global and local growth rate in COVID-19 cases

Keywords
COVID-19; climate conditions; socioeconomic-governmental nexus; semantic segmentation analysis; deep learning; satellite images

1. Introduction
The ongoing novel coronavirus disease 2019 (COVID-19) has already been considered as one of the greatest challenges faced by humanity in the 21st century. On 11 Mar 2020, COVID-19 was officially declared as a global pandemic by the World Health Organization (Mbbs et al. 2020). As compared to the 2003 SARS outbreak in Asia, COVID-19 is much more infectious, hence bringing about new norms of lifestyles such as enforced safe distancing in public places, compulsory masks wearing, etc. At present, the number of confirmed COVID-19 cases worldwide has already crossed 190 million, while the number of deaths due to this new virus exceeds 4 million globally. While vaccinations are gradually deployed worldwide, there are still unanswered questions pertaining to the nature of this complex disease and its variants. One specific area of concern is the unknown relationship between climate conditions and the transmission rate of COVID-19. For example, Germany, France, and the UK imposed their 2nd nationwide lockdown in late 2020 due to a spike in their numbers of confirmed COVID-19 cases which, at the same time, coincided with their winter season.

The spread of coronavirus (SARS-related) in urbanized cities was previously shown to be dependent on a multitude of climate parameters, specifically temperature, wind speed, and humidity (Dalziel et al. 2018; Maiti et al. 2021; Viezzer and Biondi 2021). Yuan et al. (2006) demonstrated that the 2003 SARS outbreak in Beijing, China peaked at the following conditions: (a) mean temperature of 16.9°C; (b) mean relative humidity of 52.2%; (c) wind speed of 2.8 m/s. Most recently, Prata et al. (2020) highlighted that the daily number of confirmed COVID-19 cases in Brazil had an inverse relationship with temperatures ranging between 16.8°C and 27.4°C from Feb to Mar 2020, while Şahin (2020) observed a correlation between the spread of COVID-19 in several cities of Turkey with wind speed values recorded 14 days ago and the temperature of the actual day for the total number of confirmed cases reported. Other similar studies for specific countries, such as China, Spain, and the United States (Abdollahi and Rahbaralam 2020; Liu et al. 2020; Runkle et al. 2020; Shi et al. 2020), have also been performed since the virus’ inception.

There have also been recent studies made that informed positive or minimum correlations (Ahmadi et al. 2020; Bashir et al. 2020; Tosepu et al. 2020; Xie and Zhu 2020; Yao et al. 2020) between the transmission rate of COVID-19 and climate conditions, especially for cases where data is generally limited. Analysis for large-scale countries based upon multiple climate zones was also restricted and inconsistent in several reported results (Awasthi et al. 2020; Baker et al. 2020; Chiyomaru and Takemoto 2020; Islam et al. 2020; Le et al. 2020; Pan et al. 2021; Sobral et al. 2020). On a broader aspect, Wu et al. (2020) and Guo et al. (2021) investigated the effects of ambient temperature on the spread of COVID-19 worldwide where they reported a negative correlation between both parameters, hence highlighting the likelihood that higher temperatures can be useful to control the global transmission rate of COVID-19.

In general, the above-mentioned studies indicate that the transmission rate of COVID-19 can be directly or indirectly associated with different climate conditions which, however, remains an open and important research question to be investigated further. In March 2021, the United Nations highlighted the possibility of COVID-19 becoming a seasonal disease by pointing out that respiratory-related viral infections are often seasonal, especially during autumn-winter peak period for influenza and temperate conditions resulting in cold-related coronaviruses. On the other hand, it is also worth noting that almost all documented pandemics in humanity’s history are originated from the exploitation of the natural environment (e.g., wildlife) which is directly linked to climate change (Rodó et al. 2021). For example, due to economic development and globalization for the past several
decades, the emergence of global warming has caused wildlife to abscond to the colder poles to avoid the increasing global temperatures. Consequently, there is a likelihood that they interact with other wildlife which they usually would not, hence creating opportunities for unknown pathogens and viruses to get into the new hosts. Likewise, continuous deforestation forces animals to migrate and contact with other wildlife, hence increasing the tendency for transmission of unknown pathogens and viruses which can subsequently lead to a spillover of infections from wildlife to humanity (Rodó et al. 2021). Overall, we can recognize that climate can be linked to COVID-19, or to pandemic occurrences in general, both on short- and long-term basis, hence underlining the need for natural environment protections, as part of the present global sustainability mission, to minimize the risk of future outbreaks/pandemics.

While temporal climate conditions do affect (directly or indirectly) the spread of COVID-19, the exact seasonality and dynamic association between COVID-19 and climate conditions are still not well understood quantitatively in today’s context. Continuous studies are therefore required to correlate the relevant climate conditions to the present virus spread. It is hopeful that epidemiological prediction models which accounts for climate and weather processes can assist humanity to identify populations who are at risk towards COVID-19, and thus improve our overall surveillance and control measures (Kroumpouzos et al. 2020). At the same time, it is of common knowledge that socioeconomic factors and restrictive governmental policies do also control the transmission rate of COVID-19 in the different countries. Hence, it is also imperative that any novel prediction models for COVID-19 must also account for the relevant socioeconomic-governmental factors, atop of selected climate features, for modelling the spreading rate of COVID-19 both on the global and local contexts. In our present study, the focus lies in developing an alternative prediction tool which can fuse the different types of data features pertaining to climate conditions, socioeconomic factors, and restrictive governmental policies to model and forecast the spatiotemporal growth rate in the number of confirmed COVID-19 cases at both the global and local scales. We note that the global scale relates to an aggregation of 251 countries globally in our analysis, as subject to the data availability, while the local scale corresponds to any specific country based upon the modeler’s choice. In addition, we propose an optimization workflow for the data fusion process in the model development phase which can attain a balance among the computational runtime for training the model, logic flow, and the model’s resulting predictive accuracy/precision.

This paper is thus structured as follows. Section 2 reviews related studies which model the complex relationship between the transmission rate of COVID-19 and multiple factors pertaining climate, socioeconomic conditions and/or restrictive governmental policies. Section 3 describes the selected data features, pertaining to the combined climate-socioeconomic-governmental factors, and the output targets for our proposed image-to-image semantic segmentation analysis to model and forecast the spatial classes distribution for the rate of increase in the number of confirmed COVID-19 cases across the different countries spatially, via a proposed G parameter. Section 4 summarizes the prediction results derived from the model’s testing step, in relation to the proposed classification task for the G parameter via segmentation analysis using the selected combination(s) of data features. Finally, we investigate several major countries from the different continents to examine the level of agreement between the actual and predicted classes for the respective G parameter on a local scale, hence ensuring that the proposed prediction model can provide useful insights at both the global and local contexts. Section 5 compares the model’s prediction capability with other relevant studies, followed by discussing on the implications of our proposed prediction tool and how the tool can be
useful to the relevant stakeholders to build towards a post-Covid era. Finally, Section 6 succinctly summarizes the significance of our study and the key findings obtained by far.

2. Literature review

As the global battle against COVID-19 continues, new information analysing the complex relationship between the transmission rate of COVID-19 and a multitude of factors which include, but not limited to, climate conditions, social control measures, economic stability, etc., has emerged continuously. In the following, we provide a literature overview on the critically important and relevant studies carried out by far on the above-listed relationship(s), which have been useful to contribute to the scientific community’s qualitative and quantitative understanding on the current topic.

Investigations between climate conditions and the spread of COVID-19 have, by far to the best of our knowledge, been focused on using empirical one-dimensional (1D) quantitative formulation which aggregates multiple meteorological parameters/features via calibrated weightages using available data. For example, Wu et al. (2020) proposed a novel core generalized additive model (GAM) equation which predicts the number of daily new cases or the number of deaths in the different countries with a lead-time of 1 day as a function of wind speed, the population’s median age, time, weather variables, etc. It is, however, likely that the selected input features into their prediction model are constricted to an extent where the features remain static during the model’s computational runs, hence not maximising the model’s predictive accuracy. Guo et al. (2021) used multiple average meteorological parameters, which are first weighted via population density, for the modelling step at the city- or country-level. Likewise, the number of input features does not augment during the computational runs, hence constricting the data-driven analysis. On a broader aspect for analysing infectious diseases, Kim (2021) leveraged on exploratory spatial data analysis and one-dimensional (1D) spatial regression method to model infectious diseases, specifically measles and the 2015 Middle East Respiratory Syndrome (MERS), emergence data and 14 urban characteristics by analysing 225 spatial urban units in South Korea. The author’s results demonstrated that new infectious diseases such as COVID-19 can be influenced by ecological factors, i.e., climate change. At the same time, the author also highlighted on the need to model the spatial relationship between weather conditions and the spread of COVID-19 in their future studies to better quantify their complex relationship.

To further underline the important relationship between climate conditions and the transmission rate of COVID-19, it is well-known that the virus, as caused by SARS-CoV-2, was first detected in low temperature areas of China towards the end of 2019, followed by major outbreaks in Japan, South Korea, Northern Italy, and Germany which too were experiencing their own winter periods (Kroumpouzos et al. 2020). Observed cold weather patterns pertaining to COVID-19 (5°C - 11°C air temperature, and 47% - 70% relative humidity) are known laboratory conditions that also favour CoV survival (4°C air temperature, 20% - 80% relative humidity) (Casanova et al. 2010). On the other hand, the gradual spread, and subsequent epicentre development, to India, Thailand, Singapore, Middle East, etc., which usually experience higher temperatures during the year, is attributed to global travelling and close human contact before the respective national lockdowns and air-travel controls. Therefore, while it is important for us to recognize that climate conditions do affect the transmission rate of COVID-19 globally, socioeconomic factors, and/or restrictive
governmental policies also influence the level of virus spread in the different countries globally. For example, the virus transmission rate generally depends on the living environments, population densities, hygiene, space availability, etc., which can significantly affect the virus spread in relatively warmer areas (Huang et al. 2020; Li et al. 2020).

Since the pandemic’s inception, several important studies have been made to investigate the relationship between different socioeconomic factors, as highlighted above, and the transmission rate of COVID-19 on a whole. For example, Li et al. (2021) performed clustering analysis and developed a novel structural model equation to quantify the relationship(s) between built environment attributes and the cluster size of COVID-19 in Hangzhou, China. Their results, by far, demonstrated statistically significant influences of commercial strength and the availability of transportation infrastructures on the number of confirmed cases in any infectious cluster. Similar clustering analysis has also been performed by Das et al. (2021) in the context of Kolkata megacity, India by correlating the level of living environment deprivation on the spatial clustering of COVID-19 hotspots in Kolkata. Multiple data regression models were also leveraged for the modelling analysis, where zero-inflated negative binomial regression (ZINBR) was found to best correlate the living environment deprivation factor with the spatial clustering of COVID-19 hotspots in Kolkata. Sannigrahi et al. (2020) also used different 1D regression models to determine that total population size and income level are most important to regulate the number of death cases due to COVID-19 in the entire European region. At the same time, the authors also highlighted on the need to model the influences of environmental conditions and socioecological statuses on the transmission and death rates of COVID-19 in future studies. Other similar studies focusing on investigating the quantitative relationships between defined socioeconomic-environmental nexus and the transmission rate(s) of COVID-19 in different countries or city states such as Washington, America (Hu et al. 2021), New York City, America (Fu and Zhai 2021), or America as a whole (Maiti et al. 2021) have been well-documented. Those studies generally report that income level, level of social distancing, and stay-at-home behaviours are key socioeconomic factors for controlling the transmission and death rates of COVID-19 in urbanized cities.

