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.
Discovering optimal strategies for mitigating COVID-19 spread using machine learning: Experience from Asia

Yue Pan a,b, Limao Zhang a,*, Zhenzhen Yan c, May O. Lwin d, Miroslaw J. Skibniewski e, f, g

a School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
b Shanghai Key Laboratory for Digital Maintenance of Buildings and Infrastructure, Department of Civil Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China
c School of Physical and Mathematical Sciences, Nanyang Technological University, 1, Nanyang Link, Singapore 637371, Singapore
d Wee Kim Wee School of Communication and Information, Nanyang Technological University, 3, Nanyang Link, WKWSCI Bldg, 637718, Singapore
e Department of Civil and Environmental Engineering, University of Maryland, College Park, 9 MD 20742-3021, USA
f Chaoyang University of Technology, Taichung 4413310, Taiwan
g Polish Academy of Sciences, Institute for Theoretical and Applied Informatics, 44-100 Gliwice, Poland

ARTICLE INFO

Keywords:
COVID-19
Random forest regression
Feature importance analysis
Multi-objective optimization
Sustainability

ABSTRACT

To inform data-driven decisions in fighting the global pandemic caused by COVID-19, this research develops a spatiotemporal analysis framework under the combination of an ensemble model (random forest regression) and a multi-objective optimization algorithm (NSGA-II). It has been verified for four Asian countries, including Japan, South Korea, Pakistan, and Nepal. Accordingly, we can gain some valuable experience to better understand the disease evolution, forecast the prevalence of the disease, which can provide sustainable evidence to guide further intervention and management. Random forest with a proper rolling time-window can learn the combined effects of environmental and social factors to accurately predict the daily growth of confirmed cases and daily death rate on a national scale, which is followed by NSGA-II to find a range of Pareto optimal solutions for ensuring the minimization of the infection rate and mortality at the same time. Experimental results demonstrate that the predictive model can alert the local government in advance, allowing the accused time to put forward relevant measures. The temperature in the category of environment and the stringency index belonging to the social factor are identified as the top 2 important features to exert a greater impact on the virus transmission. Moreover, optimal solutions provide references to design the best control strategies towards pandemic containment and prevention that can accommodate the country-specific circumstance, which are possible to decrease the two objectives by more than 95%. In particular, appropriate adjustment of social-related features needs to take priority over others, since it can bring about at least 1.47% average improvement of two objectives compared to environmental factors.

1. Introduction

Since the first confirmed case of the novel coronavirus disease 2019 (COVID-19) reported at the end of December 2019 in Wuhan, Hubei province, China, COVID-19 has rapidly spread across more than 200 countries and territories of the world. The severe pneumonia outbreak has triggered a public health emergency of international concern, which has been announced as a global pandemic by the World Health Organization (WHO) on March 11, 2020. According to open data from the Johns Hopkins University tracker, the total confirmed cases and reported deaths globally have reached over 82 million and 1.8 million up to 31st December 2020. Notably, the genome and symptoms of COVID-19 show apparent similarity to the severe acute respiratory syndrome coronavirus (SARS-CoV) that occurred during 2002–2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) firstly appeared in 2012 (Ali & Alharbi, 2020). In contrast to SARS and MERS, COVID-19 demonstrates a lower case-fatality rate and stronger transmission capability, resulting in a larger number of infections worldwide. The mutation in the virus has led to changes in viruses’ ability to cause the pandemic, which is estimated to be more transmissible than the
original version. As the epidemic continues progressing, it negatively affects the resilience of the global society from every aspect of daily life, environment, economy, and others, and thus it raises urgent attention from policymakers and health planners internationally. In order to address the unprecedented challenge, the scientific evidence can be derived from the deep exploration of the dynamic evolution of COVID-19 transmission. Accordingly, potential strategies and interventions can be formulated at an early stage for limiting or even blocking the sustained propagation, contributing to reducing the infectious and mortality rates.

Apart from existing bioinformatics and medical methodologies about virus mechanism study, data science as another feasible solution has shown its advantages of battling against the pandemic (Barda, Riesel et al., 2020; Gao, Cai et al., 2020). In this regard, a novel and convincing way is to continuously investigate the impacts of important factors especially concerning environmental and social aspects on the spread of COVID-19, resulting in data-oriented modeling and prediction to make treatment or prevention decisions (Lim, Yun et al., 2021; Lwin, Lu et al., 2020). For one thing, environmental dynamics, like temperature, humidity, atmospheric pressure, wind speed, meteorological variability, air quality, and others, have been regarded as the top predictors of coronavirus illnesses (Das, Ghosh et al., 2021). They possibly play critical roles in influencing the viability, transmission, and mortality of COVID-19 (Bashir, Ma et al., 2020). As expected, the actionable environmentally-informed COVID-19 risk monitoring and prediction can meet its great potential in supporting political agendas for pandemic comprehension and response (Zaitchik, Sweijd et al., 2020). For another, when no reliable vaccines and antiviral medication is available, a feasible response plan for curbing the spread of the virus relies on a suitable combination of various government intervention strategies, such as home quarantines, social distance measures, personal protective measures, and others, which could impinge on people’s behavior and daily life, including work mode, travel distance, and engagement in social activities (Haug, Geyerhofer, Londle, & Dervic, 2020). It has been found that the speed of the region to act on outbreaks and the reduction in travel is crucial to eliminate the infectivity, transmissibility resurgence of COVID-19 (Siegenfeld & Bar-Yam, 2020). Besides, strategies based on social distancing and ventilation have been emphasized broadly in limiting the spread and risk of COVID-19 (Sun & Zhai, 2020). In short, the dynamic of the COVID-19 pandemic is highly associated with multiple factors from the environmental and social aspects. Most of the previous studies have paid a lot of attention to climate change, especially for the temperature and humidity, while the impact of other environmental and social factors has not been fully explored (Nakada & Urban, 2020). In particular, Hu et al. (2021) indicated that social environmental factors could be strongly correlated with COVID-19 death. Due to the interrelated nature of the pandemic situation with various factors, additional factors beyond the climate aspects should be considered and examined systematically, aiming to more reliably track the evolution and forecast the future trend of virus spread. Therefore, a question can be raised on how to create a realistic multi-variable prediction model to accurately estimate and understand the daily infection rate and death due to COVID-19 at the nationwide scale, in order to potentially facilitate responsive and preventive actions ahead of time.

For tracking and predictive purposes, the susceptible–infectious–recovered (SIR) epidemiological model is the most basic and intuitive tool applied to describe the dynamic evolution of virus transmission across populations. Although it is relatively easy to understand and implement only relying on three components of susceptible (S), infected (I), and removed (R) individuals, its prediction performance is limited by the assumptions, such as the well-mixed population, homogeneity of the population, exponentially distributed duration of infection, large population (Huppert & Katriel, 2013). In other words, the SIR model has difficulty in adapting to changeable external conditions, which is insufficient to handle the COVID-19 epidemic/pandemic development in the nature of complexity and nonlinearity. For this reason, machine learning involving complex mathematical calculations has been employed as a reliable alternative to provide more accurate time-series forecasting, which can seek nonlinear relations in a set of input variables to discover the hidden patterns by learning the increasing availability of historical data (Pan & Zhang, 2021). It is believed that big data practice based on machine learning models is helpful to intelligently and dynamically support decision-making for fighting against COVID-19 towards compiling the sustainable development goal (Rahman, Zaman et al., 2020). Recent research reveals that machine learning-based methods can make full use of spatio-temporal data to achieve more promising results with the superiority of adaptive learning, trend-based recalibration, and great flexibility, contributing to fully understanding the current situations and predicting future trends about COVID-19 (Bansal, Padappayil et al., 2020). For example, Yesilkacan (2020) adopted random forest to predict the daily cases of COVID-19 reaching R2 above 0.845, which was proven useful in estimating and map the spatio-temporal distribution of COVID-19 outbreak. Malki et al. (2020) built several state-of-the art machine learning models, including k neighbors regressor, random forest, decision tree, support vector machine, and others, to learn features about climatic conditions for studying the spreading rate of COVID-19, which validated that weather variables were more relevant to the accurate prediction of mortality rate and thus these variables required more attention. Zhan et al. (2021) developed a hybrid machine learning model of random-forest-bagging broad learning system to forecast the COVID-19 trend and compare the prediction performance with other popular than machine learning algorithms, which verified that random forest had great robustness and was flexible in feature importance analysis. According to the above-mentioned studies, it is worth noting that an ensemble learning-based model named random forest composed of several decision trees has shown outstanding performance in estimating the daily number of confirmed COVID-19 cases.

