COVID 19 Pandemic, Socio-Economic Behaviour and Infection Characteristics: An Inter-Country Predictive Study Using Deep Learning

Srinka Basu1 · Sugata Sen2

Accepted: 4 November 2021 / Published online: 26 January 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
This work aims to develop a data driven multi-horizon incidence forecasting model considering the inter-country variability in static socio-economic factors. The specific objectives of this study are to predict the future country-wise COVID 19 incidences, to locate the influences of individual socio-economic factors on the predictions, to analyze the clusters of countries on the basis of influential explanatory variables and thus to search for intra-cluster and inter-cluster characteristics. To that respect this study has used the deep neural network based temporal fusion transformer for the predictions, Pearson correlation to understand the influence of socio-economic variables on incidence and hierarchical clustering for cluster-analysis. The findings conclude that the inter-country infection related predictions vary widely over spatio-temporal variability and different socio-economic variables have different influences over this inter-country variability. It is observed that greater the population size, stronger the global connectedness, larger the social cohesion, higher the population density and meaningful the gender based discrimination higher will be the future spread. On the other hand greater the development level, higher the nutritional status, greater the access to quality health services, greater the urban population and greater the material poverty lesser will be the future spread. Definite spatial pattern of influence of the explanatory variables emerged from cluster analysis. To minimize the vulnerability towards unforeseen biological calamities modern and sustainable development policies are needed; affluence may not guarantee less infection. But these policies should vary between economies due to the variation in socio-economic status of the countries worldwide.

Extended author information available on the last page of the article
Keywords Infection prediction · Static socio-economic variates · Deep neural network

1 Introduction

Prediction about a disaster plays an important role to minimize the damage and economic losses. The importance of prediction increases infinitely when the disaster appears as unique and starts to spread at a fast rate. With the emergence of good quality high volume data, Artificial Intelligence (AI) has appeared as a potent tool of prediction. But in case of unprecedented disasters like COVID 19 minimizing the errors in multi-horizon prediction is one of the greatest challenges of AI based models. These challenges heighten further when the dynamic planning to mitigate the menace largely depends upon the static non-pharmaceutical interventions (NPI). With time this pandemic has exposed that in the absence of medical interventions non-pharmaceutical community measures can be the only policy initiatives which the societies can formulate to overcome the disaster. But the success of these NPI depends upon an array of socio-economic variables. Naturally the errors in predictions vary from counties to countries as the distribution of the influences of the stated variables varies between economies. So this work has tried to develop an AI based multi-horizon predictive model considering the simultaneous variability within the static explanatory variables as well as the longitudinal historical data of infection. Due to exceptional nature of the current disaster selecting the causal variables appeared as the additional challenge of this prediction. To that respect this effort has used an attention based deep neural network termed as Temporal Fusion Transformer (TFT) in selection of the relevance of the explanatory variables. For predictive activities TFT, unlike the existing statistical and machine learning models, combine the static variates, historical cumulative incidence count and derived temporal variates like days since first occurrence of infection to estimate the future temporal target cumulative incidence count. Apart from developing a predictive model this work has delivered country-wise long term future infection levels related to COVID 19 pandemic. The finding concludes that the inter-country infection related predictions vary widely over spatio-temporal variability and different socio-economic variables have different influences over this inter-country variability. Based on the similarity of the influencing explanatory variables the countries of North America, South America, Africa, Asia and Australia exhibit certain patterns. For example, it is observed that none of the North American countries are found in clusters where urban population, poverty, availability and accessibility to quality health care (expressed through Healthcare Availability and Quality Index or HAQI), human development status (expressed through Human Development Index or HDI) and sex ratio have taken most influential role to form the cluster. On the contrary, the South American and Asian countries are not found in the clusters where Trade Openness Index and HAQI have taken the most influential roles respectively while Australian countries are not found in the clusters where HAQI and HDI have played the most influential role to form the clusters.
The first use of computations to establish the incidence of epidemics dates back to 1760 when Bernoulli used it to predict the life expectancy of French population (Marathe and Vullikanti 2013). But the actual improvement in the application of computation based techniques on the studies of epidemics started in the early years of 1900s. In 1911 Ronald Ross developed a model to study the spread of malaria with the help of a system of differential equations (Bacáir 2011a). Kermack and McKendrick furthered this model to a mass-action model (Bacáir 2011b). This development is regarded as the foundation of modern computational epidemiology. The process ultimately culminated in the emergence of Artificial Intelligence (AI) and Machine Learning (ML) as the foremost supports to the professionals and researchers related to development of drugs, improvement of diagnosis and health related policy formulation (Lalmuanawma et al. 2020).

Naturally AI and ML based modeling has started to play a major role in predictive studies on COVID 19 pandemic. These types of modeling have also appeared as the main tool of social policy formulation on the background of COVID 19 pandemic.

Ribeiro et. al has developed an AI based forecasting model on Brazilian data using Stacking-ensemble with support vector regression algorithm to predict short term cumulative incidence of infection (Ribeiro et al. 2020). Chakraborty and Ghosh have proposed a real time prediction model using wavelet-base forecasting and autoregressive integrated moving average time series analysis method. Using demographic data this model has tried to forecast the short term outcome of COVID 19 incidences (Chakraborty and Ghosh 2020). Zheng et. al. have used Improved Susceptible Infected (ISI) model instead of fixed infection rate to study the short term infection process of COVID 19. For developing a hybrid AI based model to incorporate effects of control measures they have embedded Natural Language Processing module and Long Short Term Memory (LSTM) network into the ISI model (Zheng et al. 2020). Yang et.al. have also used LSTM model to predict new infections. Here probability of transmission, incubation rate, probability of recovery or death and number of contacts have been used as explanatory variables (Yang et al. 2020). Li et al. (2020) have tried to develop a predictive model on the transmission process of corona infection using Gaussian distribution. This work has tried to infer the evolution trend of the COVID 19 infection history overtime to predict the future trend. Yan et al. (2020) have tried to develop effective biomarkers to estimate the risk of fatality using machine learning algorithms. To that respect they have used biological, demographic and clinical data of the sample respondents. The study used supervised XGBoost classifier to cluster the respondents with risk of death. Velásquez and Lara (2020) have tried to forecast infection, fatality and recovery rates with the help of Reduced-Space Gaussian Process Regression associated to Chaotic dynamic system using history data. Bertozzi et al. (2020) have tried to estimate the course of infection and the influence of stringency measures on the growth of infection with the help of three interconnected macroscopic models and time series data. Judging the usefulness of different models in different stages this study has interchangeably used exponential growth model, the self-exciting branching process and the...
susceptible-infected-resistant compartment model. They found that reproduction rate is variable over time and location.

Incidentally it is found that most of the AI based predictive models on COVID 19 have failed to incorporate major socio-economic variables or in other words it can be said that predictive models purely based on socio-economic factors are almost absent. It is to be kept in mind that UNDP has mentioned that the COVID 19 pandemic is not a global health emergency but a human development crisis. They have concluded that higher the state of development, higher the level of preparedness to face the pandemic (Kovacevic and Jahic 2020). The above stated preparedness of the countries has differently been studied through an index called HAQI (Health care Access and Quality Index) (Lai et al. 2020). It is not only the human capital but also the social capital which plays an important role to mitigate the spread of infection¹, (Chuang et al. 2015; Koh and Cadigan 2008). Interestingly, apart from the development divergence it is the variation in gender that has played a crucial role in the spread of current pandemic (Conti and Younes 2020). Some other demographic factors which have left a distinct mark in the rate of infection are population density and urbanization (Desai 2020). Recent data from World Health Organization have shown that the initial infection rates and trajectories of infection varied widely for the whole of the world ². In the absence of pharmaceutical interventions it is undoubtedly the effectiveness of the socio-economic variables - which we discussed above, have created some nations more competent and some not.