The effect of social distancing has been investigated by several studies. For example, Ugail et al. (2021) proposed a novel methodology which leverages on a well-known circle packing problem to re-configure common physical environments to generate the optimal design space in near real-time for the objective of minimizing the infectious rate of COVID-19 among the group of individuals within the defined space. On the other hand, Sun and Zhai (2020) innovated on the traditional perfect-mixing-based Wells-Riley model by introducing 2 new indices which modelled the effects of social distancing and space ventilation on the infection probability of COVID-19 in building spaces and transportation vehicles. Overall, their studies concluded that increasing social distance can significantly reduce the infection rate by an average of 30% during the first 30 min with appropriate ventilation conditions.

In summary, the above-outlined research studies have been very useful to improve the community’s qualitative and quantitative understanding of the relationship between the different combinations of environmental conditions and socioeconomic factors, and the transmission rate of COVID-19 as the target parameter. However, going forward, several research gaps, which have also been highlighted by previous research works summarized above, can be delineated at this stage, namely:
The previous studies, as discussed above, mainly focus on individual category of features (e.g., socioeconomic factors under one main category) to model either the transmission or/and death rate(s) of COVID-19. However, we are aware that there is a complex climate-socioeconomic-governmental nexus (and may be larger) which influences the level of COVID-19 spread in the different countries. By far, to the best of our knowledge, we note that there has been no prediction tool being developed to predict/classify the spatiotemporal growth rate in the number of confirmed COVID-19 cases, both on the global and local scales, as a function of climate conditions, socioeconomic factors, and governmental restrictive policies as an entirety.

Most of the previous research studies focused on specific case studies which include, but not limited to, America, India, China, etc. On the contrary, we are hopeful that a singular competent prediction model can be developed, and built upon beyond this study, to predict/classify the growth rate in the number of confirmed COVID-19 cases for any specific country of the modeler’s choice with a defined lead-time.

To the best of our knowledge, there have been no studies made, by far, which leverage on two-dimensional (2D) input features to model the spatiotemporal growth rate in the number of confirmed COVID-19 cases both on the global and local scales. In general, 2D or 3D model features such as image representations for defined data features can provide more insights into their complex dynamics and how they may better correlate with the transmission rate of COVID-19 worldwide due to their higher data dimensions. On the contrary, 1D time-series regression modelling, which is commonly used in the above-discussed studies, may limit the data-driven analysis due to lower data dimensions involved in the input features. We are thus hopeful that by using 2D input features for our proposed deep learning method can model the complex and high-dimensional correlation between the spatiotemporal growth rate in the number of confirmed COVID-19 cases (as the target objective) and the combination of climate conditions, socioeconomic factors, and restrictive governmental policies (as the data input features) on both the global and local scales.

In this study, we thus develop a personalized encoder-decoder deep learning architecture which is capable of extracting high-level features from 2D image representations for a range of input data parameters which include defined climate parameters, socioeconomic factors, and restrictive governmental policies. For convenience, we term the combination of the different data features as “climate-socioeconomic-governmental features”. The extracted high-level features from the encoder component of the deep learning model are then further processed to reconstruct 2D map representations which quantify the spatial classes distributions for the rate of increase in the number of confirmed COVID-19 cases worldwide, via a proposed G parameter. To the best of our
knowledge, this form of data exploitation for modelling the spread of COVID-19 has not been investigated before in the literature since the virus’ inception. The team is thus hopeful that the formulated approach of using 2D image representations of the climate-socioeconomic-governmental features to model and forecast the proposed G parameter, both on the global and local scales, can support resilience and sustainability development in a post-Covid era where COVID-19 is expected to become an endemic in our society. For example, if the COVID-19 virus is found to correlate with seasonal weather variations and would result in higher infectious rates during certain periods, the quantitative information from the proposed prediction model can thus assist healthcare management to better prepare for possible spikes in community cases, especially when various socioeconomic factors, which are modelled as other input data parameters in the proposed tool, such as economic activities and global travelling are back to normality.

3. Data and methodology

In the following, we first describe, in-detail, the selected data parameters pertaining to the defined climate-socioeconomic-governmental features which will be leveraged to train our prediction tool to model and forecast the proposed G parameter, as representative of the rate of increase in the number of confirmed COVID-19 cases, both on the global and local scales via classification analysis. Details relating to our deep learning model development, coupled with the proposed data fusion optimization process, are then subsequently provided.

3.1. Data description of climate features

The dynamic climate conditions/features are quantified using open-source two-dimensional (2D) satellite images which are first extracted from NASA Earth Observations (NEO) portal (https://neo.sci.gsfc.nasa.gov/). Our current analysis focuses on multiple climate parameters, namely: (a) Land Surface Temperature Day (LSTD); (b) Land Surface Temperature Night (LSTN); (c) Cloud Fraction (CF); (d) Cloud Optical Thickness (COT); and (e) Cloud Particle Radius (CPR), as subjected to their data availability ranging between Jan 2020 and Oct 2020 which aligned with the global COVID-19 pandemic situation. Table 1 describes the characteristic of each climate feature and their possible relationship with the transmission rate of COVID-19 globally. Note that NASA satellite images for the different climate parameters are collected via Moderate Resolution Imaging Spectroradiometer (MODIS), which is an instrument used for NASA’s Terra and Aqua satellites. As summarized in Table 1, it is likely that various climate features may have some level of overlap relationship with one another. For example, CF and COT features directly affect the amount of sunlight entering earth’s lower atmosphere and ground level, hence influencing the LSTD and LSTN variations over space and time. As such, it is important to first evaluate which climate feature quantitatively is most influential in affecting the transmission rate of COVID-19 to avoid features overlapping, before fusing with the other defined socioeconomic-governmental factors as part of the data fusion optimization process which will be described in subsequent sub-section(s).

For each of the selected climate features, 257 quantities of 2D NASA satellite images, ranging between 22 Jan 2020 and 4 Oct 2020 daily, having dimensions of 180 (width) and 360 (height) pixels/grids respectively have been collated for our deep learning analysis. Examples of the 2D images for LSTD, LSTN, CF, COT, and CPR are shown in Figure 1a to 1e, respectively. Note that a 2D image can be represented numerically as a 3D array with the 3rd parameter pertaining to its
number of features to be modelled, while the 1st and 2nd parameters define the width and height of the corresponding image itself. For example, for typical red-green-blue (RGB) images, the 3rd parameter is represented as a value of 3 to represent the total number of RGB features where each colour feature consists of their own 2D array. As such, the original 2D image can be quoted as \( w \times h \times 3 \) where \( w \) and \( h \) represent the width and height of the image respectively.

In our study, we instead convert the RGB colour representation of each 2D image into their normalized quantitative values, with respect to the known range of values from Table 2, for each specific climate feature by associating with their respective colour tone at each grid/pixel in the original image. As a rule of thumb, the brighter colours (red, orange) generally represent the higher ranges of values for each climate feature as summarized in Table 2, and vice versa for the darker colours (blue). By doing so, we can thus analyse every 2D satellite image with a single feature in their corresponding 3D array, where the 3rd dimension is now represented by a single value which represents the colour intensity value instead of the conventional RGB features.

When processing each 2D image, we compute the mean and standard deviation values for each grid/pixel by using their respective range of values (from Table 2) for the corresponding climate feature. Each valid value in every grid/pixel location is then normalized via the following scaling formula:

\[
\overline{X}_J = \frac{x_j - \mu}{s} \quad (1)
\]

where \( \overline{X}_J \) represents the normalized value for grid/pixel \( J \), \( x_j \) represents the original value for pixel/grid \( J \) which excludes the background representation (black colour in Figures 1a-1e), \( \mu \) represents the mean value of all valid values in the original 2D image, and \( s \) represents the standard deviation value of all valid values in the same 2D image.

| Climate feature | Description | Possible relationship with the spread of COVID-19 globally |
|-----------------|-------------|----------------------------------------------------------|
| Land Surface Temperature Day (LSTD) | LSTD represents the skin temperature of whatever is on the land (snow, ice, human-made structures, etc.). Note that LSTD is not indicative or equivalent to surface air temperature values. LSTD can also be correlated to changes in weather and climate patterns. For example, how does the increase in atmospheric greenhouse gases affects LSTD variations over time. | The temperature has been speculated to influence the spread of COVID-19, where countries with colder land temperatures may face greater challenges in controlling its spread rate. For example, in early Nov 2020, Germany, France, and the UK imposed their 2nd lock-down period which aligned with the start of their winter season. |
| Land Surface Temperature Night (LSTN) | Identical in its data collection and investigative purpose as that of LSTD, with the exception that LSTN data are collected during the nights. | Same speculated relationship to COVID-19 as that of LSTD. |
| Cloud Fraction (CF) | CF is generally indicative of the amount of sunlight energy that reaches the Earth from the Sun as well as the amount of energy that radiates from Earth and back into space, hence affecting Earth’s temperature | Building upon LSTD and LSTN, we need to investigate any possible relationship between CF and the transmission rate of COVID-19 globally |
and humidity in different locations. As a rule of thumb, the higher the CF values, the greater the amount of sunlight is being reflected back into space, and at the same time, higher CF values can also trap a greater amount of heat radiated from the earth’s surface and preventing them from escaping back into space.

Cloud Optical Thickness (COT): COT generally measures the amount of sunlight passing through the clouds for reaching the Earth’s surface. As a rule of thumb, the higher the COT values, the greater the amount of sunlight is being scattered and reflected by the available clouds. Since COT and CPR affect the amount of sunlight reaching the Earth’s surface, the earth’s land temperature is again affected. Therefore, it is again useful to investigate the relationship between COT/CPR and the spread of COVID-19 globally as discussed in the preceding climate features above.

Cloud Particle Radius (CPR): Likewise, to COT, CPR contributes to the amount of sunlight passing through the clouds to reach the Earth’s surface. Generally, with smaller particles, brighter and more reflective clouds are generated which reflect sunlight back into space and cool the planet. Clouds with larger particles absorb more shortwave infrared light and, conversely, clouds with smaller particles absorb less shortwave infrared light.

*Climate features image data extracted from [https://neo.sci.gsfc.nasa.gov/](https://neo.sci.gsfc.nasa.gov/)

**Figure 1. Typical RGB satellite images for multiple climate features: (a) LSTD; (b) LSTN; (c) CF; (d) COT; (e) CPR. (Image data extracted from [https://neo.sci.gsfc.nasa.gov/](https://neo.sci.gsfc.nasa.gov/))**

**Table 2. Valid ranges of values for each climate feature.**

| Climate feature | The valid range of values       |
|-----------------|---------------------------------|
| LSTN            | -25.0°C to 45.0°C               |
| LSTD            | -25.0°C to 45.0°C               |
| CF              | 0.0 to 1.0 (dimensionless)      |
| COT             | 0.0 to 50.0 (dimensionless)     |
| CPR             | 4.0µm to 40.0µm                 |

*Climate features image data extracted from [https://neo.sci.gsfc.nasa.gov/](https://neo.sci.gsfc.nasa.gov/)

**3.2. Data description of socioeconomic factors & restrictive governmental policies**

As discussed, besides considering the climate features as the input data for the prediction model, other relevant socioeconomic factors, and restrictive governmental policies, which are associated with the transmission rate of COVID-19 both on the global and local scales, are also included as the model’s input features as part of the data fusion step in our analysis. The relevant socioeconomic-governmental factors (total of 11 additional data features) considered are extracted from another open-source database in [https://ourworldindata.org/coronavirus-data](https://ourworldindata.org/coronavirus-data), as summarized in Table 3. For all listed factors in Table 3, except for “Government stringency index”, their representations are
categorical via the use of unique classes for different ranges of values. For example, the “Level of income support” factor is quantified via 3 labelled classes of, namely: (Class 1) no support for loss salary; (Class 2) support > 50% of loss salary; and (Class 3) support < 50% of loss salary. Hence, to set up the common basis to correlate with the climate features, which are already represented as normalized floating values, all categorical data features are also normalized into discrete values via the same normalization formula from Equation (1), where the mean and standard deviation values are now computed via the data distribution of the number of classes for each of the data features. We also note that the same normalization procedure is undertaken for the discrete values of the “Government stringency index” factor to set up the same scaling of values for all data features, inclusive of the different climate features.