From the advent of COVID-19, most studies mainly put efforts on building the intelligent model for predicting the pandemic trend and investigating the contribution of different factors to the virus spread for a single country or a state (Maiti, Zhang et al., 2021; Mansour, Al Kindi et al., 2021; Sannigrahi, Pilla et al., 2020). However, a challenge arises that it is impossible to directly inform preparedness response strategies under the combination of ideal socio-environmental conditions from the prediction. As reviewed, there has been limited research on applying optimization algorithms to the constructed machine learning model that provides the relationship between inputs and outputs. The possible cause might be the considerably high degree of complexity and uncertainty arising from multiple influential factors and their complicated impact on the transmission dynamics of COVID-19. It is observed that simulation of the dynamic pandemic behavior is nonlinear and time-variant in nature, which can be regarded as a multi-objective prediction and optimization problem (Thakur, 2020). Therefore, another research gap to be narrowed is to seek the optimal setting of environmental and social factors for depressing together the growth of the confirmed cases and deaths as much as possible in the targeted region. For this concern, we intend to incorporate the random forest regression model with certain optimization algorithms in the area of single or multicriteria decision-making. To be more specific, single-objective optimization determines the best solution to reach the maximum/minimum of the single objective function, while the multi-objective optimization problem for the task is to find the set of optimal trade-off solutions to handle conflicting objectives simultaneously. As a result, knowledge from the optimum solutions offers clear insights into the complex nature of the COVID-19 outbreak, which provides decision-makers with the profile of the influential factors that can satisfy the desired requirements. Accordingly, mixed and suitable control measures can be designed to combat the pandemic.

Since the COVID-19 pandemic has placed humanity at extremely high health and social risk, this research aims to seek novel solutions
based on data science in pursuit of reducing the COVID-19 severity and dissemination. It paves a new way to control the ongoing pandemic COVID-19 for promoting sustainable societies. We concentrate on the following two specific objectives: one is from the view of prediction by developing a machine learning-based prediction model with the proper lead-time to forecast the daily growth rate and mortality rate of COVID-19 towards possible circumstances in the next few days; the other is from the view of optimization by optimizing the established model for the determination of optimal control and preventive strategies. That is to say, the research offers an opportunity to fully mine the valuable information behind the large amount of accumulated data related to environmental and social dimensions. Results from the spatiotemporal analysis are expected to provide insights to policymakers for putting forward early preparations and preventive measures. In this regard, there are three main research questions to be answered: (1) How to build the machine learning model to characterize the dynamics of COVID-19 cases and return convincing predictions about its evolution trend; (2) How to explain the tree-based predictor and precisely measure the critical roles of various factors in impacting the disease prevalence; (3) How to perform optimization to reasonably reveal the optimal environmental and social conditions for minimizing the growth of confirmed cases and deaths.

2. Data and method

2.1. Framework of the proposed method

For two goals of prediction and prevention, the proposed approach is executed based on three major steps, including dataset preparation, COVID-19 predictor creation, and optimal strategy formulation, which has been visualized in Fig. 1. At first, a high-quality dataset is carefully prepared, which incorporates a large number of continuous features from several public datasets. The scope of features that we have employed is broad, covering the environmental and social categories to describe the spatial dynamic of COVID-19. Secondly, random forest regression is trained on the prepared dataset to model the non-linear relationship between the input influential factors and the targeted objective, aiming to conduct the time-series prediction. As a result, it can return the promising estimation of the daily increase about the COVID-19 confirmed cases and deaths within 95% CI at the country level. In particular, hyper-parameters of the model are fine-tuned by the PSO algorithm to ensure high prediction accuracy. For better interpretability of the random forest model, MDI and SHAP-based feature important measures are calculated to numerically infer the effects of socio-environmental factors on the outcomes. Thirdly, a multi-objective optimization algorithm is applied to the prediction model to minimize two objectives about the growth rate and death rate simultaneously. As a result, it can yield Pareto fronts as the optimal solution for determining the effective control strategies under ideal settings of decision variables. Several single objective optimization algorithms are also carried out to minimize each objective separately. In the end, the full understanding of the data analysis results assists governments in decision-making efforts to reduce COVID-19 risk for harmonizing the goal of sustainable development, which can decrease dependence on subjective judgment.

2.2. Data preparation

To begin with, the preparation of the machine model training is to collect a comprehensive dataset. As reviewed, the dependency of COVID-19 on environmental and social factors has been verified. In order to gain a more full understanding of the epidemiology of COVID-19, we can therefore rely on the daily growth of confirmed cases and deaths along with related factors potentially impacting the virus spread. All the desired information is available on the following four open sources: (1) The confirmed and death cases per day in each country are continuously recorded by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Hopkins, 2020). (2) The environmental data, including the air pollutant species and meteorological characteristics, is collected from the world air quality index project (WAQI, 2020). (3) Social Distancing Stringency Index and Policy
Indicators are obtained via the Oxford Covid Government Response Tracker (OxCGRT) (Hale, Webster et al., 2020), which aggregate scores to quantify the extent of government responses. (4) The travel popularity data is summarized in the Google Community Mobility Reports (Google 2020), as a reflection of the citizens’ movement trends over time by geography. In subsequence, data from the above-mentioned disparate sources can be systematically merged into a single repository. Since it is found that multiple data sources in combination can demonstrate a strong ability in representing an event or domain of interest (Samuelsen, Chen et al., 2019), our prepared dataset with data integration will have implications for predicting the potential cases and explore the influence from the socio-environmental aspects. In addition, we focus on the COVID-19 transmission country-wise, including Japan, Korea, Pakistan, and Nepal, during the period of Feb 15, 2020 – Nov 27, 2020. Detailed description of the prepared dataset has been provided in Tables 1 and 2.

### 2.3. Random forest model development and evaluation

Due to the powerful capability of returning considerably accurate and robust predictions, a machine learning algorithm named random forest (RF) (Breiman, 2001) is adopted in this research for establishing a non-linear regression model. Random forest, known as an ensemble of multiple base estimators by averaging. Each decision tree, occurs with the combination of multiple weak decision trees that are easily implemented and interpreted (Pan, Zhang et al., 2020). That is to say, RF with the notion of an ensemble forms a strong learner to demonstrate a much better approximation to the ground truth. To be more specific, RF begins with the bootstrap resampling technique, which makes random sampling from the original training data. After the random sample with replacement, subsets of the training data in the same size are generated. It should be noted that around 36.8% of the data is left out for the validation purpose in one bootstrap resampling, which is called out-of-bag (OOB) samples. Then, each bootstrapped training set is responsible for building an unpruned decision tree, which learns a limited number of features that are randomly selected from all alternatives. Finally, the overall prediction of random forest can be determined based on the aggregation of results from multiple base estimators by averaging.

Notably, hyper-parameter tuning can be the key to build an optimal random forest model with the goal of minimizing the error difference between the actual and predicted, which exerts a direct effect on the model performance. However, the task of hyper-parameter optimization is not straightforward, since parameters cannot be directly learned from data. To address the issue, we carry out a prevalent optimization technique named particle metaheuristic algorithm (PSO) in 3-fold cross-validation, and thus the hyper-parameters of random forest can be determined in an automatic process. To ensure a promising performance, the hyper-parameters of the random forest, which are unable to be directly learned from data, are automatically determined by a prevalent optimization technique named particle metaheuristic algorithm (PSO) in 3-fold cross-validation. PSO in the family of the evolutionary algorithm has been proven useful in efficiently finding near-optimal or optimal solutions in large configuration space in a set of public datasets (Yang & Shami, 2020). In short, the superiority of PSO lies in its fast convergence rate and easy implementation under relatively low computational complexity. A brief introduction of PSO is given in the part of single-objective and multi-objective optimization. Herein, four important hyper-parameters are chosen to be tuned for well-fitting the random forest model into the prepared COVID-19 dataset, including the number of trees in the forest, the maximum depth of the tree to grow, the optimum number of data to split an internal node, and the minimum number of data to gain a leaf node. In the end, three common metrics in Eq. (1) – (3), namely the mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are adopted to figure out the optimal hyper-parameter configuration and evaluate the model performance.

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (y_{p,i} - y_{a,i})^2
\]

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_{p,i} - y_{a,i}|
\]

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_{p,i} - y_{a,i}}{y_{a,i}} \right|
\]

where \( N \) is the total number of data points, \( y_{p,i} \) and \( y_{a,i} \) are the predictive values.