So at the end of this section it can be said that as the success of control measures depends purely on the socio-economic macro variables, AI based models with socio-economic variables as the explanatory variables are the need of the time. These models will be able to deduce the influence of individual socio-economic variables on the spread of infection at national levels. On the other hand to formulate acceptable social policies AI and ML based models with ability to deliver long-term predictions are also needed. Thus this work wants to substantiate on the following.

- To develop a deep neural network based multi-horizon predictive model for the spread of COVID 19 infection on the basis of static socio-economic macro variables as well as historical incidence record.
- To predict about 270 days ahead inter-country incidence of COVID-19 using the developed model.
- To locate the influence of individual socio-economic factors on the estimated country-wise predictions.
- To cluster the countries on the basis of influential explanatory variables and thus to search for intra-cluster and inter-cluster characteristics.

¹ https://egrove.olemiss.edu/population_brief/2/, https://www.africa.undp.org/content/rba/en/home/blog/2020/covid-19-leveraging-on-social-capital-to-flatten-the-curve-in-e.html
² https://covid19.who.int/
2 Related Work

Time series prediction for the understanding of future trends and tendencies is an important problem in a wide range of applications that have attracted researchers over decades resulting in the development of numerous models. The traditional statistical tools for prediction like ARIMA, ARIMAX and their variants are extensively used for forecasting in various domains including the prediction of the development of infectious diseases, for example: Ahmar and Del Val (2020) used ARIMA and SutteARIMA to predict the short-term of confirmed cases of COVID-19 in Spain, Tandon et al. (2020) and Roy et al. (2021) employed it for forecasting future COVID-19 cases in India, Ding et al. (2020) use ARIMA to predict the spread of the Italian COVID-19 epidemic. In spite of the popularity, the traditional models are limited in their ability to process big data with high dimensionality, to capture the intricate relations between multiple static and temporal variates, to efficiently represent complex non-linear functions.

With greater availability of massive and potentially useful infectious diseases transmission related data through multiple channels, including sentinel reporting systems, national syndromic surveillance systems, internet search key trends, Twitter data, outbreak investigation reports from various agencies, powerful ML tools started being used for effective analysis of epidemic data. Bai and Jin in 2005 (Bai and Jin 2005) used artificial neural network (ANN) with back propagation learning algorithm to predict SARS (severe acute respiratory syndrome) epidemic in Beijing and Shanxi, China. Since then a lot of researchers have explored the potential of ANN and their variants for epidemic prediction (Aburas et al. 2010; Siettos and Russo 2013). Xu et al. (2017) employed a multi-layer feed-forward network to forecast new cases of influenza-like-illness (ILI) in general outpatient clinics in Hong Kong using online social media and web search queries (Google search) data.

The researchers also have proposed hybrid models based on neural networks. Xue et al. (2018) used back propagation neural network and the genetic algorithm to predict ILI using Google Flu Trend data and epidemic data collected from Center for Disease Control (CDC). A hybrid model of statistical model ARIMA and neural network model is used by Chakraborty et al. (2019) to forecast dengue endemic in San Juan, Iquitos, and the Philippines. Ensemble methods based on statistical and ML models that work on temporal epidemic data are also developed in attempt to improve the prediction accuracy (Soliman et al. 2019).

Recently, deep learning has emerged as an effective approach that can alleviate the limitations of the traditional predictive models. A significant number of research papers have come out that use deep neural network model LSTM and its variant to predict incidence count, death count or recovery count of infectious diseases. Chae et al. (2018) used DNN and LSTM models for prediction of infectious diseases namely, Malaria, Chickepox and Scarlet Fever in South Korea based on temporal incidence records and other temporal non-clinical data such as - search query data from the Naver Data Lab, the daily number of tweets mentioning infectious disease, daily average temperature and daily average humidity data. Arora et al. (2020) used
LSTM variants namely, Deep LSTM, Convolutional LSTM and Bi-directional LSTM models for predicting the number of reported positive COVID-19 cases for 32 states and union territories of India. Alazab et al. (2020) used a deep convolutional neural network (CNN) to detect COVID-19 positive patients using chest X-ray images. Additionally, they used the prophet algorithm (PA) (Taylor and Letham 2018) and LSTM to predict the week-ahead numbers of COVID-19 confirmations, recoveries, and deaths. Ayyoubzadeh et al. (2020) used LSTM to predict the incidence of COVID-19 in Iran from the Google Trends website data. Parallelly, a good number of research used deep neural network comprising of several convolutional layers and dense layers for the diagnosis of COVID 19 from clinical data (Panwar et al. 2020; Ning et al. 2020).

Though the deep learning based predictive models helped to achieve higher predictive accuracy over the traditional models, they failed to incorporate the static socio-economic variables while play important roles in the growth of epidemic.

There are only limited work that have tried to quantify the impacts of socio-economic factors on the incidence count of infectious disease. The few existing works have used the state-space models, for example, Redding et al. (2019) used a variant of SEIR model to assess the impacts of socioeconomic factors on epidemic potential of Ebola in Africa, Carter et al. developed a compartmental model with causal relation to predict the effects of improved socio-economic health determinants on the tuberculosis epidemic (Carter et al. 2018). The state-space models suffer from the limitations, namely: (i) they are not scalable for large data, (ii) they can not capture the heterogeneity in the data.

3 Materials and Methods

For the current study we focus on the incidence records and socio-economic parameters for the 174 countries for which the complete data were available, as listed in the Fig. 1. The socio economic parameters considered in the current study are namely, social capital (as a measure of social cohesion)3, population density(person/square Km)4, percentage of population below poverty line (as an indicator of incidence of material poverty) [44], Human Development Index (HDI) Value (2019) [45], life expectancy at birth (as a measure of nutritional status) [46], sex ratio (as an indicator of gender based discrimination) 5, Health care Access and Quality Index (HAQI) (as an estimator of availability and accessibility of quality health services) 6, Trade Openness Index (as an indicator of global connectedness) 7, percentage of urban population out of total population (as a measure of extent of

3 http://solability.com/the-global-sustainable-competitiveness-index/the-index/social-capital
4 http://web.archive.org/web/20161116131638ifhttps://esa.un.org/unpd/wpp/Download/Standard/Population/
5 http://www.cia.gov/library/publications/the-world-factbook/fields/351.html
6 https://ourworldindata.org/grapher/healthcare-access-and-quality-index
7 https://www.theglobaleconomy.com/rankings/trade_openness
urbanization)\(^8\) and population. The daily incidence count and cumulative incidence data are collected from World Health Organisation (WHO) situation reports dated 21.01.2020 till 31.07.2020\(^9\).

The missing values are imputed using moving average technique (Durbin 1959). It is important to note here that, the time series data formed by the daily incidence count is non-stationary, as proved by the Augmented Dicky Fuller Test. Further, as intuitively, the growth rate of incidence depends on the number of infected people itself, the time series data is nonlinear. From the cumulative incidence records we further compute the real valued temporal variable, days lapsed since the first infection is reported. This derived temporal variable is known for historical as well as future time steps. Thus it augments the input dataset for better prediction.

\(^8\) http://data.worldbank.org/indicator/SP.URB.TOTL.in.zs
\(^9\) https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200702-covid-19-sitrep-164.pdf?sfvrsn=ac074f58_2

**Fig. 1** List of the countries used in the study
The major challenges in the study are about incorporating the effects of the static socio-economic factors on varying length non-stationary nonlinear temporal cumulative incidence data. Amongst the existing data-driven prediction techniques, deep neural network based models allow many flexibilities. TFT (Lim et al. 2019) is one of the recently proposed potential model that is shown to capture the effects of various static factors as well as known dynamic variants on a temporal target variable while demonstrating high prediction accuracy. TFT can also capture non-stationarity and nonlinearity in data. Further, TFT do not require all the time series to be of same length for its training. For the multi-horizon incidence prediction we use the TFT model. For the analysis of the influence of the socio-economic factors on the estimated country-wise predictions we use the weights of the context vectors generated by TFT. The correlation between the weights of the context vectors and the total incidence count are analyzed using Pearson correlation (Benesty et al. 2009).