Table 3. List of socioeconomic factors and restrictive government policies for data fusion with climate features.

| Type                      | Data features                             | Value range                                                                                      |
|---------------------------|-------------------------------------------|--------------------------------------------------------------------------------------------------|
| Socioeconomic factors     | a. Level of income support                | 3 classes (no support, support > 50% loss salary, support < 50% loss salary)                     |
|                           | b. Level of debt relief                   | 3 classes (no relief, moderate relief, large relief)                                             |
|                           | c. Face covering                          | 5 classes (no policy, recommended, required in some public places, required in all public places, always required outside of homes) |
|                           | d. Cancellation of public events and gatherings | 3 classes (no cancellations, recommended cancellations, compulsory cancellations)             |
|                           | e. Testing and contact tracing            | 4 classes (no testing, testing for those with symptoms and belonging to higher-risk groups, testing for anyone with symptoms, open public testing) |
|                           | f. Public information campaigns           | 3 classes (none, public officials urging caution, coordinated information campaigns)            |
| Restrictive government policies/factors | g. Government stringency index            | 0-100 (100 being the highest stringency score)                                                    |
|                           | h. Schools and workplaces closures         | 4 classes (no closures, recommended, recommended at some levels, required at all levels)         |
|                           | i. Stay-at-home restrictions               | 4 classes (no measures, recommended, required except for running essentials, required with few exceptions) |
|                           | j. Controls for domestic travels          | 3 classes (no measures, recommended movement restrictions, restricted movement)                |
|                           | k. Controls for international travels      | 5 classes (no measures, screening, quarantine from high-risk countries, ban on high-risk countries, total border closure) |

*Data features data extracted from [https://ourworldindata.org/coronavirus-data](https://ourworldindata.org/coronavirus-data)
As discussed earlier, each of the listed climate feature from Table 1 is represented as a 2D image, i.e., 3D array, of \(w \times h \times 1\) where the value of 1 for the 3rd dimension represents the single colour intensity feature. Data fusion of the above-listed 11 socioeconomic-governamental factors is then performed by constructing an initial 3D array for each of the climate feature, followed by concatenating with their respective colour intensity feature, hence resulting in a 3D array size of \(w \times h \times 12\) for each of the climate feature from Table 1. Note that \(w\) (width) and \(h\) (height) of each 3D array are still fixed as 180 and 360 which controls the total number of pixels/grids in each image (3D array).

3.3. Data description of COVID-19 confirmed cases

The global recorded numbers for the confirmed COVID-19 cases are extracted from another open-source database (https://ourworldindata.org/coronavirus-data). As mentioned earlier, we propose a \(G\) parameter \((G_{\text{global}})\), as defined in Equation (2), which measures the daily growth rate in the total number of confirmed cases on the global scale. At this stage, the global scale encompasses 251 countries in total, where the respective total number of the confirmed cases for each country can be extracted from the same above-stated database. Figure 2a illustrates the cumulated number of confirmed cases globally for the selected period between 22 Jan 2020 and 4 Oct 2020, while Figure 2b represents the global variation of the proposed \(G_{\text{global}}\) parameter for the same period.

\[
G_{\text{global}} = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100\% \quad (2)
\]

where \(Y_t\) represents the global number of confirmed COVID-19 cases at time \((t)\), and \(Y_{t-1}\) represents the global number of COVID-19 cases at time \((t-1)\) from the previous day. For example, the \(G\) value on 31 January 2020 \((t)\) is computed by dividing the difference between the respective number of COVID-19 confirmed cases on 31 January \((t)\) and 30 January 2020 \((t-1)\) by the number of COVID-19 confirmed cases from the latter’s date.

At the same time, we also compute the individual \(G_{\text{local}}\) values for all 251 countries using their respective reported numbers of the confirmed COVID-19 cases between 22 Jan 2020 and 4 Oct 2020. The computed \(G_{\text{local}}\) values (in percentage quantities) are then group into 6 main classes, as summarized below:

- Class 0 – \(G_{\text{local}} \leq 1\%
- Class 1 – 1% < \(G_{\text{local}} \leq 2.5\%
- Class 2 – 2.5% < \(G_{\text{local}} \leq 5.0\%
- Class 3 – 5.0% < \(G_{\text{local}} \leq 10.0\%
- Class 4 – 10.0% < \(G_{\text{local}} \leq 25.0\%
- Class 5 – \(G_{\text{local}} > 25.0\%

**Figure 2.** Global variations in the total confirmed number of COVID-19 cases for the period between 22 Jan 2020 and 4 Oct 2020: (a) absolute numbers; and (b) proposed \(G_t\) parameter. (Data extracted from https://ourworldindata.org/coronavirus-data)

3.4. Data fusion and optimization workflow
As part of our data fusion process, the multiple parameters pertaining to the defined climate-socioeconomic-governmental factors are fused accordingly to construct a new 2D input image (3D array) of the defined size of 180 x 360 x d (64,800 x d grids) where each grid represents the normalized value(s) for the defined number of data features (d) being fused together. The resulting output target is also a 2D map representation having the same dimensions of 180 x 360 x 1, where the 3rd parameter now represents a specific class type (Class 0 – 5) as indicative of the rate of increase in the confirmed cases in earlier defined $G_{t,\text{local}}$ parameter. The 251 countries are then mapped accordingly to assigned grids/pixels on the 2D input and target maps, based upon their known longitudinal and latitude global coordinates for the defined size of the 2D maps.

For example, Figure 3 illustrates that if Country X occupies a selected pool of pixels in the defined 2D map of 180 x 360, then the same pool of pixels in the input and output maps of identical width and height dimensions will be tagged respectively with the normalized values for the modelled data features (fusion of climate-socioeconomic-governmental factors) and the resulting classes for $G_{t,\text{local}}$ as shown. The 2D input maps for the fused features will then be fed into our personalized deep learning encoder-decoder model to model the defined 2D output maps for the $G_{t,\text{local}}$ classes for all 251 countries at their respective pixels/grids, as part of the direct image-to-image segmentation analysis. More details pertaining to our proposed deep learning model will be provided in the next section.

**Figure 3.** Example of generating input and output 2D maps, i.e., 3D arrays, based upon assigned pixels/grids for specific countries.

To optimize the features selection process during the prediction model’s training phase, Figure 4 illustrates a 2-step optimization workflow to determine the optimal set of fused data features which can best maximize the trained model’s predictive performance on a defined testing dataset. Descriptions of the 2 steps involved in the proposed optimization workflow from Figure 4 are as follows:

**Step I:** Training of our proposed deep learning prediction model with each of the selected 5 climate features (LSTD, LSTN, CF, COT, and CPR), and determine which optimal climate feature can maximize the model’s precision score on the defined testing dataset by attaining at least 70% precision score as our first optimization score. Again, we note that this step is performed to avoid the tendency of overlapping climate features, as described earlier, where the multiple features can be associated with one another. For example, the dependence of LSTN and LSTD on COT and CF as summarized in Table 1.

**Step II:** Data fusion of optimal climate feature with the 11 listed socioeconomic-governmental factors (from Table 3), followed by performing the same model training process from Step I to further optimize the prediction model’s predictive performance on the same testing dataset.

**Figure 4.** Example of generating input and output 2D maps, i.e., 3D arrays, based upon assigned pixels/grids for specific countries.

**Step I: Optimize selection of climate feature**

To determine the optimal climate feature from the initial model training step in Step I, we investigate a series of Scenarios for mapping the defined 2D maps of the different independent climate feature to 2D target maps, having the same dimensions, which represent the spatial classes distributions of the proposed $G_{t,\text{local}}$ metric for the same period between 22 Jan 2020 and 4 Oct.
2020. Table 4 summarizes the series of Scenarios where each main Scenario investigates the relationship of specific climate feature with the different classes of $G_t$ for all 251 countries over time, while the Sub-Scenario within the main Scenario investigates the effects of varying lead-times of the same climate feature on the modelled output target. We again note that the locations of the different countries are appropriately mapped to specific pixels on the defined 2D input and output maps based upon their respective known longitudinal and latitude global coordinates, even though the original satellite images for CF, COT, and CPR features do not contain the specific contours of the 251 countries as that of Figure 1a and 1b. As illustrated in Figure 4, the optimization objective in Step I is also to determine the optimal set of hyperparameters (batch size, epochs, type of Sub-Scenario) which can result in specific climate feature capable of maximizing the model’s precision score to be above 70% at least when evaluating the model’s predictions on the defined testing dataset. The selected hyperparameters will then be used the primary set for training the same prediction model using the fused data features in the subsequent Step II, before additional tuning to improve the model’s resulting predictive capability.

**Table 4.** Summary of Scenarios and Sub-Scenarios for mapping defined 2D maps for each climate feature to the same representations for spatial variations of $G_t$ classes for all 251 countries.

| Scenarios | Main Scenario 1 (CF) | Main Scenario 2 (COT) | Main Scenario 3 (CPR) | Main Scenario 4 (LSTD) | Main Scenario 5 (LSTN) |
|-----------|---------------------|-----------------------|-----------------------|------------------------|-----------------------|
| Sub-Scenario a | Using the original 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 3rd day ($X_{t-3}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |
| Sub-Scenario b | Using the average 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 3 days ($X_{t-3}, X_{t-4}, X_{t-5}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |
| Sub-Scenario c | Using the original 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 5th day ($X_{t-5}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |
| Sub-Scenario d | Using the average 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 5 days ($X_{t-5}, X_{t-6}, X_{t-7}, X_{t-8}, X_{t-9}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |
| Sub-Scenario e | Using the original 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 7th day ($X_{t-7}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |
| Sub-Scenario f | Using the average 2D map of the specific climate feature (CF, COT, CPR, LSTD, or LSTN) from the previous leading 7 days ($X_{t-7}, X_{t-8}, X_{t-9}, X_{t-10}, X_{t-11}, X_{t-12}, X_{t-13}$) to model the current state of the spatial variations of the $G_t$ parameter ($Y_t$) classes |

**Step II: Data Fusion**
Upon determining the optimal climate feature from Step I of our proposed optimization workflow from Figure 4, we proceed to fuse the selected climate feature with the 11 listed socioeconomic-governmental factors from Table 3 to construct a new 3D array of (180 x 360 x 12) in size as previously shown in Figure 3. Note that our first approach models all 11 socioeconomic-governmental factors, coupled with the optimal climate feature, to train and validate our proposed deep learning prediction model, following testing the trained model on the same defined testing dataset from Step I. Again, in this Step II, we continue to tune on the model’s hyperparameters to best maximize the model’s resulting predictive performance in terms of the precision and overall accuracy scores. Finally, Algorithm 1 summarizes the key computational steps involved in the entirety of Step (I-II) for our proposed workflow to optimize the data fusion process.

Algorithm 1. 2-Step Optimization Workflow for Data Fusion Process.