### Table 1

Characteristics of features in Japan and Korea datasets.

| Category          | Feature description                                      | Variable | Japan      | South Korea |
|-------------------|----------------------------------------------------------|----------|------------|-------------|
|                   |                                                          |          | Mean (SD)  | Mean (SD)   | P-value     |
|                   |                                                          |          | [Min, Max] | [Min, Max]  |             |
| Environment       | Carbon monoxide (CO) (mg/m$^3$)                          | x1       | 2.99 (0.50) | [1.95, 4.58]| 4.83 (0.72)| [3.03, 7.63]| <0.001 |
|                   | Relative humidity (%)                                     | x2       | 70.40 (10.80)| [24.8, 91.58]| 72.08 (15.22)| [38.17, 97.67]| 0.021 |
|                   | Nitrogen dioxide (NO$_2$) (μg/m$^3$)                    | x3       | 6.25 (2.09) | [3.04, 16.15]| 13.15 (5.08) | [5.32, 34.60]| <0.001 |
|                   | Ozone (O$_3$) (μg/m$^3$)                                 | x4       | 25.25 (7.61) | [8.98, 51.44]| 24.80 (10.44) | [5.75, 49.15]| 0.014 |
|                   | Particulate matter with aerodynamic diameter ≤10 μm (PM$_{10}$) (μg/m$^3$) | x5       | 13.39 (5.16) | [6.29, 39.48]| 27.57 (10.44) | [6.17, 55.00]| <0.001 |
|                   | Particulate matter with aerodynamic diameter ≤2.5 μm (PM$_{2.5}$) (μg/m$^3$) | x6       | 38.11 (11.91) | [18.23, 84.35]| 56.61 (18.89) | [13.00, 115.33]| <0.001 |
| Pressure (hPa)    |                                                          | x7       | 1008.66    | [993.98, 1032.90] | 1013.87     | [991.00, 1032.00]| <0.001 |
|                   |                                                          |          | (5.93)     | (7.17)      |             |             |         |
| Sulfur dioxide (SO$_2$) (μg/m$^3$) | x8          | 2.51 (0.41) | [1.83, 5.16] | 4.72 (0.50) | [3.60, 6.40] | <0.001 |
| Temperature (°C)  |                                                          | x9       | 18.77 (6.39) | [4.27, 29.65] | 16.97 (6.49) | [1.03, 28.48] | <0.001 |
| Wind gust (m/s)   |                                                          | x10      | 6.33 (2.32) | [2.47, 15.90] | 9.32 (1.45) | [1.50, 15.40] | <0.001 |
| Wind speed (m/s)  |                                                          | x11      | 2.77 (0.75) | [1.58, 8.47]  | 2.40 (0.86) | [1.08, 6.32] | <0.001 |
| Social            | Stringency index                                         | x12      | 36.05 (6.75) | [19.44, 47.22] | 51.73 (10.51) | [31.48, 82.41] | <0.001 |
|                   | Change of people in retail and recreation (%)            | x13      | –13.23     | [–48, 21]   | –11.72     | [–45, 15]   | 0.131 |
|                   |                                                            |          | (10.97)    |             |             |             |         |
|                   | Change of people in grocery and pharmacy (%)            | x14      | 0.24 (4.08) | [–15, 14]   | 6.06 (10.00) | [–18, 73]   | <0.001 |
|                   |                                                            |          | (10.97)    |             |             |             |         |
|                   | Change of people in parks (%)                            | x15      | –1.25 (16.50) | [–52, 76]   | 30.79 (27.71) | [–38, 158] | <0.001 |
|                   |                                                            |          | (12.75)    |             |             |             |         |
|                   | Change of people in transit stations (%)                 | x16      | –25.67     | [–69, 2]    | –8.55 (9.24) | [–36, 16] | <0.001 |
|                   |                                                            |          | (12.75)    |             |             |             |         |
|                   | Change of people in workplaces (%)                       | x17      | –14.49     | [–73, 4]    | –6.89      | [–81, 3] | <0.001 |
|                   |                                                            |          | (11.16)    |             |             |             |         |
|                   | Change of people in residential (%)                      | x18      | 7.12 (4.73) | [0, 30]     | 3.84 (3.48) | [–2, 17] | <0.001 |

Note: SD, min, max are the abbreviation of the standard deviation, the minimum, and the maximum, respectively. P-value for the 18 continuous variables is calculated by the Wilcoxon signed ranks test.
value and actual value, respectively. The smaller value of MSE, MAE, and MAPE indicates the better prediction performance.

### 2.4. Feature importance measurement

Based upon the machine learning model, more investigation is required to examine the cause-effect between the COVID-19 spread and a series of factors to consider. For this purpose, two methods of feature importance calculation are deployed in this research to assign the importance scores to input features for quantifying their contribution to the model’s decision. The first one is a popular tree feature attribution approach known as mean decrease impurity (MDI), which is a built-in important score in the scikit-learn package for the random forest model. MDI directly measures the importance of a feature to the tree that is measured by the variance reduction using MSE. If a variable assumes that it acts as an important feature to define the predictions. As an efficient solution, a modified algorithm for fast SHAP value estimation has been introduced in the paper (Lundberg, Erion et al., 2018) to exponentially lower the computational complexity of Eq. (5) from \( O(TL2M^T) \) to \( O(TLDT) \), where \( T \) is the number of trees in the tree ensemble model, \( L \) is the maximum number of leaves in a tree, \( M \) is the number of features, and \( D \) is the maximum depth of a tree. For another, the SHAP interaction index, as an extension of SHAP values, can directly capture the pairwise interaction effects between local features on the model’s prediction, which can be obtained from Eq. (6) to generate a matrix of feature attribution (Lundberg, Erion et al., 2020). This can be explained as the difference between the SHAP value of the \( ij \)-th feature when the \( j \)-th feature is presented or not.

\[
\phi(f, x) = \sum_{i \in \mathcal{X}(f)} \left[ \frac{\left| \left( S \setminus \{i\} \right) \cdot f_i(S) - \left( S \cup \{i\} \right) \cdot f_i(S) \right|}{\sum_{j \in \mathcal{X}(f)} \left| \left( S \setminus \{i\} \right) \cdot f_j(S) - \left( S \cup \{j\} \right) \cdot f_j(S) \right|} \right], \text{ when } i \neq j
\]

where \( \mathcal{X}(f) \) denotes the subsets of all \( M \) input features. It should be noted that \( \phi(f, x) = \phi_0(f, x) \), since the SHAP interaction index of the \( ij \)-th feature can be equally split between each feature.

### 2.5. Single-objective and multi-objective optimization

Notably, the determination of optimal COVID-19 control solutions in a data-driven manner can be considered as a single or multi-objective optimization problem. The ultimate goal of proper optimization algorithms is minimizing the rate of increase in COVID-19 confirmed cases or deaths or both of them. In particular, these targeted objectives are closely associated with various influential features, and their complicated relationships have been mathematically modeled by machine learning. In consequence, a reasonable profile of variables of interest can be provided, which is believed to benefit epidemic management and prevention.
For the single-objective optimization, three typical methods belonging to the family of meta-heuristic optimization techniques are deployed in this research, namely genetic algorithm (GA), particle metaheuristic algorithm (PSO), and biased random key genetic algorithm (BRKGA). They all own the powerful ability of performing a global search over the promising regions to address non-convex problems (Pan & Zhang, 2021; Yan, Gao et al., 2018). The main procedure of the three algorithms is presented below: (1) In GA (Gen & Lin, 2007), a fitness function, which is the objective function defined by the random forest regression model herein, is employed to assess the performance of individuals in the current generation. The desired individuals that are deemed to have better survival capability and adaptability will have more chances to survive and produce the next generation based on operators of selection, crossover, and mutation. The above-mentioned process will repeat several times until the termination criterion is satisfied. As a result, the best individual which GA is convergent to symbolizes the global optima. (2) In BRKGA (Gonçalves & Resende, 2011), a solution of the combinatorial optimization problem is encoded by a vector of randomly generated numbers named the random keys in the range [0,1]. To take the place of evolutionary operators, a decoder is designed to take the random keys as inputs to produce a discrete solution. Solutions are ranked by the fitness calculation and partitioned into the elite and non-elite ones. Later, one parent is always chosen from the elite solutions, while the other is selected from the non-elite. Based on the parameterized uniform crossover, the two parent solutions can be combined for generating the offspring. (3) In PSO (Kennedy & Eberhart, 1995), its obvious distinction from GA is that it does not implement any evolutionary operators termed crossover and mutation. PSO carries out the search using a swarm of particles. The velocity of a particle is regarded as the potential solution, which is greatly influenced by the personnel best and global best locations. At each iteration, the velocity and the position of the particle can be updated by Eqs. (8) and (9) to drive the particle towards the desired locations.

\[
\begin{align*}
    v_{id}(t+1) &= wv_{id}(t) + c_1r_1(p_{id}(t) - x_{id}(t)) + c_2r_2(g_{id}(t) - x_{id}(t)) \\
    x_{id}(t+1) &= x_{id}(t) + v_{id}(t+1)
\end{align*}
\]

where \(w\) is the inertia weight, \(v_{id}(t)\) and \(x_{id}(t)\) are the \(d\)th coordinates of \(t\)th particle’s velocity and position at the \(t\)th iteration, respectively, \(p_{id}(t)\) and \(g_{id}(t)\) are the \(d\)th coordinates of \(i\)th particle’s local best and global best value. \(c_1\) can \(c_2\) are constant to balance exploiting \(p_{id}(t)\) and \(g_{id}(t)\). \(r_1\) and \(r_2\) are two random values between 0 and 1.