To cluster the countries on the basis of influential explanatory variables and to search for intra-cluster and inter-cluster characteristics we use hierarchical clustering (Ward 1963). The optimal cut on the hierarchical clustering is decided based on the Silhouette score (Reddy and Vinzamuri 2013).

### 3.1 Deep Neural Network for Prediction

We use the deep neural network architecture termed as Temporal Fusion Transformer (TFT) (Lim et al. 2019) for the cumulative incidence prediction of each of the countries.

Let the static covariates, the historical cumulative incidence record and the derived temporal feature namely days lapsed since the first occurrence for a given country be represented by \( I = \{s, x_{0}, ..., t\} \) and \( z_{0}, ..., t, t+s \) respectively, where \( t \) is the period for which incidence data is available and \( \tau \) is the future period for incidence prediction. The multi-horizon forecasting estimates the cumulative incidence using quantile output that generates prediction intervals along with the point prediction by the simultaneous prediction of 10\(^{th}\), 50\(^{th}\) and 90\(^{th}\) percentiles at each time step. Quantile forecasts, for a specified quantile \( q \) is generated using the following equation that linearly transform the output from the temporal fusion decoder:

\[
\hat{y}(q, t, \tau) = W_q \tilde{y}(t, \tau) + b_q
\]

where \( W_q \in R^{1 \times d}, b_q \in R \) are linear coefficients for \( q \), \( \tilde{y}(t, \tau) \) is the TFT generated output and \( d \) is same as the dimension of \( \tilde{y}(t, \tau) \). Forecasts are generated for future horizons where \( \tau \in \{1, ..., \tau_{\text{max}}\} \).

To learn the interdependence between the static socio-economic factors, the varying length non-stationary nonlinear temporal cumulative incidence data, the derived temporal variable and the future incidence value, temporal fusion transformer makes use of several layers of non-linear computation, self-attention and dense neural network. Figure 2 shows the high level architecture of Temporal Fusion Transformer(TFT). The inputs to the TFT are the \( \{s, x_{0}, ..., t\} \) and \( z_{0}, ..., t, t+s \) for each country and the outputs are the prediction intervals.
Fig. 2 The figure shows the architecture of the TFT model. For each country, inputs to the model are static covariates, the historical cumulative incidence record and the derived temporal feature and the outputs are the prediction interval \( (y_{t+1,10}, y_{t+1,50}, y_{t+1,90}) \) for each time step.
\{ (y_{t+1.10}, y_{t+1.50}, y_{t+1.90}), \ldots, (y_{t+\tau,1.10}, y_{t+\tau,50}, y_{t+\tau,90}) \} \text{ for } \tau \in \{1, \ldots, \tau_{\text{max}} \}. \] The major layers of the deep neural network TFT and their constituents are discussed in the following section.

### 3.1.1 Gated Residual Network

Given the worldwide COVID-19 incidence data and the static socio economic data, it is difficult to determine whether a linear processing is sufficient or a non-linear processing is required as the relationship between the exogenous input variables and the output incidence variable is unknown. TFT (Lim et al. 2019) uses the Gated Residual Network (GRN) that apply non-linear processing only if needed. The GRN generates a context vector \( c \) from the input data represented as \( I \) using the following equations:

\[
\text{GRN}_{\omega}(I, c) = \text{LayerNorm}(I + \text{GLU}_{\omega}(\eta_1)),
\]

\[
\eta_1 = W_{1,\omega} \eta_2 + b_{1,\omega},
\]

\[
\eta_2 = \text{ELU}(W_{2,\omega} I + W_{3,\omega} c + b_{2,\omega}),
\]

where \( \eta_1, \eta_2 \in R^{d_{\text{model}}} \) are intermediate layers, \( d_{\text{model}} \) is the hidden state size used in TFT and \( \omega \) is an index to denote weight sharing. Layer normalization denoted by \( \text{LayerNorm}(.) \) re-centers and re-scales its activations using the normalization terms: the mean and the variance of the summed inputs within each layer (Ba et al. 2016).

The function \( \text{ELU}(.) \) used in the GRN is the exponential linear unit (ELU) defined as below:

\[
\text{ELU}(x) = \begin{cases} 
  x & \text{if } x > 0 \\
  \alpha(\exp(x) - 1) & \text{if } x \leq 0
\end{cases}
\]

where \( \alpha > 0 \). When \( x = W_{2,\omega} a + W_{3,\omega} c + b_{2,\omega} \gg 0 \), the ELU activation acts as an identity function and when \( x = W_{2,\omega} a + W_{3,\omega} c + b_{2,\omega} \ll 0 \), the ELU activation generates a constant output \( \alpha \), yielding a linear behavior. The advantage of using ELU activation is that the common problem of vanishing gradient effect is reduced in ELU activation as the positive part of the function is an identity function with derivative as one.

To suppress a part of the architecture that are not required for the given dataset, TFT uses the component gating layers based on Gated Linear Units (GLUs) (Dauphin et al. 2017).

Letting \( \gamma \in R^{d_{\text{model}}} \) be the input, the GLU takes the following form:

\[
\text{GLU}_{\omega}(\gamma) = \sigma(W_{4,\omega} \gamma + b_{4,\omega}) \odot (W_{5,\omega} \gamma + b_{5,\omega}),
\]

where \( \sigma(.) \) is the sigmoid activation function, \( W_{(\ldots)} \in R^{d_{\text{model}} \times d_{\text{model}}} \), \( b_{(\ldots)} \in R^{d_{\text{model}}} \) are the weights and biases, \( \odot \) is the element wise Hadamard product, and \( d_{\text{model}} \) is the hidden state size used in TFT.

GLU controls the extent to which the GRN contributes to the original input \( I \). As the GLU outputs could be significantly small and close to 0, it can help to skip over the layer entirely if necessary.
3.1.2 Variable Selection Networks

Amongst input dataset of the real valued static socio-economic factors, categorical variable that is country identifier, real valued historical incidence record and computed real valued temporal variables like days lapsed since the first occurrence, the relevance and specific contribution of the variables in predicting the output variable are unknown. Variable selection network of TFT (Lim et al. 2019) is used to select relevant input variables from both static variables and time-dependent variables at each time step.

Entity embeddings (Guo and Berkhahn 2016) is used for the categorical variable as feature representations while linear transformations is used for continuous real variables. Each input variable is transformed into a \( d_{\text{model}} \)-dimensional vector that matches the dimensions in subsequent layers for skip connections. All static, historical and future inputs for derived variables make use of separate variable selection networks.

The working principle of the variable selection network for historical inputs is discussed below: Let \( x_t^{(j)} \in \mathbb{R}^{d_{\text{model}}} \) denote the transformed input of the \( j \)-th variable at time \( t \), with \( X_t = [x_t^{(1)^T}, \ldots, x_t^{(H)^T}] \) being the flattened vector of all past inputs at time \( t \). Variable selection weights are generated by feeding both \( X_t \) and an external context vector \( c_s \) through a GRN, followed by a Softmax layer as:

\[
vt = \text{Softmax} \left( \text{GRN}_{vt}(X_t, c_s) \right)
\]

where \( vt \in \mathbb{R}^H \) is a vector of variable selection weights, and \( c_s \) is obtained from a static covariate encoder as described in Sect. 3.1.3. For the static variables, it can be noted that the context vector \( c_s \) is omitted as it already has access to static information.

At each time step, an additional layer of non-linear processing is applied by feeding each \( x_t^{(j)} \) through its own GRN as given by the equation below:

\[
x_t^{(j)} = \text{GRN}_{x_t^{(j)}}(\tilde{x}_t^{(j)})
\]

where \( \tilde{x}_t^{(j)} \) is the processed feature vector for variable \( j \). We note that each variable has its own \( \text{GRN}_{x_t^{(j)}} \), with weights shared across all time steps \( t \). Processed features are then weighted by their variable selection weights and combined as:

\[
\tilde{x}_t = \sum_{j=1}^{H} v_t^{(j)} \tilde{x}_t^{(j)}
\]

where \( v_t^{(j)} \) is the \( j \)-th element of vector \( vt \).