1. **Input:** 2D input maps ($X_{climate}$) for each of the 5 climate features, 2D input maps ($X_{SEG}$) compacting all 11 socioeconomic-governmental factors, size of lead-time ($T_{lead}$) for Sub-Scenarios (a-f), values of model’s hyperparameters ($H_1, H_2, \ldots$), threshold values for precision metric ($P_{T1, P_{T2}}$), validation dataset, testing dataset

2. **Output:** 2D output maps ($Y$) representing spatial classes distribution for all 251 countries with defined lead-time using the proposed approach

3. **for each** $X_{climate}$ ← Step I of Optimization Process
   - do
     - a. initialize training of deep learning prediction model with user-defined $H_1, H_2, \ldots$ values & $X_{climate}$ with selected Sub-Scenario (a-f) for defined $T_{lead}$ in Step I optimization process
     - b. derive $Y$ predictions on validation and testing datasets via trained model with defined $T_{lead} \& H_1, H_2, \ldots$ values
     - c. model evaluation of predictions on testing dataset via $P$ score
     - if $P \geq P_{T1}$
       - proceed
       - d. store $P$ score with corresponding $H_1, H_2, \ldots$ values & type of Sub-Scenario for defined $T_{lead}$

4. select optimal $X_{climate}$ with highest $P (> P_{T1})$ score with corresponding $H_1, H_2, \ldots$ values & type of Sub-Scenario for defined $T_{lead}$

5. **do**
   - a. fuse optimal $X_{climate}$ with $X_{SEG}$
   - b. initialize training of deep learning prediction model with selected $H_1, H_2, \ldots$ values & type of Sub-Scenario for defined $T_{lead}$ from (4) ← Step II of Optimization Process
   - c. derive $Y$ predictions on validation and testing datasets with trained model
   - d. model evaluation of predictions on testing dataset via $P$ score
   - if $P \geq P_{T2}$
     - proceed
     - store best prediction model for near real-time predictions
   - else
     - re-train prediction model with additional hyperparameters ($H_1, H_2, \ldots$) tuning

6. **end**

3.5. Proposed deep learning model

The direct semantic segmentation analysis for both Steps (I) and (II) in our proposed optimization workflow (Figure 4) is performed with our own personalized encoder-decoder
algorithm as illustrated in Figure 5. At each epoch (i.e., iteration), the input data feature into the encoder-decoder model is a defined quantity (batch-size) of 3D arrays (2D maps) representing a specific climate parameter, while the output target is the same number of 3D arrays where each array represents the spatial classes distribution of the proposed $G_t$ parameter for all 251 countries in their respective assigned pixels.

Table 5 summarizes the details of the different hidden layers inherent to our proposed encoder-decoder model. Learning rate and the number of epochs runs are maintained at the values of 0.0001 and 50 respectively, and no dropout probability is involved in the model’s training step at this stage. The batch-size of the dataset is varied among 4, 8, 16, and 32 for training the model, while Adam optimizer is used to optimize the batch-size selection and the cross-entropy cost function is selected for the model training and validation steps. Our proposed model uses the mini-batch gradient descent method to minimize the selected cost function, where the different mini-batch sizes (4, 8, 16, 32 in quantities) of the defined 2D input and output maps are designated appropriately with respect to the different lead-times Sub-Scenarios analysed from Table 4. Finally, the model is implemented using Keras version 2.2.4 and TensorFlow flow version 1.14.0 as developed by Google (Tensorflow.org).

**Figure 5.** Personalized encoder-decoder deep learning model, with pipeline parallelism protocol, to perform direct image-to-image segmentation analysis of defined 2D maps of climate features to 2D map representing the spatial classes distribution of the $G_t$ values for all 251 countries.

The principles of convolutional neural network (CNN) establish the basis for our proposed encoder-decoder model, which has shown promising results in many classification tasks using image-based features (Lecun et al. 1998; LeCun and Bengio 1998). A traditional CNN layer learns local features by sliding kernel filters on a feature map, as originally derived from 2D images of defined dimensions, as input features into a deep learning model. As such, there is a continuous transformation of the original low-level input features into high-level features which are expected to maximize the model’s predictive accuracy. In CNN layers, the activation function usually uses the ReLU function, which accelerates the model’s training phase and also avoids gradients disappearing (Schmidhuber 2015). A pooling layer is also used to subsample the output from a CNN layer to reduce the total computational cost, while capturing the high-level features at the same time. A batch normalization layer can also be adopted to avoid model overfitting and to improve the model’s training performance (Ioffe and Szegedy 2015).

To optimize the computational workflow of our proposed model, pipeline parallelism is performed with 2 Nvidia V100 GPU cards on Azure Cloud. The 1st and 2nd GPU cards are affiliated to the encoder (GPU-0) and decoder (GPU-1) architectures (see Figure 1) respectively. Pipeline parallelism is initiated at every epoch as follows. During the forward pass, the first mini-batch of 2D maps for the corresponding climate feature is passed into the encoder architecture and then passing the extracted high-level features into the next decoder architecture via an assigned master thread. The master thread will also store a copy of the encoder model’s initial weights for updating in the later back-propagation step. At the same time, the encoder model will immediately process the next mini-batch of 2D satellite images while the decoder model processes the first mini-batch of extracted high-level features as initially derived from the encoder model. Note that another set of randomly initialized weights is adhered for the encoder model to process the second mini batch of data as the first backpropagation step has not yet been initiated at this stage to update the first set of initialized
weights for the encoder model. However, since significant batches of data will be leveraged for model training and validation at each epoch, we would expect relatively fast convergence of the model’s accuracy evaluation at every epoch regardless of the model’s initial weights.

After completing the first forward pass, the decoder model will then initiate the first backpropagation step by updating its initial weights using the prediction results derived from the first mini-batch of data. At this stage, GPU-0 (encoder model) will remain idle while the Master thread stores a copy of the high-level features extracted from the second mini batch of data using the encoder model. Upon completing the first back-propagation step, GPU-1 (decoder model) will start the forward pass for the second mini-batch of high-level features, as retrieved from the main memory storage from the master thread, derived from GPU-0 earlier. Concurrently, GPU-0 commences on its backpropagation step using the encoder model’s very first initial weights, as retrieved from the same main memory storage, and the respective backpropagation results derived from GPU-1 for the first mini-batch of data. Only after completing its own backpropagation step, GPU-0 will then initiate the forward pass for the 3rd mini-batch of 2D input maps using the encoder’s updated weights from the completed backpropagation step. The same data copying, and operational steps then repeat for both the encoder and decoder models. The summarized difference between the runs with pipeline parallelism (with 2 GPU cards) and without, i.e., with just 1 GPU card, for the first 2 mini-batches of data is outlined in Table 6. Note that both GPU cards are maintained within the same machine, hence no inter-machines communication time is incurred.

Finally, the team proposed an encoder-decoder model for the classification analysis, instead of a traditional classification model, for the following reasons:

- All 251 countries vary in their respective sizes which thus affect their assigned number of pixels within the 2D map. As such, it becomes difficult to design a common input vector size as representative for all countries to perform the modelling step via traditional classification models. To do so, a significant amount of zero paddings is likely required for smaller scale countries which may affect the model’s resulting accuracy for the local scale analysis of those countries.

- By analysing all countries within a single 2D map for each day, it becomes possible to correlate neighbouring countries for the modelled target as we would expect climate conditions to be reasonably close or similar for countries within the same continent or region. The convolutional layers deployed in the encoder-decoder model can capture the boundary relationship among the neighbouring grids as a representation of countries within the same continent or region.

- The team is hopeful of developing a prediction model, via the proposed encoder-decoder model, which is capable of informing users (the community) the rate of increase in the confirmed COVID-19 cases on both the global and local scales. For the global scale, it basically informs users on the general rate of increase in the number of cases across the different continents or regions, while the local scale analysis informs users of the exact class
of the rate of increase in the local cases for specific countries of their choice. The team believes that the latter application provides the most useful insights at this stage.

### 3.6. Classification performance evaluation

A total of 3 performance indicators are used to evaluate the performance of our proposed encoder-decoder model, namely: (a) precision; (b) recall; and (c) F1-score. Precision, $P$ (see Equation 3) is defined as the number of true positives (TPs) divided by the number of true positives plus the number of false positives (FPs). Recall, $R$ (see Equation 4) is defined as the number of TPs divided by the number of TPs plus the number of false negatives (FNs). F1-score, $F1$ combines precision and recall is the harmonic mean of precision and recall and is defined in Equation 5. In our analysis, TP indicates that a specific class (0 to 5) for the computed $G_e$ parameter pertaining to the different grids/pixels is correctly predicted by the model, which is the most desirable outcome for multi-classification analysis.

$$P = \frac{TPs}{TPs + FPs} \quad (3)$$

$$R = \frac{TPs}{TPs + FNs} \quad (4)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (5)$$

where $P$ measures the average precision magnitude within a time interval; $R$ measures the completeness of the classification performance within a time interval; $F1$ measures the weighted average of the precision and recall, which reaches its best value at 1 and worst score at 0.

**Table 5. Details of proposed encoder-decoder architecture.**

| Layer No. | Input | Filter size applied | No. of filters | 2D Pooling layer | Pooling filter size | Activation function |
|-----------|-------|---------------------|----------------|------------------|---------------------|---------------------|
| Convolutional layer 1 + Batch Normalization | 3D arrays (i.e., 2D satellite images) | (3x3) | 128 | Nil | | ReLU |
| Pooling layer 1 | Convolutional layer 1 | Nil | | Maximum (2x2x2) | Nil |
| Convolutional layer 2 + Batch Normalization | Pooling layer 1 | (3x3) | 256 | Nil | | ReLU |
| Pooling layer 2 | Convolutional layer 2 | Nil | | Maximum (2x2x2) | Nil |
| Layer Type | Previous Layer | Convolutional Size | Output Channels | Activation | Next Layer |
|------------|----------------|--------------------|-----------------|------------|------------|
| Conv. layer 3 + Batch Normalization | Pooling layer 2 | (3×3) | 512 | ReLU | Up-pooling layer 1 |
| Up-pooling layer 1 | Concatenate the transpose of convolutional layer 3, and convolutional layer 2 | (3×3) for transpose of convolutional layer 3 | 256 for transpose of convolutional layer 3 | ReLU | De-convolutional layer 4 |
| De-convolutional layer 4 | Up-pooling layer 1 | (3×3) | 256 | ReLU | De-convolutional layer 5 |
| Up-pooling layer 2 | Concatenate the transpose of convolutional layer 2, and convolutional layer 1 | (3×3) for transpose of convolutional layer 3 | 128 for transpose of convolutional layer 3 | ReLU | De-convolutional layer 6 |
| De-convolutional layer 5 | | (3×3) | 128 | | De-convolutional layer 6 |
| De-convolutional layer 6 | De-convolutional layer 5 | (3×3) | 6 (no. of classes) | Softmax | |

"same" padding adopted for architecture → spatial dimensions of inputs and outputs are identical after each operation.

**Table 6.** Overview of pipeline parallelism with 2 GPU cards.

| GPU-0 | FP-Batch 1 | FP-Batch 2 | Idle | BP-Batch 1 | FP-Batch 3 | BP-Batch 2 | FP-Batch 4 | BP-Batch 3 |
|-------|------------|------------|------|------------|------------|------------|------------|------------|
| GPU-1 | Idle       | FP-Batch 1 | BP-Batch 2 | FP-Batch 2 | BP-Batch 2 | FP-Batch 3 | BP-Batch 3 | FP-Batch 4 |

| Master Thread | FP-Batch 1 | FP-Batch 2 | Idle | BP-Batch 1 | FP-Batch 3 | BP-Batch 2 | FP-Batch 4 | BP-Batch 3 |
|---------------|------------|------------|------|------------|------------|------------|------------|------------|
| Store random initial weights of encoder model and outputs | Retrieve initial weights data and outputs associating to encoder model for use in BP-Batch 1 and | No operations | Erase initial weights data and outputs associating to encoder model from | Retrieve initial weights data and outputs associating to encoder model for use in BP-Batch 2 and | Erase initial weights data and outputs associating to encoder model from | Retrieve initial weights data and outputs associating to encoder model for BP-Batch 3 and FP- |
from GPU-0 for FP-Batch 1 from GPU-0 for FP-Batch 2 respectively Batch 1 runs FP-Batch 3 respectively Batch 2 runs Batch 4 respectively

Store initial weights of encoder model and outputs from GPU-0 for FP-Batch 3 respectively Batch 2 runs Store initial weights of encoder model and outputs from GPU-0 for FP-Batch 4 respectively

| Without pipeline parallelism |
|-----------------------------|
| GPU-0                      |
| FP- Batch 1 | Idle | Idle | BP-Batch 1 | FP-Batch 1 | Idle | Idle | BP-Batch 2 |
| GPU-1                      |
| Idle | FP- Batch 1 | BP-Batch 1 | Idle | Idle | FP-Batch 2 | BP-Batch 2 | Idle |

* FP represents forward-pass (e.g., FP-Batch 1 means forward pass for mini-batch 1).