For the multi-objective optimization, a popular evolutionary algorithm called the non-dominated sorting genetic algorithm (NSGA-II) is adopted, which obeys the general procedure of GA with a modified mating and survival selection. There are three special characteristics involved in NSGA-II, namely fast non-dominated sorting approach, fast crowded distance estimation procedure, and simple crowded comparison operator, which are conducive to guarantee computational efficiency and population diversity (Deb, Pratap et al., 2002). The general steps of NSGA-II are given below (Yusoff, Ngadiman et al., 2011). To begin with, individuals in the population are sorted according to the non-domination criteria and assigned the crowding distance value. For producing the next generation, three critical operators are applied, which are the binary tournament selection under the crowding comparison-operator, simulated binary crossover, and polynomial mutation. Next, a combination of the obtained offspring population and the current generation is created to ensure elitism, from which several solutions can be chosen by the non-dominated sorting to form the next generation. In other words, the next generation is continuously filled by each front until its size is larger than the current population size. After several iterations, the last generation can approximate the Pareto front on behalf of a representative set of trade-off optimal solutions to the defined problem.

3. Results

3.1. Baseline characteristics of the prepared dataset

The COVID-19 pandemic has brought unprecedented health and economic crisis to the world, which unfortunately weakens the sustainability and resilience in both developed and developing countries. To make a comprehensive prediction and optimization in this research, we obtain access to COVID-19-related data of two developed countries in East Asia and two developing countries in South Asia (Pakistan and Nepal) from February 15, 2020 to November 27, 2020. Thus, there are four prepared datasets for the four targeted countries, and each dataset owns 267 lines of valid records to describe the condition during the study period. For one thing, Japan and South Korea are taken as two typical developed countries, which are the first ones to confront the novel coronavirus beyond China. Both of them have taken suitable measures and policies to well control the disease and prevent the massive outbreaks that have hit the United States and some parts of Europe. Lessons learned from them are possible to lead the global recovery efforts. For another, Pakistan and Nepal are two representatives of developing countries under relatively poor economic conditions. The location of the four countries has been visualized in the map of Fig. 2a, along with their accumulated number of COVID-19 confirmed cases and death cases by November 27, 2020. Observably, Pakistan and Nepal have reported the largest number of confirmed cases, which are around 11.76 and 6.8 times more than South Korea. To some extent, the COVID-19 outbreak tends to put the relatively less developed countries at higher health risk mainly due to the lack of sufficient economic resources, advanced medical technology, and far-reaching government responses for risk mitigation.

Different kinds of features affecting the COVID-19 spread countrywise are gathered mainly from the environmental and social perspectives, as listed in Tables 1 and 2. To be more specific, environmental features are composed of the meteorological variables (i.e., the daily average of relative humidity, pressure, temperature, wind gust, and wind speed) and air pollutant variables (i.e., mean concentrations of CO, NO\(_2\), O\(_3\), PM\(_{10}\), PM\(_{2.5}\), and SO\(_2\)). Social features also incorporate two categories of information. One is the stringency index to quantify the level of government responses between 0 (low strictness) to 100 (high strictness). The other is the Google mobility index in the form of a positive or negative percentage to reflect changes in people’s movement affected by the governmental policy implementation, which is measured by comparing with the medium value from the period of January 3, 2020 – February 6, 2020 as the baseline. Due to data availability, datasets of Japan and Korea own 18 features in total, which has 6 more features than the resting datasets for Pakistan and Nepal. The difference lies in the additional 5 features about the concentration of air pollutants (including CO, NO\(_2\), O\(_3\), PM\(_{10}\), PM\(_{2.5}\), and SO\(_2\)) and 1 feature regarding wind gust.

To provide insights into conditions of the environment, social, and pandemic evolution across four targeted countries, Tables 1 and 2 outline the statistical characteristics of features in terms of mean value, standard deviation, and upper and lower bound. All the prepared features are in the numerical type, which are valuable information for describing the patterns in pandemic evolution over time. Besides, the Wilcoxon rank-sum test, a nonparametric version of the two-sample \(t\)-test under the assumption of independence and equal variance in two populations, is carried out in Tables 1 and 2 to statistically examine the differences from the observations between Japan and South Korea, Pakistan and Nepal. Almost all the \(P\)-value is less than 0.05, except for the variable x13 in Table 1. It means that although the pairs of countries are proximal to each other, there is a significant difference in both the environmental and social aspects, which can be highlighted as the major contributors to the varying degree of COVID-19 negative effects across countries. The selected features are weakly correlational to the output variables calculated by the Pearson correlation coefficient in Supplementary Figure 1. Later, the random forest can flexibly learn these
prepared data to fit different functions for different countries. All these features will act as the decision variables in the single-objective or multi-objective optimization task.

### 3.2. Random forest model performance

For the purpose of predictive controlling and prevention, these influential features described in Tables 1 and 2 can then be fed into the appropriate machine learning model to return reliable predictions about the future trend. Herein, two random forest regression models are built for each country to capture the complex dynamics of COVID-19 spread, resulting in accurate predictions on the daily growth rate of COVID-19 confirmed cases and death cases. The hyper-parameters of models are tuned by the optimization technique called PSO in 3-fold cross-validation to further guarantee the promising and stabilize model performance. The developed models optimized by PSO can always quickly reach the slightly smaller MSE and MAE than the traditional grid search method (Supplementary Table 1), and thus PSO is a reasonable choice for fine-tuning parameters of models in this research.