3.1.3 Static Covariate Encoder

The TFT (Lim et al. 2019) model integrate information from static data, using separate GRN encoders to produce four different context vectors that are wired into
the following layers in the temporal fusion decoder where static variables play an important role in processing:

- Variable selection networks for temporal features,
- Local processing layer for temporal features, and
- Static enrichment layer for enriching the temporal features with static information.

### 3.1.4 Temporal Fusion Decoder

The temporal fusion decoder uses a series of layers that learn the temporal relationships present in the dataset. In time series data, use of pattern information in addition to point values can help to achieve better performance. For example, leveraging the relationship of a point value in relation to its surrounding values can help us identify the change-points. To that end, TFT use a sequence-to-sequence layer of LSTM encoders and decoders by feeding the historical values \( \tilde{x}_{t-k:t} \) into the encoder and future values \( \tilde{x}_{t+1:t+s_{\max}} \) into the decoder. This subsequently generates a set of temporal features \( f(t,n) \in \{f(t,-k), \ldots, f(t,s_{\max})\} \) with \( n \) being a position index, that serve as inputs into the temporal fusion decoder.

A gated skip connection is used over this layer as:

\[
\tilde{f}(t,n) = \text{LayerNorm}(\tilde{x}_{t+n} + \text{GLU}(f(t,n))),
\]

where \( n \in [1:-k, s_{\max}] \) is a position index.

To consider the influence of the static variables on the temporal dynamics, TFT uses a static enrichment layer of the following form:

\[
\theta(t,n) = \text{GRN_0}(\tilde{f}(t,n), c_e)
\]

where the index is \( n \), the weights of \( \text{GRN_0} \) are shared across the entire layer, and \( c_e \) is a context vector from a static covariate encoder.

The output of the static enrichment layer is applied to the temporal self-attention layer to learn the long term relationships from the historical temporal incidence data and historical derived temporal features only that scale the input values depending on the inter-temporal relationships. This layer filters out the effect of future values of the incidence records on the output.

An additional non-linear processing is applied to the outputs of the self-attention layer in the Position-wise Feed-forward Layer using GRN followed by a LayerNorm.

The interested readers may refer to (Lim et al. 2019) for detail functionality of the layers.

### 3.1.5 Training and Loss Functions

We partition all the time series data into - a training set for learning, a validation set for hyper parameter optimization, and a hold-out test set for performance evaluation. Further, each time series data is converted into a format containing a
collection of input features and output value. The input features are the static variates, seven days historical cumulative incidence records and seven days past and seven days future computed temporal variable, days lapsed since the first occurrence. The output value is seven days ahead cumulative incidence records. TFT is trained by minimizing the quantile loss, summed across all quantile outputs as:

\[
L(\Omega) = \sum_{y_t \in \Omega} \sum_{q \in Q} \sum_{\tau=1}^{\tau_{\text{max}}} \frac{QL(y_t, \hat{y}(q, \tau), q)}{M^{\tau_{\text{max}}}}
\]

where \( \Omega \) is the domain of training data containing \( M = 174 \) sample countries, \( Q \) is the set of output quantiles \( Q = \{0.1, 0.5, 0.9\} \) in our experiments and \( \tau_{\text{max}} \) is the horizon of future prediction. The variables \( y_t \) and \( \hat{y}(\cdot) \) are the actual incidence value and the \( q \)th quantile predicted output while \( (\cdot)_+ = \max(0, \cdot) \).

### 3.1.6 Hyper Parameter Optimization

The validation data set is used for hyper parameter tuning. Hyper parameter optimization is carried out by using a random search with 1000 iterations out of all possible combinations of all the tunable hyper parameters that minimize the loss function over the validation dataset. The search ranges for all hyper parameters are same as that mentioned in Lim et al. (2019). The list of hyper parameters and their optimal values are mentioned in the Fig. 3.

### 3.1.7 Measure for Prediction Accuracy

To assess the accuracy of the prediction we compute the mean absolute percentage error (MAPE) on the test dataset. For a given time series \( \{A_i\}_{i=1}^{n} \) of length \( n \) and its predicted values by an algorithm as \( \{F_i\}_{i=1}^{n} \), MAPE computes the accuracy as a ratio defined by the formula:

| Hyperparameter          | Value | Hyperparameter          | Value |
|------------------------|-------|------------------------|-------|
| State size             | 240   | Number of epochs       | 100   |
| Dropout rate           | 0.3   | Early stopping patience| 10    |
| Minibatch size         | 256   | Multiprocessing workers| 10    |
| Learning rate          | 0.0001| Number of encoder steps| 7     |
| Maximum gradient norm  | 0.01  | Total Time Steps       | 14    |
| Number of heads        | 1     |                        |       |

Fig. 3 Optimal hyper-parameter values for TFT model
\[ M = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \]

For a database of time series with \( D = \{A^j\}_{j=1,...,k} \) where \( A^j = \{A^j_i\}_{i=1,...,n_j} \) is a time series of length \( n_j \) and its predicted values by an algorithm as \( \{F^j_i\}_{i=1,...,n_j} \) the MAPE is computed as given by the formula:

\[ M^D = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{i=1}^{n_j} \left| \frac{A^j_i - F^j_i}{A^j_i} \right| \]

For the current study, \( D \) is the test dataset, \( k \) is the number of countries and \( n_j \) is the length of time series data for the \( j \)th country in the test dataset. The 50th percentile prediction of TFT is used for MAPE computation.

### 3.1.8 Out-of-Sample Prediction Using TFT

A rolling forecast mechanism is used for out-of-sample prediction. For every country, a new test data with random feature values is fed into the trained TFT model that yields the seven days ahead cumulative incidence prediction. The predicted values are appended in the training dataset to re-train the model. The process is repeated until 270 days ahead predictions are obtained.

The out-of-sample prediction gives an absolute number of incidence count for each country. We propose the following normalization terms that would enable us to compare the normalized incidence across countries by controlling the effect of population. The first normalized term \( I_{\text{Norm}} \) is defined as below.

\[ I_{\text{Norm}} = \frac{I + P}{P + T} \]

where \( I \) is the total incidence count, \( P \) is the population of the country and \( T \) is the world population. The proposed normalization factor imposes higher penalty to the countries with fewer population as compared to the countries with larger population. For example in a country with about 95,000 population and in another country with about 250,000,000 population if the total incidences are 5000 and 1315000 respectively, the \( I_{\text{Norm}} \) would yield 0.052 for both the countries. However, a total count of 1315000 infected people has globally higher impact on further spread, assuming other variates as same for both the countries. On the contrary, considering world population as 7.5 billion, the \( I_{\text{Norm}} \) values for the said countries are \( 0.13 \times 10^{-4} \) and \( 0.35 \times 10^{-2} \) respectively, indicating the country with larger population has higher impact. The other normalized term used in the study is z-transformed \( I_{\text{Norm}} \) that allow us to compare the countries by adjusting the scale variance of \( I_{\text{Norm}} \).
3.2 Correlation

To estimate the linear correlation between the total incidence count and the influence of the static socio-economic factors we use Pearson correlation coefficient ($r$) (Benesty et al. 2009), that measures the linear correlation between two variables $X$ and $Y$. Pearson’s correlation coefficient ($r$) is the fraction of the covariance of the two variables $X$ and $Y$ and the product of their standard deviations. The value of $r$ ranges between $+1$ and $-1$ with $+1$ indicating a total positive linear correlation, and $-1$ is total negative linear correlation. The value $r = 0$ however indicates that there is no linear correlation between the variables.