** BP represents back-propagation (e.g., BP-Batch 1 means backpropagation for mini-batch).

4. Results and discussions

Adhering to our proposed 2-step optimization workflow (see Figure 4 and Algorithm 1) for the data fusion process, this section systematically describes the results obtained from Steps I and II of the proposed workflow. Again, we highlight that the best possible set of hyperparameters, as determined from Step I, will be leveraged as the 1st set of primary hyperparameters for training (and validating) our proposed prediction model in Step II which involves the data fusion between the optimal climate feature and all 11 socioeconomic-governmental factors, followed by additional model tuning to further maximize the model’s predictive capability if possible. Finally, we provide some high-level details on the application of the proposed prediction model to possibly control the spread of COVID-19 in the local communities in near real-time.

4.1. Step I: Climate feature selection

4.1.1. Model evaluation on a global scale

For each of the main Scenarios as summarized in Table 4, we first evaluate the model’s predictive performance on a global scale by computing the average precision, recall, and F1-score values derived from the model’s testing step, with respect to Equations (3 – 5). The dataset is first split into a 70:30 ratio, where the first 70% of the data pool, starting from 22 Jan 2020, is used for model training and validation. The remaining 30% is then used for model testing. Within the former 70% of data, we further split the dataset internally into 80:20 for training and validating the proposed
Tables 7 to 11 summarize the average precision, recall, and F1-score values computed from the model’s testing step by varying the selected batch sizes (4, 8, 16, and 32 in quantities) for model training and validation in each main Scenario and their corresponding Sub-Scenario as listed in Table 4. The obtained results indicate that the model’s precision score from the testing step can be ranked accordingly with respect to the multiple climate features being modelled: LSTD > CPR > COT > LSTN > CF. Readers are also referred to Figure 7 which illustrates the training loss and validation accuracy plots for the best Sub-Scenario determined for each of the climate features using the corresponding hyperparameters as shown. The current maximum precision score, as part of Step I optimization objective, attained from the model’s testing step is approximately 72.1% by using the average 2D satellite image for the LSTD parameter, as derived from the previous leading 7 days, for the modelling step (see Table 10, sub-Scenario 4f) which highlights the usefulness of reducing the “noises” in satellite images via the prior averaging step. In addition, the results align with that of recent studies (Wu et al., 2020 and Guo et al., 2021) which concluded that temperature does have some correlation with the transmission rate of COVID-19. In our analysis, however, we focused on land temperature instead of ambient of which both parameters are expected to be correlated to an extent.

Comparing LSTD and LSTN parameters, the former has a stronger relative correlation with the rate of increase in COVID-19 cases worldwide by an approximate 10% difference between their respective best precision scores (see Tables 10 and 11, Scenarios 4f) from the model’s testing step, which could possibly be attributed to the following reasons. Firstly, there are generally fewer interactions among individuals during the nights; hence lowering the probability of infection spread via human contact. On the contrary, human activity, even during lockdown periods in different countries, tend to be more significant during the day. The observed LSTD may thus coincide with the transmission rate of COVID-19 cases as it is likely that the bulk number of infections occurs during the day. At the same time, the land temperature during the day tends to be higher which may suppress the spread of COVID-19 cases worldwide, as reported in previous studies (Wu et al., 2020 and Guo et al., 2021), hence resulting in more variations in the computed $G_t$ values over space and time. Consequently, the greater variations in the measured $G_t$ values may better match with the classes’ predictions when using LSTD as the input feature. This pointer will be further verified in the later sections where we quantitatively investigate the relationship between the LSTD parameter and the $G_t$ classes’ variations for specific major countries at the local scale analysis.

When using CPR, COT, and CF as input features, the resulting model performance is less ideal as compared to that of LSTD. The results make intuitive sense since CPR, COT, and CF are expected to affect the amount of sunlight being absorbed by the ground surface (see Table 1), hence there is a degree of relationship between CPR/COT/CF and LSTD/LSTN, where the temperature is the dependent feature. This observation aligns with our previous statement made that there may be overlapping relationship(s) among the different climate features, hence underlining the need to first identify the most suitable, i.e., optimal, climate feature which can best maximize the prediction model’s precision score on the defined testing dataset as part of Step I in our proposed optimization workflow for the data fusion process. On a whole, the better predictive performance derived from using LSTD confirms the following: (i) temperature (land and likely ambient) is complexly affected by the other climate features (CPR, COT, and CF), hence the data parameter is likely to be most
useful to model the rate of increase in the number of COVID-19 cases across the 251 countries; and (ii) temperature can therefore be taken as the key climate feature which can significantly affect the rate of increase of COVID-19 cases globally, thus reducing the dimensions of the current problem being modelled in Step II of the optimization workflow.

**Figure 6.** Training loss and validation accuracy plots for the best Sub-Scenarios determined for each of the climate feature modelled: (a) Sub-Scenario 1a - CF; (b) Sub-Scenario 2a - COT; (c) Sub-Scenario 3f - CPR; (d) Sub-Scenario 4f - LSTD; (e) Sub-Scenario 5f - LSTN.

**Table 7.** Summary of $P$, $R$ and $F1$ values for Scenario 1a – 1f (Cloud Fraction, CF).

| Sub-Scenario | Batch size | Epochs | Precision | Recall | F1-Score |
|--------------|------------|--------|-----------|--------|----------|
| 1a (fixed 3 days’ lead-time) | 8 | 25 | 0.615 | 0.759 | 0.679 |
| | 16 | | 0.607 | 0.761 | 0.675 |
| | 32 | | 0.588 | 0.761 | 0.663 |
| 1b (average 3 days’ lead-time) | 8 | 25 | 0.604 | 0.723 | 0.658 |
| | 16 | | 0.613 | 0.751 | 0.675 |
| | 32 | | 0.598 | 0.759 | 0.669 |
| 1c (fixed 5 days’ lead-time) | 8 | 25 | 0.590 | 0.744 | 0.658 |
| | 16 | | 0.589 | 0.746 | 0.658 |
| | 32 | | 0.598 | 0.760 | 0.663 |
| 1d (average 5 days’ lead-time) | 8 | 25 | 0.604 | 0.703 | 0.650 |
| | 16 | | 0.593 | 0.736 | 0.657 |
| | 32 | | 0.588 | 0.750 | 0.659 |
| 1e (fixed 7 days’ lead-time) | 8 | 25 | 0.591 | 0.744 | 0.659 |
| | 16 | | 0.589 | 0.753 | 0.661 |
| | 32 | | 0.596 | 0.761 | 0.668 |
| 1f (average 7 days’ lead-time) | 8 | 25 | 0.601 | 0.733 | 0.660 |
| | 16 | | 0.595 | 0.725 | 0.654 |
| | 32 | | 0.602 | 0.756 | 0.670 |
Table 8. Summary of $P$, $R$ and $F1$ values for Scenario 2a – 2f (Cloud Optical Thickness, COT).

| Sub-Scenario | Batch size | Epochs | Precision | Recall | F1-Score |
|--------------|------------|--------|-----------|--------|----------|
| 2a (fixed 3 days’ lead-time) | 4          | 25     | 0.588     | 0.741  | 0.656    |
|          | 8          |        | 0.584     | 0.750  | 0.657    |
|          | 16         |        | 0.589     | 0.755  | 0.662    |
|          | 32         |        | 0.592     | 0.731  | 0.654    |
| 2b (average 3 days’ lead-time) | 4          | 25     | 0.661     | 0.700  | 0.680    |
|          | 8          |        | 0.598     | 0.753  | 0.667    |
|          | 16         |        | 0.623     | 0.715  | 0.666    |
|          | 32         |        | 0.606     | 0.691  | 0.646    |
| 2c (fixed 5 days’ lead-time) | 4          | 25     | 0.587     | 0.746  | 0.657    |
|          | 8          |        | 0.587     | 0.742  | 0.655    |
|          | 16         |        | 0.610     | 0.718  | 0.660    |
|          | 32         |        | 0.585     | 0.692  | 0.634    |
| 2d (average 5 days’ lead-time) | 4          | 25     | 0.604     | 0.695  | 0.646    |
|          | 8          |        | 0.614     | 0.691  | 0.650    |
|          | 16         |        | 0.631     | 0.725  | 0.675    |
|          | 32         |        | 0.605     | 0.675  | 0.638    |
| 2e (fixed 7 days’ lead-time) | 4          | 25     | 0.587     | 0.749  | 0.658    |
|          | 8          |        | 0.591     | 0.735  | 0.655    |
|          | 16         |        | 0.585     | 0.739  | 0.653    |
|          | 32         |        | 0.597     | 0.742  | 0.662    |
| 2f (average 7 days’ lead-time) | 4          | 25     | 0.659     | 0.716  | 0.686    |
|          | 8          |        | 0.632     | 0.692  | 0.661    |
|          | 16         |        | 0.614     | 0.717  | 0.662    |
|          | 32         |        | 0.599     | 0.678  | 0.636    |

Table 9. Summary of $P$, $R$ and $F1$ values for Scenario 3a – 3f (Cloud Particle Radius, CPR).

| Sub-Scenario | Batch size | Epochs | Precision | Recall | F1-Score |
|--------------|------------|--------|-----------|--------|----------|
| 3a (fixed 3 days’ lead-time) | 4          | 25     | 0.592     | 0.739  | 0.657    |
|          | 8          |        | 0.590     | 0.760  | 0.664    |
|          | 16         |        | 0.602     | 0.736  | 0.662    |
Table 10. Summary of P, R and F1 values for Scenario 4a – 4f (Land Surface Temperature Day, LSTD).