Besides, it is worth noting that a rolling time-window in the size of $n$ days is designed in random forest regression to model and predict the COVID-19 evolution in Japan and South Korea gains a pretty clear agreement between the ground truth and the predicted value in all eight figures. The great discrepancy is more likely to appear in the first fifty days is designed in random forest regression to model and predict the weekly trend. COVID-19 transmission during this period. On the contrary, the remaining days see a great fitting and the predicted rate can even be slightly higher than the actual data to manifest a relatively conservative estimate of the model. Table 3 measured the performance of the model with the optimal parameters by MSE, MAE, and MAPE. It turns out that the generated predictions. The defined range of values supports a reliable estimation of uncertainty from the random forest regression. The model performance after 3-fold cross-validation is demonstrated in Table 3 lists the most optimal set of hyper-parameter configurations and the rolling time-window, which are adopted to build the best predictive models for learning the collected data. Also, we have made attempts to compute variance and find a 95% confidence interval (CI) of the generated predictions. The defined range of values supports a reliable estimation of uncertainty from the random forest regression. The model performance after 3-fold cross-validation is demonstrated in Fig. 2(b) – (i), where the predicted value under 95 CI is plotted with the actual trend reflecting the daily increase rate of confirmed cases and deaths for a clear comparison. Observably, there is a high level of agreement between the ground truth and the predicted value in all eight figures. The great discrepancy is more likely to appear in the first fifty days in Japan, Korea, and Pakistan and the second fifty-day interval in Nepal mainly due to the unpredictably high and sudden increase of COVID-19 transmission during this period. On the contrary, the remaining days see a great fitting and the predicted rate can even be slightly higher than the actual data to manifest a relatively conservative estimate of the model. Table 3 measured the performance of the model with the optimal parameters by MSE, MAE, and MAPE. It turns out that MSE and MAE are small enough in all conditions, highlighting the great reliability and generalization of the developed model. For the two prediction purposes, models in the Japan dataset achieve the most accurate results under the lowest MSE of $3.51 \times 10^{-4}$ (95% CI: $2.41 \times 10^{-4}$ – $4.62 \times 10^{-4}$) and $2.12 \times 10^{-3}$ (95% CI: $6.67 \times 10^{-3}$ – $3.56 \times 10^{-3}$), and the lowest MAE of $2.12 \times 10^{-3}$ (95% CI: $9.45 \times 10^{-3}$ – $1.30 \times 10^{-2}$) and $1.89 \times 10^{-2}$ (95% CI: $1.39 \times 10^{-2}$ – $2.38 \times 10^{-2}$), probably because of the smaller outlier value in Japan than other countries. Through comparison with other relevant studies, it can be found that the prediction performance of our model is
better. In terms of MAPE value for model evaluation, experiment result based on the random forest from Malki, Atlam et al., (2020) for predicting confirmed cases globally was 3.313, and the improved adaptive network-based fuzzy inference system (CESBAS-ANFIS) from Zivkovic, Bacanin et al., (2021) for predicting confirmed cases in China was 4.08. As for our established model in this case, the MAPE value is in the range of 0.593 and 2.416, which is significantly lower than the above-mentioned studies. Particularly, Japan and Nepal demonstrate a higher prediction accuracy, whose MAPE value for the growth rate and death rate can be all below 1.0. Overall, both Fig. 2 and Table 3 reveal the success of the random forest-based model in forecasting the growth of infections and mortality in the near future under acceptable prediction error, which can play a valuable role in perceiving the potential risk at an early phase. To further validate the prediction performance of random forest, we compare it with another popular machine learning algorithm named support vector machine (SVM), as shown in Supplementary Table 2. As a more detailed explanation, it reveals that random forest can bring an improvement of 10.2%, 38.6%, 23.3%, 24.3% in terms of MAE in predicting the growth rate and an improvement of 31.7%, 30.2%, 32.1%, 22.8% in terms of MAE in predicting the death rate for Japan, South Korea, Pakistan, and Nepal, respectively.

Table 3

| Country  | Prediction | Optimal parameter | MSE (95% CI) | MAE (95% CI) | MAPE |
|----------|------------|-------------------|--------------|--------------|------|
| Japan    | Growth rate | $n = 14, d = 21, s = 8, L = 2, w = 7$ | $3.51 \times 10^{-4}$, $(2.41 \times 10^{-4}, 0.00 \times 10^{-3})$ | $1.12 \times 10^{-2}$, $(0.94 \times 10^{-3}, 1.30 \times 10^{-3})$ | 0.763 |
|          | Death rate  | $n = 24, d = 31, s = 6, L = 10, w = 7$ | $2.12 \times 10^{-3}$, $(6.67 \times 10^{-3}, 3.56 \times 10^{-3})$ | $1.89 \times 10^{-2}$, $(1.39 \times 10^{-2}, 2.38 \times 10^{-2})$ | 0.366 |
| South Korea | Growth rate | $n = 38, d = 35, s = 15, L = 1, w = 7$ | $4.31 \times 10^{-2}$, $(0.00 \times 9.87 \times 10^{-3})$ | $1.73 \times 10^{-2}$, $(0.97 \times 10^{-2}, 2.47 \times 10^{-2})$ | 1.371 |
|          | Death rate  | $n = 29, d = 18, s = 8, L = 14, w = 7$ | $1.62 \times 10^{-2}$, $(0.00 \times 4.31 \times 10^{-2})$ | $3.07 \times 10^{-2}$, $(1.61 \times 10^{-2}, 4.52 \times 10^{-2})$ | 1.895 |
| Pakistan | Growth rate | $n = 20, d = 15, s = 2, L = 8, w = 3$ | $2.63 \times 10^{-2}$, $(5.17 \times 10^{-3}, 7.45 \times 10^{-3})$ | $5.57 \times 10^{-2}$, $(3.83 \times 10^{-2}, 7.40 \times 10^{-2})$ | 2.416 |
|          | Death rate  | $n = 48, d = 38, s = 11, L = 7, w = 3$ | $7.12 \times 10^{-3}$, $(0.00 \times 1.46 \times 10^{-2})$ | $2.95 \times 10^{-2}$, $(2.02 \times 10^{-2}, 3.87 \times 10^{-2})$ | 2.070 |
| Nepal    | Growth rate | $n = 25, d = 17, s = 8, L = 10, w = 3$ | $1.03 \times 10^{-2}$, $(2.08 \times 10^{-3}, 1.85 \times 10^{-2})$ | $4.03 \times 10^{-2}$, $(2.94 \times 10^{-2}, 5.12 \times 10^{-2})$ | 0.746 |
|          | Death rate  | $n = 25, d = 31, s = 13, L = 7, w = 3$ | $8.75 \times 10^{-3}$, $(2.18 \times 10^{-4}, 1.73 \times 10^{-3})$ | $3.33 \times 10^{-2}$, $(2.30 \times 10^{-2}, 4.35 \times 10^{-2})$ | 0.593 |

Note: $n$ is the number of trees in the forest in the range [5, 50], $d$ is the maximum depth of the tree to grow in the range [5, 50], $s$ is the minimum number of data to split an internal node in the range [2, 50], $L$ is the minimum number of data to gain a leaf node in the range [1, 20], $w$ is the size of the rolling time-window defined by the integer (3, 5, 7).

Fig. 3. Summary of the top 20 or top 10 important features in all the eight models for predicting the increase rate of confirmed cases or the death cases. The numeral above each line represents the number of times a certain feature will appear in the top list. The feature importance in (a) and (c) is measured by the mean decrease impurity MDI, while (b) and (d) are based on the TreeSHAP.
3.3. Feature importance analysis

To better understand the inner workings behind the random forest, the importance of prepared features is measured and ranked by two different approaches (an impurity-based method and a SHAP-based method). The potential influence features ranked at the top can be assumed to have greater predictive power. Since a n-day rolling time-window is added to predict the successive daily increase rate based upon its past n values, multiple lagged features with values at the prior time steps can be generated from the time series data. That is to say, each feature in Japan and South Korea datasets will be transformed into 7 lagged features due to the window width of 7, while datasets of Pakistan and Nepal own the 3 lagged features. In view of the number of total lagged features used for model training, Fig. 3 highlights the top 20 features with the highest importance for Japan and Korea and the top 10 important features for Pakistan and Nepal.

It can be assumed all features summarized in Fig. 3 can contribute more to the accuracy of the prediction model, and thus they deserve more investigation. Observably, both the impurity-based score and SHAP value reveal that features representing the temperature (x9 in Japan and South Korea, x4 in Pakistan and Nepal) and stringency index (x12 in Japan and South Korea, x6 in Pakistan and Nepal) tend to appear more frequently in the top 20 or top 10 important feature list. In other words, temperature from the environmental category and stringency index belonging to the social category can be reasonably perceived as the two most influential and relevant features of the growth rate and death rate apart from Nepal. Especially for Japan and South Korea, the two identified features take up even over 40% among the top 20 important features in the two prediction purposes. With regard to the environmental factors, it is found that the relative humidity (x12 in Japan and South Korea, x6 in Pakistan and Nepal) performs as the second most important, since the number of times it turns up in the top list is followed by the feature of temperature. Besides, it can be seen that the degree of contribution from the environmental and social variables is diametrically opposite in the pairwise countries of Japan and South Korea, Pakistan, and Nepal. More specifically, there are more than 12 features concerning the environmental aspects in the top 20 list for the Japan dataset, indicating that environmental factors play a leading role in impacting the prediction performance. Conversely, social-related features accounting for more than 77.78% of all 20 top features will take over the prediction in the South Korea dataset. A similar situation can be seen in Pakistan and Nepal. Therefore, it is of necessity to take both the categories of environment and society into account, aiming to ensure the development of a generalized and robust prediction model for the monitoring and controlling of COVID-19.