3.3 Hierarchical Clustering

To analyze how different countries form groups in terms of the similarity between the influencing explanatory variables we use hierarchical clustering. Hierarchical clustering is one of the popular techniques for identifying groups in the dataset. The advantage of hierarchical clustering over other common clustering techniques is that it does not require to specify the number of clusters to be generated. In addition, hierarchical clustering yields a tree-based representation of the observations, called a dendrogram. The working principle of the hierarchical agglomerative clustering is as follows.

- Initially each data point is considered to belong to its own cluster.
- At each successive steps, the most appropriate pair of clusters are merged into one cluster. The appropriateness of a pair of clusters for merging is decided by certain criterion.
- The iterative merging process continues until there is only one cluster.

Ward’s minimum variance method (Ward 1963) a popular hierarchical agglomerative clustering technique, suggests the selection of the pair of clusters for merging at each step that minimize the total within-cluster variance. During implementation, this is achieved by finding the pair of clusters that leads to minimum increase in total within-cluster variance ($E$) after merging. This increase is a sum of squared distances between cluster centers. Initially, all clusters are singletons and the initial distance between individual objects is the squared Euclidean distance. $E$ at this stage is computed as 0.

In hierarchical clustering the task of estimating the right number of clusters from an input dataset remains open. One approach to estimate the true right number of clusters is by optimizing a cluster validity index.

The cluster validity indexes that use just the clustering result of an algorithm and the similarity or dissimilarity measurements among the data points instead of true class information of each data point, are generally termed as internal cluster validity indexes. Silhouette Index is one of the most successfully used internal cluster validity indexes (Rousseeuw 1987). For a clustering solution of a dataset, the more the average Silhouette Score of all the data points closer to $+1$, the better the clustering solution is. For hierarchical clustering, we consider the various cuts...
resulting in different number of clusters \((k)\). For all possible cuts and for all values of \(k\) the one that leads to the maximum average Silhouette Index is considered as an estimate of the number of natural clusters in the dataset.

## 4 Results

In this section we discuss the results obtained for COVID-19 prediction study and the analysis of the influencing explanatory variables using the TFT model, correlation analysis and hierarchical clustering.

### 4.1 Prediction of Cumulative Infected Population

We show the MAPE values for every country using a world map in sub Fig. 4a and sub Fig. 4b shows the distribution of the MAPE computed for all the 174 countries. As shown in the Fig. 4b, MAPE follows a positively skewed density distribution

![Maps and graphs showing MAPE distribution and cumulative incidence plots for countries with MAPE above and below average.](image)

**Fig. 4** Sub figure a shows the MAPE for 174 countries. Sub figure b shows a positively skewed density distribution plot for prediction error in terms of MAPE with the mean highlighted with blue dotted line. Sub figure c shows the cumulative incidence plots for the countries having MAPE below average follow a logistic growth model. Sub figure d shows the cumulative incidence plots for the countries having MAPE above the average follow an exponential growth model. The color mapping for MAPE values are shown in the legends.
with the median and mean of MAPE as 0.0815 and 0.165 respectively. 80% of the countries have MAPE below 0.3. This shows that TFT exhibits high prediction accuracy. As shown in sub Fig. 4a the majority of European union (EU), south-east Asia, Australia, United States and some portion of Africa shows MAPE below average. The countries of South American continent, Mexico, India, Kazakhstan and most of the countries of Africa show MAPE higher than the average with Botswana, Namibia and Seychelles showing MAPE higher than 0.6. It may be noted that for countries showing high MAPE, the total incidence count and the days lapsed since first infection are very low as compared to the other countries. We next analyze the association between the MAPE error and the characteristics of the data. The MAPE error do not exhibit any strong correlation neither with parameters like days elapsed since first occurrence nor with total incidence count. Interestingly enough, when we plot the cumulative incidence plot for the countries having MAPE lower than the average and that higher than the average as shown in the sub Fig. 4c and d, it is observed that there are clearly two categories of growth model: (i) approximately 33% of total 174 countries have higher MAPE than average show exponential growth model , (ii) rest 66% of total 174 countries have lower MAPE than average show logistic growth model. The model seems to show bias towards the logistic growth model that is majority in the data. As a result, the prediction using TFT model is more accurate for the countries that show logistic growth model in cumulative incidence count than with the former one. Further, it may be concluded that the countries showing logistic growth model are either have successfully implemented preventive strategies like lock down or social distancing to flatten the natural exponential spread model of COVID-19 or they are at the very early stage of propagation with comparatively low total incidence count. From the results, while the first assumption is shown to be true for the EU countries, the second case is true for the African countries. The countries namely, India, Mexico, Columbia, Brazil, Argentina, Iraq, South Africa show exponential model of growth of cumulative incidence count which indicate that social distancing based preventive strategies might not have been adequate to contain the spread in such countries.

We next plot the predicted cumulative incidence range along with the actual cumulative incidence counts for the top 20 countries having minimum MAPE errors as shown in Fig. 5, and the predicted cumulative incidence range along with the actual cumulative incidence counts for the top 25 countries with highest cumulative incidence count till the experiment forecast date as shown in Fig. 6. It can be clearly seen that for the top 20 countries with minimum MAPE below the average value, the actual cumulative incidence count lies well with the prediction range as shown in the Fig. 5. From Fig. 6 it can be seen that for the countries namely, India, Mexico, Columbia, Brazil, Argentina, Iraq, South Africa the prediction could not match the actual cumulative incidence count. The reason for this, as explained before is the exponential behavior of the growth. The TFT model being biased towards the countries showing logistic growth model.
Future Prediction of Total Incidence

After achieving a satisfactory level of accuracy for prediction within test horizon, we use the TFT model for 270 days ahead future prediction. The basic assumption here is that the cumulative incidence count will grow following the same growth model that has been demonstrated during the training phase.

The cumulative incidence count as on 31 March 2021 along with the normalized incidence scores for countries with top 10 future incidence count, minimum 10 incidence count and 10 around the median are listed in Fig. 7. The top two most infected countries are namely, the US and Brazil with over 54 million and 21 million infected people while for India the predicted count on number of infected people is nearly 8.1 million. It is important to note that the $I_{\text{norm}}$ and z-normalized score for India are higher than that of other countries, indicating India being a highly populated country has higher impact on the spread of infection, followed by the US.

![Cumulative incidence count for the top 20 countries with minimum MAPE error during the experiment forecast period](image)

**Fig. 5** Cumulative incidence count for the top 20 countries with minimum MAPE error during the experiment forecast period

### 4.2 Future Prediction of Total Incidence

After achieving a satisfactory level of accuracy for prediction within test horizon, we use the TFT model for 270 days ahead future prediction. The basic assumption here is that the cumulative incidence count will grow following the same growth model that has been demonstrated during the training phase.

The cumulative incidence count as on 31 March 2021 along with the normalized incidence scores for countries with top 10 future incidence count, minimum 10 incidence count and 10 around the median are listed in Fig. 7. The top two most infected countries are namely, the US and Brazil with over 54 million and 21 million infected people while for India the predicted count on number of infected people is nearly 8.1 million. It is important to note that the $I_{\text{norm}}$ and z-normalized score for India are higher than that of other countries, indicating India being a highly populated country has higher impact on the spread of infection, followed by the US.
and Brazil. The countries with least incidence count are generally the country with less than 1 million population with the exception of Lao People’s Democratic Republic having 7.2 million population. The $I_{norm}$ score of Lao People’s Democratic Republic is 0.00093 that is significantly higher than that of the rest of the countries in least 10 incidence count. The median of future incidence is 87168.5 and mean of future incidence is 1230515. Bulgaria has future incidence just above the median. The other countries with future incidence around the median are Ethiopia and Bosnia and Herzegovina.