| Sub-Scenario                  | Batch size | Epochs | Precision | Recall | F1-Score |
|-------------------------------|------------|--------|-----------|--------|----------|
| 3b (average 3 days’ lead-time)| 4          | 25     | 0.590     | 0.744  | 0.658    |
|                               | 8          |        | 0.594     | 0.721  | 0.651    |
|                               | 16         |        | 0.639     | 0.735  | 0.684    |
|                               | 32         |        | 0.609     | 0.644  | 0.626    |
| 3c (fixed 5 days’ lead-time)  | 4          | 25     | 0.594     | 0.730  | 0.655    |
|                               | 8          |        | 0.584     | 0.754  | 0.658    |
|                               | 16         |        | 0.652     | 0.725  | 0.687    |
|                               | 32         |        | 0.595     | 0.754  | 0.665    |
| 3d (average 5 days’ lead-time)| 4          | 25     | 0.617     | 0.674  | 0.644    |
|                               | 8          |        | 0.608     | 0.737  | 0.666    |
|                               | 16         |        | 0.597     | 0.733  | 0.658    |
|                               | 32         |        | 0.603     | 0.668  | 0.634    |
| 3e (fixed 7 days’ lead-time)  | 4          | 25     | 0.595     | 0.717  | 0.650    |
|                               | 8          |        | 0.586     | 0.748  | 0.657    |
|                               | 16         |        | 0.626     | 0.746  | 0.681    |
|                               | 32         |        | 0.587     | 0.739  | 0.654    |
| 3f (average 7 days’ lead-time)| 4          | 25     | 0.641     | 0.699  | 0.669    |
|                               | 8          |        | 0.658     | 0.711  | 0.683    |
|                               | 16         |        | 0.678     | 0.734  | 0.705    |
|                               | 32         |        | 0.600     | 0.736  | 0.661    |
| Sub-Scenario | Batch size | Epochs | Precision | Recall | F1-Score |
|--------------|------------|--------|-----------|--------|---------|
| 4c (fixed 5 days’ lead-time) | 4 | 25 | 0.615 | 0.746 | 0.674 |
| | 8 | | 0.593 | 0.758 | 0.665 |
| | 16 | | 0.594 | 0.762 | 0.668 |
| | 32 | | 0.591 | 0.761 | 0.665 |
| 4d (average 5 days’ lead-time) | 4 | 25 | 0.643 | 0.707 | 0.673 |
| | 8 | | 0.621 | 0.754 | 0.681 |
| | 16 | | 0.606 | 0.755 | 0.672 |
| | 32 | | 0.596 | 0.760 | 0.668 |
| 4e (fixed 7 days’ lead-time) | 4 | 25 | 0.600 | 0.753 | 0.668 |
| | 8 | | 0.591 | 0.760 | 0.665 |
| | 16 | | 0.657 | 0.762 | 0.706 |
| | 32 | | 0.593 | 0.758 | 0.665 |
| 4f (average 7 days’ lead-time) | 4 | 25 | 0.638 | 0.735 | 0.683 |
| | 8 | | **0.721** | **0.744** | **0.732** |
| | 16 | | 0.603 | 0.756 | 0.671 |
| | 32 | | 0.610 | 0.758 | 0.676 |

Table 11. Summary of \( P \), \( R \) and \( F1 \) values for Scenario 5a – 5f (Land Surface Temperature Night, LSTN).
In summary, the optimal configuration from the completion of Step I for the optimization workflow is defined as follows: (i) Land Surface Temperature Day, LSTD climate feature; (ii) batch size of 8; and (iii) average 2D image representation using 7 days of lead-time from Scenario 4f. In the later Step II, the optimal feature in LSTD is fused with the other 11 socioeconomic-governmental factors to construct a new 3D array of 180 x 360 x 12 in size to undergo model training (and validation) with the same model configurations from Scenario 4f, followed by testing the trained model on the same testing dataset.

4.1.2. Correlation analysis for climate features

We further investigate the possible correlations between the defined input feature (climate feature) and target output (spatial classes distributions for $G_t$) at multiple lead-times of 3 days, 5 days, and 7 days, via the traditional cosine similarity (CS) metric as defined in Equation (6). This metric enables us to evaluate the level of similarity in the temporal trend of the defined input feature and target output a priori. While we do not expect to observe very high similarity between 2 disparate quantities (climate conditions and transmission rate of COVID-19), the CS computations for each climate feature may provide a preliminary indication of their respective level of statistical relationship with the spread of COVID-19 globally. In addition, the analysis may be helpful to complement with our finding, from the preceding sub-section, on which specific climate feature can best model the proposed G parameter, with respect to the testing dataset, for the optimal combination of the hyperparameters and Sub-Scenario for the defined lead-time.

For 2 defined images of size of $x \times y$, they can first be flattened into respective 1D arrays. In our analysis, the resulting size of the 1D arrays for each singular climate feature is 64800 by flattening the corresponding 2D maps for the input data feature and target output a priori. The CS index for each pair of 1D arrays $(X, Y)$ can be quantified as follows:

|   |   |   |   |   |
|---|---|---|---|---|
| 32 | 0.598 | 0.762 | 0.670 |
| 5d | 4  | 25 | 0.597 | 0.745 | 0.663 |
|     | 8  | 0.599 | 0.757 | 0.669 |
|     | 16 | 0.604 | 0.759 | 0.673 |
|     | 32 | 0.588 | 0.754 | 0.661 |
| 5e | 4  | 25 | 0.617 | 0.757 | 0.680 |
| (fixed 7 days’ lead-time) | 8  | 0.599 | 0.759 | 0.670 |
|     | 16 | 0.600 | 0.762 | 0.671 |
|     | 32 | 0.581 | 0.762 | 0.659 |
| 5f | 4  | 25 | 0.636 | 0.733 | 0.681 |
| (average 7 days’ lead-time) | 8  | 0.600 | 0.760 | 0.671 |
|     | 16 | 0.610 | 0.761 | 0.677 |
|     | 32 | 0.604 | 0.760 | 0.673 |
\[ CS = \|X\| \|Y\| \cos \theta \] (6)

where \(\|X\|\) is the norm value of \(X\), \(\|Y\|\) is the norm value of \(Y\), and \(\theta\) the cosine distance between \(X\) and \(Y\). As a rule of thumb, an increasing similarity between \(X\) and \(Y\) is achieved with decreasing cosine distance between them.

**Figure 7.** Temporal variations of Similarity Indexes for different climate features at multiple lead-times of 3, 5 and 7 days for the period between late Jan 2020 and Sep 2020 (200 days): (a) LSTD; (b) LSTN; (c) CF; (d) COT; (e) CPR.

For each day with respect to the defined lead-time (3, 5 or 7 days), we compute the CS values for every pair of the input feature and target output, both in 2D map representations, for all climate features being modelled. Figures 6a to 6e illustrate the temporal variations of the similarity index values for each climate feature respectively at the different lead-times of 3, 5, and 7 days during the period between late Jan 2020 and Sep 2020 (200 days). Inspections of the computed CS values for all cases indicate the following key observations:

- For all climate features, the estimated p-values are greater than 0.05 (our assumed significance value) which indicates statistical significance between the climate features and the global G parameter. In short, there is at least 95% probability chance that climate conditions do affect the rate of increase in COVID-19 cases globally. To what extent, however, is thus best investigated via data-driven deep learning analysis where we have shown previously that LSTD achieves the highest precision score from the model’s testing test as summarized in Table 10.

- Land temperatures (day and night) generally exhibit a moderate negative correlation (average similarity index value of around -0.500 to -0.600) with the transmission rate of COVID-19 over time, hence verifying previous studies that higher temperature is likely to suppress the spread of COVID-19 and vice versa. This observation thus complements our earlier finding that LSTD can best correlate with the G parameter on the global scale, where it achieves the highest possible precision score of 72% among all other climate features being selected as the data input feature.

- Cloud fraction exhibits the weakest negative correlation with the transmission rate of COVID-19 over time, while cloud optical thickness and cloud particle radius have the strongest negative correlation (average similarity index of around -0.600) with COVID-19 during the approximate period of April 2020 to May 2020. A similar increasing trend in the correlation values for the other 3 climate features can be observed for the same period. The observation is likely to make intuitive sense as the global community was still developing a scientific understanding of the unknown virus during that period and evaluating different social measures to control the virus spread, hence climate conditions were likely to be most influential in affecting the transmission rate of COVID-19 then.
Interestingly, all climate features exhibit very low correlations (similarity indexes ≤ 0.1) with the spread of COVID-19 during the approximate period of June 2020 to July 2020 which may be attributed to the fact that most countries had implemented their respective lockdown periods then to flatten the COVID-19 transmission rate curve. As such, it was probable that climate conditions had a minimum influence on the spread of COVID-19 during prolonged lockdown periods as social contact was kept at its minimum. This observation is thus useful to highlight the need to include other relevant socioeconomic factors and/or restrictive governmental policies in the modelling analysis for predicting/classifying the classes distribution in the growth rate for the number of confirmed COVID-19 cases over space and time, which is the focus of Step II in our proposed optimization workflow for the data fusion process.

Adoption of the different lead-times (3 to 7 days), i.e., using different previous days of satellite images for the climate features, does not seem to significantly influence the transmission rate of COVID-19. This observation again makes intuitive sense as we would expect climate conditions to follow quasi steady-state variations over time, unless climate anomaly (unexpected variations) took place.

To cross validate our overall finding that LSTD, or temperature as the key climate feature, we cross-referenced with previous studies, as summarized in the below pointers, to evaluate if similar findings can be made at this stage.

- Wu et al. (2020) reported that temperature and humidity can generally suppress the spread of the COVID-19 virus, where a 1°C increase resulted in 3.08% reduction in the daily number of new cases reported globally by analysing 161 countries. Their key finding supports our correlation analysis where our present results report that there is negative correlation between LSTD and the global G parameter with p-value greater than the significance value of 5%.

- Guo et al. (2021) further affirms that temperature is most important in affecting the incidence rate of COVID-19 where higher temperatures can reduce the virus spread globally by analysing 190 countries. Likewise, their overall finding aligns with our present results attained for LSTD feature.

- Our present prediction model is built upon the analysis for 251 countries, hence the scale is larger than that of Wu et al. (2020) and Guo et al. (2021), while considering the differing methodologies involved among the studies. Despite the variations, we have shown consistent finding, in terms of underlining the negative correlation between temperature and the transmission rate of COVID-19, with the above 2 key studies which are among the few that
focus on developing a model which considers the global context, instead of individual
countries/states.

- However, in those previous studies, there is no extension of the respective prediction model
to the local scale analysis, i.e., direct prediction of specific countries, based upon the
modeler’s choice, from the macroscale global model within the end-to-end prediction
framework as proposed in our present study. Hence, we are hopeful of the additional novelty
introduced by our proposed approach, coupled with the 2-step optimization process for the
data fusion step, in modelling the global and local G parameters.

4.1.3. Model evaluation at a local scale

Building upon Scenario 4f (batch-size of 8 for LSTD parameter) from Table 10, which provides
the highest precision score of 72% at this stage, we perform a local analysis for several major
countries from the different continents, namely: (a) Indonesia (Asia); (b) Brazil (South America); (c)
Germany (Europe); (d) Somalia (Africa); and (e) Australia (Australian Continent), by examining the
level of agreement between the predicted and actual classes for their respective $G_{t,local}$ classes over
time. Likewise, a lead-time of 7 days is adhered to perform the model predictions, as starting from 1
Feb 2020, using the same encoder-decoder model from Figure 5.

**Figure 8.** Comparison of observed and predicted $G_t$ classes for 200 days, as starting from 1 Feb
2020: (a) Indonesia (Asia); (b) Brazil (South America); (c) Germany (Europe); (d) Somalia (Africa);
(e) Australia (Australian Continent).