Next, we take two random forest models of Japan as an example for model interpretability. The top 20 most important features captured by the SHAP value of the tree ensembles have been provided in Figs. 4a and 5, which are deeply analyzed as follows. As for the explanation of the growth rate model, it can be seen in Fig. 4a that all the top 3 features are about temperature, which are followed by two social features representing the change of people in workplaces. The main difference between Fig. 4a and b is that Fig. 4b with the individualized feature attribution can succinctly show the positive or negative effects of the features on the prediction. From Fig. 4b, it is observed that the three highest-magnitude effects come from temperature about the previous 3 to 5 days. When the temperature drops smoothly, the daily growth rate of confirmed cases has an increasing tendency. There is a long trail on the right, indicating that the extremely low temperature is very possible to raise the risk of accelerating virus spread. Noticeably, all the seven records of temperature within the 7-day rolling window are incorporated in the top 20 list, meaning that temperature from the whole of the last week will pose a combined and significant effect in the growth prediction. Fig. 4c reveals that the relationship between temperature at the previous 7 days and the targeted output of the model is considerably complex, which is neither approximately linear nor monotonic. Besides, the influence of temperature on the predicted growth rate may differ for the strictness of policies. When Japan enforces its strongest measures

Fig. 4. Feature importance analysis by SHAP values in the baseline model of growth rate in Japan: (a) A bar chart of the absolute mean SHAP values, where the top 20 features are sorted in a descending order according to their importance for the model outcome calculated by SHAP values. The feature label x9_d1 means the feature x9 about the ith day within the rolling time-window. (b) SHAP summary plot of the top 20 features, where each point in a line stands for the value of a feature attribution and is colored by the value of a feature. The x-axis displays the range of effects from a certain feature associated with the model outcome. (c) SHAP dependence plot with interaction visualization of the interaction between the temperature and stringency index from the first day of the rolling time-window. The higher SHAP value for the temperature at the previous seven days on the y-axis indicates the larger growth rate or death rate because of this feature. The gray histogram is the data distribution of the SHAP value. (d) A partial dependence plot to visualize the marginal effect of two given features that are the temperature and stringency index from the first day of the rolling time-window and the fix of other variables.
above the value of 40, low temperature will not exert too many negative impacts on disease control. Fig. 4d provides an intuitive visualization of the interaction in the two variables of interest. It shows that the higher temperature and higher level of government response are more likely to correspond to a lower growth rate of daily confirmed cases. An increase of either of the two independent variables can have positive effects in slowing down the growth.

Similarly, the model for predicting the death rate can also be interpreted. Fig. 5a and b turn out that the stringency index and temperature have been ranked into the top 1 and 2, whose SHAP scores are at least 0.0015 greater than other features. It suggests that these two features require special attention in estimating the variation tendency of COVID-19 deaths. It is worth noting that the value of the stringency index from the past 1–5 days will play a greater role in the prediction task. Moreover, there is a sudden leap appears in the vertical direction of Fig. 5c under the value of stringency index over 40. That is to say, even the fairly rigid government responses are executed, the low temperature is still an adverse factor in mitigating COVID-19, which is prone to put the patients in therapy. Additionally, the feature importance of prediction models about the other three countries can be reasonably explored in the way of reducing morality is attempting to avoid too cold conditions for patients in therapy. Furthermore, the feature importance of prediction models about the other three countries can be reasonably explored in the same way. Their corresponding global interpretability based on the SHAP summary plot can be found in Supplementary Figures 2 and 3.

3.4. Optimal control strategies determination

A promising solution for mitigating the COVID-19 transmission is to design optimal control strategies in a data-driven manner, which characterizes a multi-objective optimization problem. To this end, the NSGA-II algorithm is implemented to search the decision space iteratively under the goal of minimizing the growth rate of confirmed cases and the death rate simultaneously. These two objectives are highly associated with environmental and social variables, which have been well fitted in the random forest regression model. All the 18 features in Japan and South Korea and the 12 features in Pakistan and Nepal are served as the decision variables. They are adjusted within the range defined by their minimal and maximal value in Tables 1 and 2 to customize the optimization process. For the four targeted countries, we build four NSGA-II models according to different initialized parameters in Supplementary Table 3.

The convergence of NSGA-II in the four models has been provided in Fig. 6, which can show the effectiveness of the multi-objective optimization algorithm. It can be perceived that the four optimization models all demonstrate a good convergence performance, especially when their iterations are added up to 600, 400, 400, and 50, respectively. After the optimization process, the approximations of the Pareto frontier in Fig. 7a to d can be generated, where red points stand for a set of trade-off optimal choices found by the algorithm. Thus, several prospective solutions under the great balance of minimizing the two objectives are available to a country for implementation, which seem to be more practical than a single solution. The most optimal solution in each figure is highlighted by a blue box, which can be considered to most effectively depress the increase in both the infections and deaths among all possible solutions. This ideal point is the most proximal one to the ideal point (0, 0) under the Euclidean distance of 1.99 × 10^{-3}, 1.39 × 10^{-3}, 1.19 × 10^{-3}, and 1.34 × 10^{-3} for Japan, South Korea, Pakistan, and Nepal, respectively.

To easily arrive at the most optimal solutions determined in Fig. 7, Table 4 lists the relevant profiles of the decision variables. For example, if Japan follows the ideal setting of variables, its growth rate of confirmed cases and death rate are expected to dramatically drop by 94.07% and 96.42% in comparison of the mean of the corresponding 287 records in the actual situation during the study period (the average growth rate of confirmed cases: 2.85 × 10^{-2}, the average death rate: 2.90 × 10^{-2}). Similarly, following the setting of decision variables in Table 4, the optimum solutions for the other three countries can also drive a more than 95% decline in the increase rate of infections and deaths. Therefore, it can be concluded that such optimal combinations of
variables have great application potential in better satisfying the two defined objectives of minimizing the infection rate and mortality together. Governments can refer to them as valuable guidance in the determination of the control policies that are adapted to the characteristics of a certain country. In that respect, the decision-making process for developing appropriate measures no longer largely depends on expert knowledge and subjective judgment. Taking two key features named temperature and stringency index as an example to describe the formulated strategies against COVID-19. According to Table 4, there is no significant difference in the temperature setting in Japan (23.24 °C), South Korea (24.83 °C), and Pakistan (26.75 °C), while the temperature in Nepal is at least 7.85 °C higher than other three countries. The strictness of government responses in the four countries can be sorted as: South Korea (39.62) > Japan (28.74) > Nepal (27.23) > Pakistan (21.90).

Social factors associated with government actions and peoples’ travel distance are preferred in the sequence of non-pharmaceutical interventions against COVID-19. As a matter of fact, it is not a straightforward task to change all the influential factors at the same time to meet all requirements of multiple objectives. Prioritization of the control strategies is another concern. Aiming to investigate the impacts of different factors on the optimization task, another three scenarios are set up through the adjustment of different decision variables, where only the environmental factors, the social factors, and the two critical variables named temperature and stringency index are considered, separately. The Euclidean distance of the most optimal solutions to (0, 0) is calculated for evaluating the optimization performance, which is compared to the distance of original conditions under no control effort to (0, 0) to measure how much the distance is shortened after optimization. The improvement from the four scenarios has been visualized in the beanplot in Fig. 8. Based on the red line in Fig. 8, it is observed that the four scenarios under different priorities of factors in all countries can reach an average improvement of distance over 30%. The improvement of distance is mainly due to the proper adjustment of the related variables, which verifies the considerable potential of the optimization approach in effectively reducing both the confirmed cases and death tolls.

A detailed explanation of Fig. 8 is given below. From the scenario 1 to 4 in each country, the mean of the optimization improvement shows a decreasing trend, while the standard deviation donating the uncertainty continues to grow. To be more specific, the control action based on all variables can always lead to the best improvement in the decrease of COVID-19 spread. This kind of optimization results in a 7.20% (Japan), 10.49% (South Korea), 20.71% (Pakistan), and 6.74% (Nepal) increase of improvement compared with changing variables in the social category. Under scenarios 2 and 3 by controlling the single kind of factors, the average improvement for the two objectives is reduced from 76.55% to 69.25% (Japan), from 65.27% to 53.73% (South Korea), from 59.02% to 46.57%, and from 88.64% to 87.17%. It means that social factors may play more important roles in processing the decreased number of COVID-19 cases and deaths. Therefore, it is essential for decision-makers to give priority to social factors in formulating the optimal strategies, which can be more beneficial for epidemic intervention. For simplicity and timesaving, a reasonable choice can rely on the examination of only the temperature and stringency index. Although it seems sufficient to lower the increase rate defined in the two objective functions for most

Fig. 6. Convergence of the NSGA-II algorithm measured by the recently proposed running metric: (a) The Japan dataset after 800 iterations. (b) The South Korea dataset after 800 iterations. (c) The Pakistan dataset after 800 iterations. (d) The Nepal dataset after 200 iterations. It is observed that there is a significant improvement at the 600th iteration, 400th iteration, 400th iteration, and 50th iteration of the four models, respectively.
cases, its improvement is relatively weaker and unstable than others. Especially for Pakistan, the average optimization improvement from the last scenario is smaller than half of scenario 1 with all variables considered.