![Fig. 6 Cumulative incidence count for the top 25 countries with highest cumulative incidence count during the experiment forecast period](image)

Table 7 shows 270 days ahead prediction with median incidence as 87168.5 and mean incidence as 1230515: Top 10 lists the 10 countries having maximum incidence, Least 10 lists the 10 countries having minimum incidence, 10 around Median lists the 10 countries around the median value of incidence. The normalized scores incidence are also reported in the table

and Brazil. The countries with least incidence count are generally the country with less than 1 million population with the exception of Lao People’s Democratic Republic having 7.2 million population. The $I_{norm}$ score of Lao People’s Democratic Republic is 0.00093 that is significantly higher than that of the rest of the countries in least 10 incidence count. The median of future incidence is 87168.5 and mean of future incidence is 1230515. Bulgaria has future incidence just above the median. The other countries with future incidence around the median are Ethiopia and Bosnia and Herzegovina.
The cumulative incidence count as on 31 March 2021 along with the normalized scores for all 174 countries are listed in Fig. 11. From Fig.11 we find that a total of nearly 214 million people across 174 countries would be infected by COVID-19 by 31 March 2021.

It is important to understand here that the model was primarily trained on the data when social distancing and other preventive mechanisms were imposed on the majority of the countries causing a flattening of the curves. However, if such social restrictions are relaxed then the incidence count would increase by following an exponential model resulting in a higher number than the projected one. On the contrary, if additional interventions are more successfully applied, then the incidence count may be fewer than the projected figure.

4.3 Influence of the Socio-Economic Parameters

One of the key features of the TFT model is that it can capture the contribution of the static variables for prediction. The variable weights from the static context vectors are used to understand the possible contribution of different socio-economic variants in different countries. The pie chart in Fig. 8a and b where the area is proportional to the the percentage of countries having a particular socio-economic explanatory variable shows the distribution of the top two most influencing static factor. In Fig. 8c and d we show the top two most influencing socio-economic factors for all 174 countries.

It can be seen from the Fig. 8a and c that Trade Openness Index is the most influencing static factor for 16.67% countries particularly in the Indian subcontinent, Russia and Mexico. While population and sex ratio are the most influencing factors for 13.8% and 12.64% countries. As shown in Fig. 8b and d the second most influencing static factors population, percentage of population below poverty line (BPL) , urban population and social capital seems to be the predominant ones. For example, in India the top two most influencing factors are Trade Openness Index and life expectancy, in China and Indonesia the top two most influencing factors are sex ratio and percentage of population below poverty line (BPL) while in United States the factors are population and Trade Openness Index respectively.

We also compute the Pearson correlation between the total incidence count and context weights of the static variates as shown in the table 1. It is observed in table 1 that total incidence count is positively correlated with population, Trade Openness Index , social capital, population density, sex ratio and negatively correlated with HDI, life expectancy, HAQI, urban population, percentage of population below poverty line (BPL). The magnitude of correlation is 0.11 with population, 0.068 with Trade Openness Index, 0.097 with social capital, 0.017 with sex ratio and 0.002 with population density, about -0.03 with percentage of population below poverty line (BPL) and between -0.09 and -0.04 with life expectancy, urban population, HDI, HAQI. Though the linear correlation between total incidence count and the static variates are low the TFT model has demonstrated that these variables have good explanatory power to explain the variations in predicted values.
Table 1

| Variate                  | Cor  | Variate                  | Cor  |
|--------------------------|------|--------------------------|------|
| Population               | 0.116| Trade Openness Index     | 0.068|
| Social capital           | 0.097| Population density       | 0.002|
| Sex ratio                | 0.017| HDI                      | -0.083|
| Life expectancy          | -0.042| HAQI                    | -0.087|
| Urban population         | -0.081| Poverty                 | -0.03 |

Fig. 8 pie chart in sub figure (a) shows the distribution of the most influencing socio-economic factor while the pie chart in sub figure (b) shows the distribution of the second most influencing socio-economic factor. Sub figure (c) shows the most influencing socio-economic factor for each country while sub figure (d) shows the second highest influencing socio-economic factor for each country.
4.4 Clustering Countries Based on Influencing Variables

We carried out a hierarchical clustering of the countries based on their similarity of the influencing explanatory variables. The influence score of the 10 explanatory variables, namely, social capital, population density, percentage of population below poverty line (BPL), HDI, sex ratio, HAQI, Trade Openness Index, urban population ratio, and population are normalized using their mean and standard deviation. Subsequently the Euclidean distance is calculated used for similarity score between two countries. Figure 12 shows the hierarchical clustering of the countries based on Euclidean distance.

In hierarchical clustering we use Silhouette score to determine the best cut. Figure 9 shows the Silhouette scores when total number of clusters \((k)\) is varied from 1, \ldots, 15. The peaks are obtained at \(k = 10\) and \(k = 12\). The cuts obtained for both \(k = 10\) and \(k = 12\) yield similar results.

The Fig. 10 shows the hierarchical clustering details based on \(k = 10\). Sub Fig. 10a shows the \(z\)-transformed static influence scores for every country annotated by the clusters and sub Fig. 10b shows the belongingness of the countries to each cluster. From Fig. 10a the intra cluster similarities and inter cluster dissimilarities can be seen.

5 Discussion

Prediction about a disaster plays an important role to minimize the damage and economic losses. With the emergence of good quality data, deep neural network has appeared as a potent tool of prediction. But the predictive research on COVID 19 started to face steep challenges as the future spread is not only dependent on the historical spread data but also on the non-pharmaceutical static socio-economic variables. This challenge became more acute due to the large variation in country wise levels of these static variables. Within these uncertainties this work has tried to
develop a data driven multi-horizon forecasting analysis based on Deep Neural Network considering the variability within the static socio-economic explanatory variables. The specific objective of this study is to predict inter-country incidences of infection arising out of COVID 19 for the next 270 days. Apart from this predictive analysis this work wants to locate the influences of individual socio-economic factors on the estimated country-wise predictions. Estimating the influence of explanatory variables, developing cluster of countries on the basis of influential explanatory variables and thus to search for intra-cluster as well as inter-cluster characteristics are also under the purview of this study. To that respect this study has used the idea of temporal fusion transformers for the predictions. For the analysis of the influence of the socio-economic factors on the estimated country-wise predictions the weights of the context vectors generated by TFT have been used. To cluster the countries on the basis of influential explanatory variables and to search for intra-cluster characteristics Hierarchical clustering has been tried. The optimal cut on the hierarchical clustering is decided on the basis of Silhouette score.

From the findings it appears that nearly 214 million people across 174 countries (the countries for which complete data was available) would be infected by COVID 19 by 31 March 2021. The top two most infected countries are the US and Brazil with over 54 million and 21 million infected people. There also exists a wide inter-country variation in the predicted spread of infection as long as the absolute figures are concerned (Highest: United States of America - 54,802,240 and Lowest: Papua New Guinea - 276 on 31st March 2021). This picture of large variation is not different for the normalized relative figures of prediction also. Incidentally it appear from the pictures of the absolute as well as relative predictions that the spread of
infection is much higher in populous countries where the infection is much lower in less populous countries.

These predictions on the basis of TFT follow higher degrees of accuracy as in this study 80% of the countries have Mean Absolute Percentage Error (MAPE) below 0.3. Majority of European Union states, south-east Asian countries, Australia, United States of America and some portion of Africa have shown MAPE below average. The countries of south American continent, Mexico, India, Kazakhstan and most of the countries of Africa have shown MAPE higher than the average with Botswana, Namibia and Seychelles showing MAPE higher than 0.6. It is observed that approximately 33% of total 174 countries with higher MAPE than average show exponential growth model where the rest 66% of total 174 countries with lower MAPE than average show logistic growth model. The study seems to show bias towards the logistic growth model that is majority in the data. As a result, this prediction using TFT model is more accurate for the countries that show logistic growth model in cumulative incidence count than with the countries that follow exponential growth model. Further, it may be said that the countries showing logistic growth model are either have successfully implemented non-pharmaceutical preventive strategies or they are at the very early stage of propagation with comparatively low total incidence count. From the results, while the first assumption is shown to be true for the EU countries, the second case is true for the African countries. The countries namely, India, Mexico, Columbia, Brazil, Argentina, Iraq, South Africa show exponential model of growth of cumulative incidence count which indicate that preventive strategies were not adequate to contain the spread in such countries. Apart from these seven countries satisfactory level of accuracy for prediction within test horizon was observed. But it is to be kept in mind that the current predictive model was primarily trained on the data when social distancing and other preventive measures were imposed on the majority of the countries causing a flattening of the curves. However, if such social restrictions are relaxed then the incidence count would increase by following an exponential model resulting in a higher number than the projected one. On the contrary, if additional interventions are more successfully applied, then the incidence count may be fewer than the projected figures. But if the status quo is maintained this large variation in the types of trajectories and inter-country infection counts may be attributed to the variation in the socio-economic and institutional factors of the concerned economies.