**Table 12.** Summary of accuracy scores for selected countries for local analysis.

| Countries               | Accuracy score |
|-------------------------|----------------|
| Indonesia (Asia)        | 75.5%          |
| Brazil (South America)  | 68.0%          |
| Germany (Europe)        | 63.5%          |
| Somalia (Africa)        | 81.0%          |
| Australia (Australian Continent) | 79.0% |

Figures 9a to 9e illustrate the classes comparison for Indonesia, Brazil, Germany, Somalia, and
Australia respectively, while Table 12 summarizes the exact accuracy scores computed from the
combined model training and testing steps. By inspecting the derived results for the 5 selected
countries from the different continents, the following key observations/pointers can be made:

- During the expected warmer period (approximately Day 90 to Day 180), we observe a
  significant number of Class 0 ($G_t \leq 1\%$) for the predictions and actual observations in
  Indonesia, Germany, and Somalia. This result aligns with most recent studies (Wu et al., 2020
  and Guo et al., 2021) which too concluded that higher temperatures can suppress the
  transmission rates of COVID-19.
• On the contrary, in the warmer period of Brazil (approximately Day 90 to Day 180), the rate of increase in the number of confirmed COVID-19 cases (Classes 1-2) was generally higher than that of Indonesia, Germany, and Somalia (Class 0), which could be attributed to the lack of social control measures enforced by the local government to control the virus spread. The observation hence suggests that there is an intricate relationship between land surface day temperature and the types of social control measures implemented by governments to effectively stem the virus spread in their local communities. At the same time, this observation further highlights the need to include other relevant socioeconomic-governmental factors in the modelling analysis for predicting/classifying the classes distribution in the growth rate for the number of confirmed COVID-19 cases over space and time. Intuitively, due to additional dimensions provided by fusing the socioeconomic-governmental factors with the optimal climate feature in LSTD, we would expect that our proposed prediction model can better forecast, with higher resulting accuracy, the specific classes for the local G parameter over time for all 5 selected major countries. This will be further investigated in Step II of the optimization workflow, as summarized in the next section.

• During the winter period in Australia for Day 100 to Day 160 approximately, there is a rise in the number of confirmed COVID-19 cases where Classes 1 and 2 are observed. This suggests that colder temperatures may accelerate the virus spread, hence underlining the importance of more stringent social measures to be implemented to control social contact during colder periods.

• Similarly, a significant rate of increase in the number of confirmed COVID-19 cases (Class 2 and above) can be observed for Day 180 and beyond in the case of Germany which could be ascribed to the increasingly colder conditions, coupled with the lack of social measures enforced.

• Our proposed prediction model for forecasting the $G_t$ classes across the different continents results in an average accuracy score of around 65% to 70% conservatively, hence confirming a level of relationship between land surface day temperature and transmission rate of COVID-19. We are thus hopeful that this information can be helpful to the global community to develop the most appropriate environmental-social policies, with respect to land surface day temperature, to control the virus spread globally in the foreseeable future, while vaccinations gradually drive up in the different countries.

4.2. Step II: Data fusion with optimal climate feature
We now proceed to fuse the selected climate feature in LSTD with the 11 listed socioeconomic-governmental factors from Table 3. Note that the new 3D array from the data fusion process has a resulting array size of (180 x 360 x 12) in size as previously shown in Figure 3. Likewise, we leveraged on the same model configuration of batch sizes ranging among 8-32, epochs of 25 and lead-time of 7 days for constructing the average 2D map representation for the LSTD parameter to train and validate the same deep learning prediction model, following testing the trained model on the same defined testing dataset from Step (I). We also note that the same 7 days lead-time is used to construct the average 2D map representation for all 11 socioeconomic-governmental factors, as part of Sub-Scenario (f) from Table 4. All other model configurations such as the size of training (validation) and testing datasets remain unchanged from that of Step I in the optimization process. In addition, we also increase the threshold precision score to a minimum value of 75% from Figure 4 to evaluate if the optimized set of hyperparameters, as previously determined from Step I, can be maintained for the same model training step in Step (II). Figure 9 illustrates the new training loss and validation accuracy profiles derived for the batch size variations among 4, 8, 16, 32 as shown. The figure generally shows that the similar validation accuracy scores can be achieved with the fused features at the varying batch sizes, when compared to Figure 6d, hence indicating that the model evaluation on the testing dataset will be more useful to determine the overall usefulness of the data fusion process.

Figure 9. New training losses and validation accuracies obtained from model training used fused features between LSTD and 11 socioeconomic-governmental factors with varying batch sizes and epochs = 25.

By increasing the total input feature dimensions by 12 folds/times in Step II of our proposed optimization process, we would expect an overall improvement in the trained model’s predictive performance on the same testing dataset as summarized in Table 13, when compared to Sub-Scenario 4f in Table 10. At this stage, the highest average precision score attained from the same batch size of 8 is 77.1%, hence fulfilling our current threshold precision score of 75%. As such, the improvement in the average precision score on the global scale indicates that the classes predictions for all 251 countries are expected to better align with their actual classes for the analyzed period. Using the latest trained prediction model, as derived from the optimal model configuration of batch size of 8 and lead-time of 7 days in Table 13, coupled with the fused data features, we then proceed to predict the respective $G_{t,\text{local}}$ classes for the same 5 major countries in Australia, Brazil, Germany, Indonesia, and Somalia as examined previously. Table 14 summarizes the local accuracy scores obtained from the trained prediction model where we can generally observe an average improvement of 5% from the previous scores in Table 12, hence highlighting the overall benefit provided by the additional dimensions derived from fusing the 11 socioeconomic-governmental factors with LSTD. Using the latest results obtained, the following new observations/pointers can be made.

- To address our previous pointer made for the classes predictions of Brazil, the inclusion of the relevant socioeconomic-governmental factors may now better explain the observed higher rate of increase in the number of confirmed COVID-19 cases during their warmer period (approximately Day 90 to Day 180). Specifically, we note that during this period, Brazil’s “Government stringency index” parameter (see Table 3) ranged between 75.0 and
81.0 which can be reasonably strict in terms of the level of government’s control measures. However, there were no strict stay-home restrictions imposed in Brazil in the same period, where the respective classes for the “Stay-at-home restrictions” parameter (see Table 3) ranged between Classes 1 and 2. Hence, the general populations were still able to travel out of home with some level of freedom which may have thus resulted in continual human contact. The gradual reduction in the rate of increase in the number of confirmed COVID-19 cases in Brazil from Day 100 onwards, as illustrated in Figure 8b, is likely ascribed to the case where Class 3 for the “Stay-at-home restrictions” parameter was maintained consistently for the next 6 months. At the same time, we also note that other social measures such as “Face covering”, “Cancellation of public events and gatherings”, “Schools and workplaces closures”, “Controls for domestic travels”, and “Controls for international travels” were maintained at their respective highest or 2nd highest classes during the same period which may have thus controlled the virus spread.

- The example of Germany in Figures 8c also highlight the usefulness of including the relevant socioeconomic-governmental factors for controlling their rate of increase in the number of COVID-19 cases at as early as from Day 25 onwards which was around mid-Feb to Mar 2020. For example, schools were recommended (Class 2 for “Schools and workspace closure” parameter from Table 3) to be closed from 25 Feb 2020 onwards, hence limiting the interactions among youths which may have been useful to control the virus spread in Germany since Day 25. In addition, we do note that the different control measures (“Stay home restrictions”, “Schools and workplaces closures”, etc.) began to be less stringent (lower classifications from Table 3) from Day 160 onwards in Germany which was close to their usual autumn/winter period. Hence, the colder temperatures observed, at the same time, may have accelerated the virus spread which aligns with our previous pointer made for Australia (Figure 8e) where the nation experienced their winter period between Day 100 and Day 160 approximately. In summary, the results again underline that higher temperatures have a greater likelihood of suppressing the virus spread, for a given set of socioeconomic-governmental factors, as compared to lower temperatures which, in turn, may accelerate the spread of COVID-19.

- While we note that the data fusion step between the optimal LSTD parameter and the 11 socioeconomic-governmental factors does improve the proposed prediction model’s accuracy in predicting the global and local classes for the defined G parameter, it is obvious that the current highest precision and accuracy scores have not exceeded the 90% score. The
reason is likely ascribed to the lack of data features which can effectively account for any “wrongdoings” by the individuals in the different local communities which may have led to unexpected increase in G parameter in the respective countries. The current prediction model cannot predict such dynamic behaviour which, however, may require finer scales of analysis within the community to do so. Examples of the defined “wrongdoings” include, but not limited to, breaking of home-restrictions, and social-distancing rules. The sudden increase in the growth rate in the number of confirmed COVID-19 cases can generally be observed for all 5 countries (Figures 8a – 8e), where there were unexpected occurrences of Class 5 at certain days or periods during the 200 days being analysed as shown in the respective figures. The inclusion of the present vaccination rates can also be considered in future modelling studies using the same concept.

Table 13. Summary of $P$, $R$ and $F1$ values from data fusion process between LSTD parameter and socioeconomic-governmental factors in Step II of optimization workflow

| Sub-Scenario | Batch size | Epochs | Precision | Recall | $F1$-Score |
|--------------|------------|--------|-----------|--------|------------|
| 4f (average 7 days’ lead-time) | 4          | 25     | 0.695     | 0.760  | 0.726      |
|              | 8          |        | 0.771     | 0.755  | 0.763      |
|              | 16         |        | 0.645     | 0.787  | 0.709      |
|              | 32         |        | 0.520     | 0.773  | 0.622      |

Table 14. Summary of accuracy scores for selected countries for local analysis from data fusion process between LSTD parameter and socioeconomic-governmental factors in Step II of optimization workflow

| Countries                  | Accuracy score | Improvement |
|----------------------------|----------------|-------------|
| Indonesia (Asia)           | 80.7%          | 5.2%        |
| Brazil (South America)     | 73.0%          | 5.0%        |
| Germany (Europe)           | 68.1%          | 4.5%        |
| Somalia (Africa)           | 86.6%          | 5.6%        |
| Australia (Australian Continent) | 85.1%      | 6.1%        |

4.3. Application of prediction model

In near real-time context, the proposed prediction model, after undergoing the required hyperparameters tuning with the fused data features during their training phase, may assist communities to build towards resilience and sustainable development in a post-COVID era where COVID-19 is expected to become an endemic. Examples of the possible application are as follows:
• Since LSTD feature is expected to affect the virus spread, especially in colder countries as based on the current results obtained, the prediction model can assist healthcare management to better prepare for likely spike(s) in the number of confirmed COVID-19 cases based upon a defined set of socioeconomic-governmental factors as the input fused features into the prediction model with a defined lead-time. In the Western countries, where the winter season usually coincides with the festive period, the increase in social gathering, i.e., lower class of the “Stay-home restrictions” parameter from Table 3, will likely supplement the virus spread as based on the current negative correlation with LSTD. Hence, the prediction model is likely to forecast a larger local G class over time in those countries. By doing so, the local government can be informed beforehand, with a defined lead-time, and undertake the best possible social control measures to manage the amount of social interactions to mitigate, to an extent, the virus spread locally. At the same time, healthcare management can better prepare medical facilities and resources to handle an expected increase of patients due to COVID-19.

• In today’s context, many Asian countries which include Hong Kong, Singapore, Australia, China, and others, are still reluctant to fully open up their economies, hence incurring the risk of being omitted from the gradual global recovery. The prediction model may thus be useful to the above countries to forecast the rate of increase in the local G parameter, given any input combination of the LSTD and other 11 socioeconomic-governmental factors into the trained prediction model. By being able to forecast with a reasonable lead-time, the local government or the different stakeholders can undertake appropriate control measures to mitigate the virus spread, hence alleviating the impact on their economies on the long run. For example, by imposing work-from-home requirements with evidence of a forecasted spike in local community cases, the nation’s economy can still function to some degree, instead of experiencing a more adverse situation where there is a sudden rise in the number of COVID-19 patients which can affect businesses and other sectors of the economy.

• The current model workflow (Figure 4, Algorithm 1) and architecture provides the flexibility to the modeler to continue the data fusion process with other data features which may further improve the present model’s precision and accuracy scores. For example, as discussed previously, the current omission of data features which can quantify “wrongdoings” by individuals in the local community which have been shown to contribute to the virus spread. The inclusion of vaccination rates in the different countries should also be considered in future modelling analysis using the same approach as proposed in the
study. We are hopeful that additional useful data features for training the same prediction model can only further improve its accuracy/precision in due time. Finally, we also note that the same approach can also be extended to forecast the transmission rate of other infectious diseases by varying the lead-times and data features involved for re-training the prediction model.