4. Discussions

As the prerequisite of optimization, the random forest regressor eventually generates two fitted objective functions for modeling relationships between inputs and two objectives about the daily growth of confirmed cases and deaths, respectively. It is known that the goal of the NSGA-II algorithm that has been well executed is to simultaneously deal with the two objective functions. Differently, when we intend to analyze each objective function independently for simplicity, the problem can be transformed into a single-objective one to minimize either the COVID-19 patients or deaths. For performing a fair comparison, the decision variables and their corresponding value range in the single-objective optimization task are the same as those in the multi-objective optimization process based on NSGA-II.

Fig. 9 presents the comparison results from one multi-objective optimization algorithm named NSGA-II and three single-objective optimization algorithms called GA, BRKGA, and PSO, where the size of each circle is proportional to the value of the minimal growth rate or death rate. Evidently, most results reveal that the single-objective optimization is likely to return a larger reduction in the unwanted cases. For example, the optimized value of growth rate in Japan, South Korea, and Nepal can significantly decline to the smallest value of $1.69 \times 10^{-3}$, $4.12 \times 10^{-3}$, and $3.15 \times 10^{-3}$, respectively, using the proper single-objective methods, which is around 1.66, 1.98, and 4.09 times lower than the results from NSGA-II. That is probably because multi-objective optimization needs to seek the tradeoff between the growth rate and mortality, whereas the single-objective task is chiefly driven by the dominating contributions. This tendency appears in almost all countries, signaling that the effectiveness of single-objective optimization is independent of a certain geographical region. Besides, PSO in this case tends to perform a little worse than the other GA-based methods, resulting in a slightly larger rate of increase. The optimal values of the

| Feature | Japan | South Korea | Feature | Pakistan | Nepal |
|---------|-------|-------------|---------|----------|-------|
| x1      | 2.52  | 4.64        | x1      | 68.92    | 69.63 |
| x2      | 78.81 | 89.98       | x2      | 102.95   | 85.50 |
| x3      | 3.71  | 12.00       | x3      | 1002.52  | 974.47|
| x4      | 34.31 | 42.09       | x4      | 26.76    | 15.39 |
| x5      | 7.64  | 28.15       | x5      | 2.33     | 1.79  |
| x6      | 20.21 | 61.90       | x6      | 21.90    | 27.23 |
| x7      | 1031.76 | 1010.01   | x7      | -9.81    | 0.13  |
| x8      | 2.74  | 4.03        | x8      | -10.97   | -12.24|
| x9      | 23.24 | 24.83       | x9      | 1.57     | 2.45  |
| x10     | 6.69  | 9.12        | x10     | 6.35     | 6.40  |
| x11     | 2.96  | 3.11        | x11     | 0.73     | 10.23 |
| x12     | 28.74 | 39.62       | x12     | -0.40    | 2.81  |
| x13     | -22.52 | 5.78       | x13     |         |       |
| x14     | 1.93  | 42.56       | x14     |         |       |
| x15     | -4.40 | 113.07      | x15     |         |       |
| x16     | -51.56 | -7.48      | x16     |         |       |
| x17     | -10.66 | -7.05      | x17     |         |       |
| x18     | 6.68  | 2.74        | x18     |         |       |

Table 4

Value setting of decision variables to reach the most optimal solution in the four countries.
Fig. 8. Beanplots about the improvement of optimal solutions through adjusting different kinds of decision variables for: (a) Japan. (b) South Korea. (c) Pakistan. (d) Nepal. As for the variable adjustment, we set four different scenarios: Scenario 1 can change all prepared variables; Scenario 2 only changes variables in the social category; Scenario 3 only changes variables in the environmental category; Scenario 4 only changes two variables named temperature and stringency index. The most optimal solution from a certain scenario is compared to the 287 original records in the dataset, and thus the individual improvement is visualized as small gray lines. Also, the estimated density about the distribution of improvement degree is summarized along with the overall average value.

Fig. 9. Comparison of the optimal results from the multi-objective optimization (NSGA-II) and three single-objective optimization algorithms (GA, BRKGA, and PSO): (a) Growth rate of confirmed cases. (b) Death rate. The value of the minimized growth rate and death rate derived from these four optimization algorithms is represented by the circle size, and thus the smaller circle means a lower value. Four targeted countries are differentiated by color.

Table 5: Ideal setting of decision variables produced by the single-objective optimization algorithm named GA, BRKGA, and PSO for the purpose of minimizing the growth rate of confirmed cases or the death rate in Japan and South Korea.

| Variable | Japan Model of growth rate | South Korea Model of growth rate |
|----------|----------------------------|---------------------------------|
|          | GA | BRKGA | PSO | GA | BRKGA | PSO | GA | BRKGA | PSO |
| x1       | 1.06 | 1.77 | 2.36 | 1.01 | 2.88 | 1.36 | 4.98 | 4.69 | 4.61 |
| x2       | 75.84 | 60.11 | 74.55 | 84.31 | 87.54 | 86.52 | 75.29 | 82.05 | 56.91 |
| x3       | 3.97 | 4.74 | 5.80 | 6.49 | 5.96 | 4.79 | 12.14 | 5.61 | 12.07 |
| x4       | 34.98 | 26.00 | 32.76 | 17.64 | 24.49 | 34.62 | 25.04 | 34.35 | 30.52 |
| x5       | 16.90 | 32.58 | 10.11 | 24.49 | 34.62 | 25.04 | 34.35 | 30.52 | 34.96 |
| x6       | 51.21 | 27.53 | 48.77 | 83.57 | 87.54 | 86.52 | 75.29 | 82.05 | 56.91 |
| x7       | 1017.80 | 1012.22 | 1020.52 | 995.73 | 1018.26 | 1013.58 | 1008.81 | 1013.32 | 1020.36 |
| x8       | 2.74 | 2.55 | 3.83 | 1.75 | 1.16 | 3.15 | 4.44 | 5.36 | 4.44 |
| x9       | 19.58 | 16.73 | 19.95 | 20.14 | 23.50 | 21.15 | 19.33 | 13.82 | 14.08 |
| x10      | 3.21 | 4.38 | 14.62 | 15.29 | 2.12 | 9.05 | 8.76 | 9.27 | 8.76 |
| x11      | 3.05 | 7.93 | 8.86 | 2.91 | 6.42 | 4.57 | 2.11 | 2.41 | 6.44 |
| x12      | 35.53 | 41.36 | 31.43 | 36.74 | 20.59 | 20.68 | 43.84 | 43.76 | 43.03 |
| x13      | 1017.80 | 1012.22 | 1020.52 | 995.73 | 1018.26 | 1013.58 | 1008.81 | 1013.32 | 1020.36 |
| x14      | 7.61 | 10.57 | 10.91 | 15.62 | 1.85 | 18.54 | 5.67 | 5.97 | 5.30 |

Y. Pan et al.
decision variables output by the three single-objective optimization algorithms have been listed in Tables 5 and 6, which can also serve as a reliable guideline for formulating government control policies. Due to the great performance of the single-objective algorithm, we can also execute it twice to the two defined objectives in sequence as an optional solution for the bi-objective problem. However, it should be noted that such a method is considerably high-cost effective compared to the multi-objective algorithm.