It is observed that total incidence count is positively correlated with population, Trade Openness Index, social capital, population density, sex ratio and negatively correlated with HDI, life expectancy, HAQI, urban population and percentage of population below poverty line. Thus it appears that greater the population size, stronger the global connectedness, larger the social cohesion, higher the population density and meaningful the gender based discrimination higher will be the future spread. On the other hand greater the development level, higher the nutritional status, greater the access to quality health services, greater the urban population and greater the material poverty lesser will be the future spread. Though the current model accurately described the variation in predictions on the basis of these socio-economic variates the correlation of the future spread is poorly correlated with these.
variables. This apparent contradiction may be attributed to the existence of the countries within the sample for which the MAPE is higher than the mean. This weak correlation can also be attributed to the variation in influence of these variables on the country-wise future predictions. This study has observed wide variation in country-wise most influencing factor and second most important influencing factor. From the distributions of top most influencing static variable and second most influencing static variable it appears that Trade Openness Index, population, urban population, population density, sex ratio, poverty and social capital have played important roles in influencing the trajectories of COVID 19 induced infections. Trade Openness Index is the most influencing static factor for 16.67% countries particularly in the Indian subcontinent and in countries like Russia and Mexico. Population and sex ratio are most influencing factors for 13.8% and 12.64% countries respectively. Within the second most influencing static factors population, percentage of population below poverty line, urban population and social capital seems to be the predominant ones. In India the top two most influencing factors are Trade Openness Index and life expectancy, in China and Indonesia the top two most influencing factors are sex ratio and percentage of population below poverty line while in United States the factors are population and Trade Openness Index respectively. The positive relationship of the future spread with population size, population density, gender based discrimination and negative relationship of the future incidence with development as well as nutritional status, accessibility to quality health care reminds that more sustainable development policy is needed to restrict these types of unforeseen biological calamities. The negative correlation between extent of urbanization and future incidence implies the fact that improvements in the modernity within the lifestyle can mitigate the menace of infection. On the other hand the negative correlation between incidence of poverty and incidence of infection demonstrates that affluence may not guarantee less infection.

Some interesting findings are observed from the hierarchical clustering of the countries based on their similarity of the influencing explanatory variables. On the basis of Silhouette score ten clusters of the 174 sample countries are developed. The belongingness of the countries to the respective clusters and z-transformed static influence scores of every country reflected some interesting outcomes about the inter-cluster and intra-cluster characteristics. It is observed that none of the countries from North America is present in cluster 1. A small number of countries from Asia, Europe, South America, Africa and Australia are found in this cluster. The major countries in this cluster are Venezuela, Botswana, Spain, Ukraine, New Zealand, Nepal, Afghanistan and Malaysia. In these countries infection is taken forward mainly by urban population. The important countries in cluster 2 are Russian Federation, Mongolia, Laos, Georgia, Turkmenistan, France, Netherlands, Denmark, Chad, Kenya and Columbia. Like Cluster 1 not a single country from North America is present here. In the same tune South America and Africa have a minimal presence here. Here the countries have been clustered in a group due to the importance of population density, poverty and social capital in carrying the
infection within them. Cluster 3 is mainly formed by Australia, India, Oman, Iraq, Finland, Algeria, Sudan, and South Sudan. Incidentally, countries from North and South America are almost absent here. Here the infection spread is carried mainly by life expectancy and loosely by social capital with urban population. Cluster 4 countries are mainly found in Africa. Here the major member countries are Nigeria, Angola, and Uruguay. In this cluster HAQI has mainly influenced the predicted infection. Cluster 5 does not have any country from South America. The important mentions here are Mexico, Libya, Niger, Pakistan, Iran, and Kazakhstan. Trade Openness Index has played the most influential role in the future spread of the infection in this cluster. The important countries which have developed Cluster 6 are Argentina, Paraguay, Japan, Saudi Arabia, Yemen, Ethiopia, South Africa, Namibia, Zimbabwe, Mozambique, Madagascar, and Czech Republic. Here the members are mostly from South America and Africa; North America is absent; Europe is almost absent and Asia has minimal presence. Infection in this cluster is influenced mainly by HDI and loosely by urban population and population size. Cluster 7 is absent in North America and has minimal presence in Africa. The countries which need mention in this cluster are Brazil, Ecuador, Chile, Gabon, Tanzania, Norway, Sweden, Belarus, Romania, China, Thailand, and Indonesia. This cluster is influenced mainly by sex ratio. In Cluster 8 the important countries are USA, UK, Germany, Belgium, Moldavia, Egypt, Zambia, and Uganda. Interestingly it is observed that this cluster is not present at Asia. Within this cluster the spread of infection can mainly be explained by population size. Second most influential variable in this cluster is social capital. Cluster 9 has the members like Canada, French Guyana, Poland, Italy, Portugal, Turkey, Vietnam, and Tajikistan with other countries. This cluster is absent in Africa; almost absent in Asia and South America. Here the predicted spread is mainly influenced by social capital and population density. Finally, we find Democratic Republic of Congo, Equatorial Guinea, and Senegal as important presence within Cluster 10. Incidentally this cluster is not present in North America, South America, Asia, Europe, and Australia. Within this cluster poverty has appeared as the most important reason behind the surge of infection. Some interesting spatial patterns in the formation of clusters are observed. Found that none of the North American countries are found in clusters where urban population, poverty, HAQI, HDI, and sex ratio have taken most influential roles to form the clusters. In the same tune South American and Asian countries are not found in the clusters where Trade Openness Index and HAQI have taken the most influential roles respectively. Australian countries are not found in the clusters where HAQI and HDI have played the most influential role to form the clusters. No such pattern is found for the European states.

6 Conclusion and Future Work

At the end of this paper we can state that the predicted level of infection from COVID 19 for the whole world is huge. The country-wise variation within this prediction is also large. But for the populous countries the impact is much higher.
Within the sample, a larger proportion of countries are following logistic growth pattern in infection spread and for these countries the prediction has become more accurate. It is observed that greater the population size, stronger the global connectedness, larger the social cohesion, higher the population density and meaningful the gender based discrimination higher will be the future spread. On the other hand greater the development level, higher the nutritional status, greater the access to quality health services, greater the urban population and greater the material poverty lesser will be the future spread.

This study has also observed wide variation in country-wise most influencing factor and second most important influencing factors. It appears from the hierarchical clustering of the countries on the basis of similarity of the influencing explanatory variables that none of the North American countries are found in clusters where extent of urbanisation, incidence of poverty, accessibility to quality health care, level of development and sex ratio have taken the most influential roles to form the cluster. In the same tune South American and Asian countries are not found in the clusters where international connectedness and accessibility to quality health care have taken the most influential role respectively. Australian countries are not found in the clusters where accessibility to quality health care and level of development have played the most influential role to form the clusters. No such pattern is found for the European states. But beyond this clustering analysis if we analyse the influencing factors worldwide more intimately, it will appear that more sustainable development policy is needed to restrict these types of unforeseen biological calamities as future spread of infection is positively related with population size, population density, gender based discrimination. The negative relationship of the future incidence with development as well as nutritional status, accessibility to quality health care also corroborates the necessity of the sustainable development programmes. At the same time the negative correlation between extent of urbanization and future incidence implies the fact that improvements in the modernity within the lifestyle can mitigate the menace of infection. But above all, the negative correlation between incidence of poverty and incidence of infection demonstrates that affluence may not guarantee less infection. Thus this study concludes that to minimize the vulnerability towards unforeseen biological calamities modern and sustainable development policies are needed. But these policies should vary from economy to economies due to the variation in socio-economic status of the countries worldwide.