5. Conclusions and future works

In this paper, we propose an alternative deep learning model, based upon semantic segmentation analysis, to quantitatively correlate climate-socioeconomic-governmental factors to the growth rate in the number of confirmed COVID-19 cases, both on a global and local scales, via a proposed G parameter. To the best of our knowledge, this form of data exploitation for COVID-19 data analytics has not been investigated in the literature by far. The global analysis encompasses a 2D map representation of 251 countries, while enabling a direct local scale analysis to any country of the modeler’s choice. Overall, the model development phase comprises of a 2-step optimization process, where Step I first determine the optimal climate feature which has the strongest correlation with the global and local G parameter. By far, when considering a period of 200 days between late January and October 2020 with respect to different lead-times, the obtained results indicate that land surface day temperature (LSTD) can best correlate with the global G parameter with an average precision score of 72% from the model’s testing step. This observation is further highlighted by our subsequent features analysis where LSTD has been shown to be moderately correlated to the spread of COVID-19, especially during the early stages of 2020 since the virus’ inception, via the cosine similarity computations. At the same time, we also examine the relationship between LSTD and the rate of increase in COVID-19 cases for several major countries, which include Somalia, Australia, Germany, Indonesia, and Brazil, across the different continents. The results obtained from Step I indicated an average accuracy score of 65 to 70% for forecasting the $G_{t, \text{local}}$ classes in those countries with respect to the LSTD climate feature.

Step II of the proposed optimization process then fuses the same LSTD climate feature with 11 selected socioeconomic-governmental factors to further improve the model’s predictive accuracy for both the global and local analyses. By doing so, the average precision score for the global analysis increases to 77% from the model’s testing step, while the local scale analysis for the 5 different countries investigated experienced an average of 5% accuracy improvement in their prediction performance. We then described qualitatively on how the LSTD climate feature correlate with the different socioeconomic-governmental factors in modelling and forecasting the predicted classes for the local G parameter in Germany and Brazil, as examples, with defined lead-times.

Going forward, the team is working on multiple items to build upon our present analysis to continue contributing to the global efforts in combating the present pandemic. Our subsequent studies generally on: (a) further fusing the climate-socioeconomic-governmental features with the community’s emotional responses towards the virus, via social media deep learning analysis, to model the transmission rates of COVID-19 on the global and local country scales; (b) developing an open-source platform tool that can provide near real-time predictions of the different classes of the $G$
parameter at both the global and localized country scales as a function of the fused data features encompassing climate conditions, socioeconomic-governmental factors.

**Declaration of interests**

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

**Acknowledgments**

The authors declare no conflict of interest. This study was supported in part by Microsoft Corporation for the AI for Health COVID-19 Azure Compute Grant of ID:00011000272 and the Start-Up Grant at Nanyang Technological University, Singapore (No. 04INS000423C120).

**References**

Abdollahi, A., and Rahbaralam, M. (2020). “Effect of Temperature on the Transmission of COVID-19: A Machine Learning Case Study in Spain.” medRxiv, 2020.05.01.20087759.

Ahmadi, M., Sharifi, A., Dorosti, S., Jafarzadeh Ghoushchi, S., and Ghanbari, N. (2020). “Investigation of effective climatology parameters on COVID-19 outbreak in Iran.” *Science of The Total Environment*, 729, 138705.

Awasthi, R., Nagori, A., Singh, P., Pal, R., Joshi, V., and Sethi, T. (2020). “Temperature and Humidity Do Not Influence Global COVID-19 Incidence as Inferred from Causal Models.” medRxiv, 2020.06.29.20142307.

Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E., and Grenfell, B. T. (2020). “Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic.” *Science*, 369(6501), 315 LP – 319.

Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., and Bashir, M. (2020). “Correlation between climate indicators and COVID-19 pandemic in New York, USA.” *Science of The Total Environment*, 728, 138835.

Casanova, L. M., Jeon, S., Rutila, W. A., Weber, D. J., and Sobsey, M. D. (2010). “Effects of air temperature and relative humidity on coronavirus survival on surfaces.” *Applied and environmental microbiology*, American Society for Microbiology (ASM), 76(9), 2712–2717.

Chiyomaru, K., and Takemoto, K. (2020). “Global COVID-19 transmission rate is influenced by precipitation seasonality and the speed of climate temperature warming.” medRxiv, 2020.04.10.20060459.

Dalziel, B. D., Kissler, S., Gog, J. R., Viboud, C., Bjørnstad, O. N., Metcalf, C. J. E., and Grenfell, B. T. (2018). “Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities.” *Science*, 362(6410), 75 LP – 79.

Das, A., Ghosh, S., Das, K., Basu, T., Dutta, I., and Das, M. (2021). “Living environment matters: Unravelling the spatial clustering of COVID-19 hotspots in Kolkata megacity, India.” *Sustainable Cities and Society*, 65, 102577.

Fu, X., and Zhai, W. (2021). “Examining the spatial and temporal relationship between social vulnerability and stay-at-home behaviors in New York City during the COVID-19 pandemic.” *Sustainable Cities and Society*, 67, 102757.

Guo, C., Bo, Y., Lin, C., Li, H. B., Zeng, Y., Zhang, Y., Hossain, M. S., Chan, J. W. M., Yeung, D.
W., Kwok, K.-O., Wong, S. Y. S., Lau, A. K. H., and Lao, X. Q. (2021). “Meteorological factors and COVID-19 incidence in 190 countries: An observational study.” *The Science of the total environment*, 757, 143783.

Hu, M., Roberts, J. D., Azevedo, G. P., and Milner, D. (2021). “The role of built and social environmental factors in Covid-19 transmission: A look at America’s capital city.” *Sustainable Cities and Society*, 65, 102580.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020). “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.” *The Lancet*, Elsevier, 395(10223), 497–506.

Ioffe, S., and Szegedy, C. (2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, JMLR.org, 448–456.

Islam, N., Shabnam, S., and Erzurumluoglu, A. M. (2020). “Temperature, humidity, and wind speed are associated with lower Covid-19 incidence.” *medRxiv*, 2020.03.27.20045658.

Kim, D. (2021). “Exploratory study on the spatial relationship between emerging infectious diseases and urban characteristics: Cases from Korea.” *Sustainable Cities and Society*, 66, 102672.

Kroumpouzos, G., Gupta, M., Jafferany, M., Lotti, T., Sadoughifar, R., Sitkowska, Z., and Goldust, M. (2020). “COVID-19: A relationship to climate and environmental conditions?” *Dermatologic therapy*, John Wiley & Sons, Inc., 33(4), e13399–e13399.

Le, N., Le, A., Parikh, J., Brooks, J., Gardellini, T., and Izurieta, R. (2020). “Ecological and Health Infrastructure Factors Affecting the Transmission and Mortality of COVID-19.” 1–14.

LeCun, Y., and Bengio, Y. (1998). “Convolutional Networks for Images, Speech, and Time Series.” *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge, MA, USA, 255–258.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). “Gradient-based learning applied to document recognition.” *Proceedings of the IEEE*, 86(11), 2278–2324.

Li, B., Peng, Y., He, H., Wang, M., and Feng, T. (2021). “Built environment and early infection of COVID-19 in urban districts: A case study of Huangzhou.” *Sustainable Cities and Society*, 66, 102685.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T. Y., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M., and Feng, Z. (2020). “Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia.” *New England Journal of Medicine*, Massachusetts Medical Society, 382(13), 1199–1207.

Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., He, X., Wang, B., Fu, S., Niu, T., Yan, J., Shi, Y., Ren, X., Niu, J., Zhu, W., Li, S., Luo, B., and Zhang, K. (2020). “Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China.” *Science of The Total Environment*, 726, 138513.

Maiti, A., Zhang, Q., Sannigrahi, S., Pramanik, S., Chakraborti, S., Cerda, A., and Pilla, F. (2021). “Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the
contiguous United States.” *Sustainable Cities and Society*, 68, 102784.

Mbbs, L. R., D, L. B. S. M., and Neurosurgery, M. S. (2020). “Coronavirus Disease Coronavirus Disease (COVID-19) Spreads.” *Who*, 75(2), 95–97.

Pan, J., Yao, Y., Liu, Z., Meng, X., Ji, J. S., Qiu, Y., Wang, W., Zhang, L., Wang, W., and Kan, H. (2021). “Warmer weather unlikely to reduce the COVID-19 transmission: An ecological study in 202 locations in 8 countries.” *Science of The Total Environment*, 753, 142272.

Prata, D. N., Rodrigues, W., and Bermejo, P. H. (2020). “Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil.” *Science of The Total Environment*, 729, 138862.

Rodó, X., San-José, A., Kirchgatter, K., and López, L. (2021). “Changing climate and the COVID-19 pandemic: more than just heads or tails.” *Nature Medicine*, Springer US, 27(4), 576–579.

Runkle, J. D., Sugg, M. M., Leeper, R. D., Rao, Y., Matthews, J. L., and Rennie, J. J. (2020). “Short-term effects of specific humidity and temperature on COVID-19 morbidity in select US cities.” *The Science of the total environment*, 740, 140093.

Şahin, M. (2020). “Impact of weather on COVID-19 pandemic in Turkey.” *Science of The Total Environment*, 728, 138810.

Sannigrahi, S., Pilla, F., Basu, B., Basu, A. S., and Molter, A. (2020). “Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach.” *Sustainable Cities and Society*, 62, 102418.

Schmidhuber, J. (2015). “Deep learning in neural networks: An overview.” *Neural Networks*, 61, 85–117.

Shi, P., Dong, Y., Yan, H., Li, X., Zhao, C., Liu, W., He, M., Tang, S., and Xi, S. (2020). “The impact of temperature and absolute humidity on the coronavirus disease 2019 (COVID-19) outbreak - evidence from China.” medRxiv, 2020.03.22.20038919.

Sobral, M. F. F., Duarte, G. B., da Penha Sobral, A. I. G., Marinho, M. L. M., and de Souza Melo, A. (2020). “Association between climate variables and global transmission of SARS-CoV-2.” *The Science of the total environment*, 729, 138997.

Sun, C., and Zhai, Z. (2020). “The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission.” *Sustainable Cities and Society*, 62, 102390.

Tosep, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. A. I., Lestari, H., Bahar, H., and Asfian, P. (2020). “Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia.” *Science of The Total Environment*, 725, 138436.

Ugail, H., Aggarwal, R., Iglesias, A., Howard, N., Campuzano, A., Suárez, P., Maqsood, M., Aadil, F., Mehmood, I., Gleghorn, S., Taif, K., Kadry, S., and Muhammad, K. (2021). “Social distancing enhanced automated optimal design of physical spaces in the wake of the COVID-19 pandemic.” *Sustainable Cities and Society*, 68, 102791.

Viezzer, J., and Biondi, D. (2021). “The influence of urban, socio-economic, and eco-environmental aspects on COVID-19 cases, deaths and mortality: A multi-city case in the Atlantic Forest, Brazil.” *Sustainable Cities and Society*, 69, 102859.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., and Liu, M. (2020). “Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries.” *Science of The Total Environment*, 729, 139051.

Xie, J., and Zhu, Y. (2020). “Association between ambient temperature and COVID-19 infection in 122 cities from China.” *Science of The Total Environment*, 724, 138201.
Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., and Wang, W. (2020). “No Association of COVID-19 transmission with temperature or UV radiation in Chinese cities.” European Respiratory Journal, 2000517.

Yuan, J., Yun, H., Lan, W., Wang, W., Sullivan, S. G., Jia, S., and Bittles, A. H. (2006). “A climatologic investigation of the SARS-CoV outbreak in Beijing, China.” American Journal of Infection Control, Elsevier, 34(4), 234–236.