This research opens up an avenue for future research on prediction based computational economics. A few important research directions are highlighted below.

- It would be interesting to study how stationary variates and multiple possibly inter-related temporal variables can be used for the predictive study. For example, how possibly geographic conditions particularly, weather along with socio-economic factors influences spread of an infectious disease like COVID 19.
The high risk systems might require an early prediction that in turn implies limited amount of data would be available for training purposes. As deep neural network require a lot of data for training the system, it would be interesting to explore how synthetic data can be generated and augmented in such cases or the concepts of dilation can be used to reduce the total number of trainable parameters. It would be also interesting to harness the potential of shallow sequential neural networks for such early prediction based analysis.

Another area of future research would be to understand and estimate the importance of governance and representative democracy in the spread of COVID 19 infection related measures namely incidence count and the shape of incidence trajectory, recovery rate, fatality rate etc, as some studies suggest that state society binary may have a distinct influence on epidemiological outcomes.

Appendix

For the interested readers we list the predicted 270 days ahead incidence for all the 174 countries in the table below along with the normalized scores. Figure 12 shows the dendogram of the hierarchical clustering formed by the 174 countries based on their influence scores of all the socio-economic variates.

![Fig. 11 Table shows 270 days ahead prediction for all 174 countries along with their normalized incidence scores](image)
Fig. 12 The hierarchical clustering of the 174 countries based on the similarity of the socio-economic explanatory variables.
References

Aburas, H. M., Cetiner, B. G., & Sari, M. (2010). Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications, 37*(6), 4256.

Ahmar, A. S., & Del Val, E. B. (2020). SutteARIMA: Short-term forecasting method, a case: Covid-19 and stock market in Spain. *Science of The Total Environment, 729*, 138883.

Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., & Alhyari, S. (2020). COVID-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications, 12*, 168.

Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals, 139*, 110017.

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health and Surveillance, 6*(2), e18828.

Bacàër, N. (2011). Ross and malaria (1911). In *A short history of mathematical population dynamics* (pp. 65–69) Springer.

Bacàër, N. (2011). McKendrick and Kermack on epidemic modelling (1926–1927). In *A short history of mathematical population dynamics* (pp. 89–96) Springer.

Bai, Y., & Jin, Z. (2005). Prediction of SARS epidemic by BP neural networks with online prediction strategy. *Chaos, Solitons & Fractals, 26*(2), 559.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4), Springer.

Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B., & Sledge, D. (2020). The challenges of modeling and forecasting the spread of COVID-19. *Proceedings of the National Academy of Sciences, 117*(29), 16732.

Carter, D. J., Glaziou, P., Lönnroth, K., Siroka, A., Floyd, K., Weil, D., et al. (2018). The impact of social protection and poverty elimination on global tuberculosis incidence: A statistical modelling analysis of Sustainable Development Goal 1. *The Lancet Global Health, 6*(5), e514.

Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International Journal of Environmental Research and Public Health, 15*(8), 1596.

Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals, 135*, 109850.

Chakraborty, T., Chattopadhyay, S., & Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications, 527*, 121266.

Chuang, Y. C., Huang, Y. L., Tseng, K. C., Yen, C. H., & Yang, L. H. (2015). Social capital and health-protective behavior intentions in an influenza pandemic. *PloS One, 10*(4), e0122970.

Conti, P., & Younes, A. (2020). Coronavirus COV-19/SARS-CoV-2 affects women less than men: Clinical response to viral infection. *Journal of Biological Regulators and Homeostatic Agents, 34*(2), 339.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In *International conference on machine learning* (pp. 933–941). PMLR.

Desai, D. (2020). Urban densities and the Covid-19 pandemic: Upending the sustainability myth of global megacities. *ORF Occasional Paper, 244*(4), 1.

Ding, G., Li, X., Shen, Y., & Fan, J. (2020). Brief Analysis of the ARIMA model on the COVID-19 in Italy. medRxiv.

Durbin, J. (1959). Efficient estimation of parameters in moving-average models. *Biometrika, 46*(3/4), 306.

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.

Koh, H. K., & Cadigan, R. O. (2008). Disaster preparedness and social capital. In *Social capital and health* (pp. 273–285). Springer.

Kovacevic, M., & Jahic, A. (2020). COVID-19 and human development. Exploring global preparedness and vulnerability. Human Development Report Office. New York: United Nations Development Programme.

Lai, C. C., Wang, C. Y., Wang, Y. H., Hsueh, S. C., Ko, W. C., & Hsueh, P. R. (2020). Global epidemiology of coronavirus disease 2019 (COVID-19): Disease incidence, daily cumulative index,
mortality, and their association with country healthcare resources and economic status. *International Journal of Antimicrobial Agents, 55*(4), 105946.

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals, p.* 110059.

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2019). Temporal fusion transformers for interpretable multi-horizon time series forecasting. arXiv preprint arXiv:1912.09363.

Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., et al. (2020). Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling, 5*, 282.

Marathe, M., & Vullikanti, A. K. S. (2013). Computational epidemiology. *Communications of the ACM, 56*(7), 88.

Ning, W., Lei, S., Yang, J., Cao, Y., Jiang, P., Yang, Q., et al. (2020). Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. *Nature Biomedical Engineering, 4*(12), 1197.

Panwar, H., Gupta, P., Siddiqui, M. K., Morales-Menendez, R., & Singh, V. (2020). Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos, Solitons & Fractals, 138*, 109944.

Redding, D. W., Atkinson, P. M., Cunningham, A. A., Iacono, G. L., Moses, L. M., Wood, J. L., & Jones, K. E. (2019). Impacts of environmental and socio-economic factors on emergence and epidemic potential of Ebola in Africa. *Nature Communications, 10*(1), 1.

Reddy, C. K., & Vinzamuri, B. (2013). A Survey of Partitional and Hierarchical Clustering Algorithms. *Data clustering: Algorithms and applications, 87*, 6039.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals, 135*, 109853.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53.

Roy, S., Bhunia, G. S., & Shit, P. K. (2021). Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Modeling Earth Systems and Environment, 7*(2), 1385.

Siettos, C. I., & Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence, 4*(4), 295.

Soliman, M., Lyubchich, V., & Gel, Y. R. (2019). Complementing the power of deep learning with statistical model fusion: Probabilistic forecasting of influenza in Dallas County, Texas, USA. *Epidemics, 28*, 100345.

Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv preprint arXiv:2004.07859.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37.

Velásquez, R. M. A., & Lara, J. V. M. (2020). Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. *Chaos, Solitons & Fractals, 136*, 109924.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236.

Xue, H., Bai, Y., Hu, H., & Liang, H. (2018). Influenza activity surveillance based on multiple regression model and artificial neural network. *IEEE Access, 6*, 563.

Xu, Q., Gel, Y. R., Ramirez Ramirez, L. L., Nezafati, K., Zhang, Q., & Tsui, K. L. (2017). Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PloS One, 12*(5), e0176690.

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

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence, 2*(5), 283.

Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., et al. (2020). Predicting COVID-19 in China using hybrid AI model. *IEEE Transactions on Cybernetics, 50*(7), 2891.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Authors and Affiliations

Srinka Basu\textsuperscript{1} \cdot Sugata Sen\textsuperscript{2}

\begin{itemize}
\item Sugata Sen
  sensugata@gmail.com
\item Srinka Basu
  srinka.basu@gmail.com
\end{itemize}

\textsuperscript{1} Department of Engineering and Technological Studies, University of Kalyani, Kalyani, West Bengal 741235, India

\textsuperscript{2} Department of Economics, Panskura Banamali College (Autonomous), Panskura, Purba Medinipur, West Bengal 721152, India