As for the multi-objective optimization solver, another superiority to notice is that it is designed to produce several optima in an approximation of the Pareto front. That is to say, the concept of domination embedded in the optimization problem can maintain diversity within a population, which can give back a choice of reasonable solutions rather than a single result. As can be seen in Fig. 7a–d, there are 8, 32, 16, and 2 desired Pareto-optimal solutions for Japan, South Korea, Pakistan, and Nepal, respectively. It is believed that strategies relying on such a list of solutions can all dramatically reduce the two objectives, which imply useful suggestions for the local government. The values of the decision variables derived from optimization have been visualized in Fig. 10a–r, and their corresponding descriptive statistics are summarized in Supplementary Tables 4 and 5. Clearly, each variable allows to fluctuate slightly within a proper interval due to the great feasibility and

**Table 6**  
Ideal setting of decision variables produced by the single-objective optimization algorithm named GA, BRKGA, and PSO for the purpose of minimizing the growth rate of confirmed cases or the death rate in Pakistan and Nepal.

| Variable | Model of growth rate | Model of death rate |
|----------|-----------------------|---------------------|
|          | GA        | BRKGA   | PSO     | GA        | BRKGA   | PSO     |
| x1       | 73.76     | 82.61   | 83.62   | 74.81     | 99.84   | 80.38   |
| x2       | 190.56    | 123.30  | 101.53  | 129.32    | 98.42   | 68.79   |
| x3       | 981.11    | 991.08  | 1003.95 | 985.70    | 916.46  | 914.70  |
| x4       | 24.81     | 26.13   | 27.08   | 17.59     | 17.24   | 16.29   |
| x5       | 3.12      | 2.40    | 3.99    | 2.09      | 2.49    | 1.05    |
| x6       | 17.15     | 22.12   | 26.27   | 17.15     | 17.94   | 16.97   |
| x7       | -4.46     | -11.68  | -8.22   | 3.89      | 4.21    | 3.30    |
| x8       | 2.58      | -2.36   | -1.87   | -28.33    | -11.80  | -12.48  |
| x9       | 17.43     | 25.63   | 0.15    | 12.63     | 7.51    | 3.12    |
| x10      | 4.92      | 4.62    | -3.19   | -7.40     | -1.60   | 7.04    |
| x11      | -59.22    | -27.40  | -16.19  | 1.76      | 7.56    | 6.76    |
| x12      | 6.92      | 4.78    | 3.50    | 1.50      | 3.45    | 1.23    |

Fig. 10. A kernel density estimate (KDE) plot for the value of decision variables in the four countries that are determined by the NSGA-II optimization algorithm. (a)–(r) Optimal value setting of decision variables x1–x18 for determining suitable control strategies. It adopts the Kernel smoothing to plot value. The peak of a curve indicates the place where values are concentrated over the interval. For labels of the x-axis, the variable outside the bracket represents the feature in Japan or South Korea dataset, while the variable in the bracket is the feature in Pakistan or Nepal dataset. Japan or South Korea dataset owns six more features than the other two countries.
flexibility of the NSGA-II optimization in an improved decision making of controlling COVID-19 evolution countrywide. The distribution shape and statistical characteristics of a variable in four countries differ from one another. That is because optimal strategies are put forward to meet the country’s own conditions. Importantly, the appropriate temperature can be set as 23.32 (IQR: [23.24, 23.62]), 22.31 (IQR: [22.37, 24.83]), 20.49 (IQR: [19.99, 20.60]), and 14.52 (IQR: [14.06, 15.39]) in Japan, South Korea, Pakistan, and Nepal, respectively. Governments of the four countries are supposed to take harsh measures at the level of 27.40 (IQR: [27.43, 28.74]), 41.31 (IQR: [39.62, 43.11]), 23.44 (IQR: [18.90, 29.95]), and 25.98 (IQR: [23.54, 28.19]). Based upon the proposed optimization process, each country can therefore make the factors balance and suitable according to the characteristics of itself so as to realize the best way for controlling the disease.

5. Conclusions and future works

In summary, this research attempts to develop a combinatorial application of random forest regression and single/multi-objective optimization algorithm, contributing to better curbing the COVID-19 spread for maintaining society sustainability. The practical value of our data-driven framework lies in two aspects. For one thing, it can well capture the transmission dynamics of the virus under consideration of environmental and social variables, resulting in an accurate prediction about the near future increase of COVID-19 cases and deaths in different countries. For another, it can systematically analyze and optimize the relevant factors on the targeted two objectives. Therefore, policymakers can directly refer to these discovered pieces of critical knowledge to realize early warning, preparation, and prevention for crisis control. In particular, the measures adopted over different countries are anticipated to adapt to local conditions, which could be well fitted in sustainable societies. The effectiveness and stability of the proposed approach have been verified in four countries, including Japan, South Korea, Pakistan, and Nepal. Some important lessons can be learned from these four Asia countries as follows.

As for the random forest regression, it is built based on the 3-fold cross-validation, PSO optimization, and proper rolling time-window to learn effective combinations of environment and social features, which is proven effective to exhibit high accuracy in estimating how COVID-19 will evolve nationally. Besides, the environmental factor named temperature and the social factor called stringency index measuring the response level of government are found to play the most dominant role in forecasting the COVID-19 evolution among all considered features. It appears that the relatively high temperature and tight control tend to support a favorable condition for suppressing the spread of the influenza virus. As for the multi-objective optimization process, the NSGA-II algorithm is responsible for generating several Pareto-optimal solutions along with a suitable combination and setting of critical variables that are adapted to the country-specific circumstance, which is beneficial to the disease mitigation. By following the determined profile of decision variables, we can see the considerable potential of cutting down the daily increase rate of infections and deaths in all countries by over 95% compared with the original condition. Accordingly, optimal results provide the local government with actionable suggestions for combating the pandemic. Moreover, it must be stressed that the optimal level based on the social factors is about 7.3% (Japan), 11.54% (South Korea), 12.45% (Pakistan), 1.47% (Nepal) larger than adjustment of environmental variables. That means that factors concerning the social aspect should be the prioritization of designing the control strategies.

Some future work will be performed to address the existing limitations in the current research. Firstly, we can add more variables of interest that significantly impact the pandemic evolution and perform proper preprocessing techniques on the collected raw data, aiming to pursue an improvement of the data quality and representativeness. We will also take more countries into consideration to further verify the feasibility and practicability of our proposed approach. The targeted countries will not be confined to Asia, which can be extended to other continents, like Europe, North America, and so on. As a result, a larger and more comprehensive dataset can be created to better fit the predictive model, aiming to provide deep insights into COVID-19 evolution in different countries. We can therefore easily look for coherent patterns and differences embedded in the spatial-temporal distribution among different countries in Asia or countries in different continents across the world. This discovered knowledge is meaningful to draw valuable conclusions for pandemic mitigation. Secondly, since the pandemic has negatively caused massive global economic disruptions, we can reasonably add a new objective concerning the economic impact of a country into the multi-objective optimization problem. This desired goal at the level of economy is entirely different from the current two objectives about the case and death increase, which deserves further investigation. It is necessary for the government to take effective actions to slow down the virus spread and in the meanwhile limit economic recession. Thirdly, we can strive to explore the real cause-effect rather than the statistical relationship between the inputs and outputs for a more solid analysis. The discovery and measurement of the causal association is another important question in mining complex time-series data.

Declaration of Competing Interest

None.

Acknowledgements

This study is supported by the Ministry of Education Tier I Grants, Singapore (no. 04MNP000279C120, no. 04MNP002126C120) and the Start-Up Grant at Nanyang Technological University, Singapore (no. 04IN5000423C120).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scs.2021.103254.

References

Aji, I. J., & Alharbi, O. M. (2020). "COVID-19: Disease, management, treatment, and social impact". Science of the Total Environment, Article 138661.

Bansal, A. (2020). "Utility of artificial intelligence amidst the COVID 19 pandemic: A review". Journal of Medical Systems, 44(6), 1–6.

Bardia, N. (2020). "Developing a COVID-19 mortality risk prediction model when individual-level data are not available". Nature Communications, 11(1), 1–9.

Bashir, M. F. (2020). "Correlation between climate indicators and COVID-19 pandemic in New York, USA". Science of the Total Environment, 138835.

Breiman, L. (2001). "Random forests". Machine Learning, 45(1), 5–32.

Das, A. (2021). "Living environment matters: Unravelling the spatial clustering of COVID-19 hotspots in Kolkata megacity, India". Sustainable Cities and Society, 65, Article 102577.

Dey, K. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". IEEE Transactions on Evolutionary Computation, 6(2), 182–197.

Gao, Y. (2020). "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19". Nature Communications, 11(1), 1–10.

Gen, M., & Lin, E. (2007). "Genetic algorithms.". Wiley Encyclopedia of Computer Science and Engineering (pp. 1–15).

Gonzales, J. F., & Resende, M. G. (2011). "Biased random-key genetic algorithms for combinatorial optimization". Journal of Heuristics, 17(5), 487–525.

Google (2020). "Google Community Mobility Reports." https://www.google.com/covid19/mobility/.

Hale, T. (2020). "Oxford COVID-19 government response tracker (OxCGRT)." https://www.hsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker.

Hopkins, J. (2020). "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)." Baltimore: Johns Hopkins University; https://www.corre.io/covid19/.

Huang, Nils, Geryofher, Lukas, Londre, Alessandro, Dervic, Elma, et al. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. Nature human behaviour, 4(12), 1303–1312.
Pan, Y., & Zhang, Limao (2021). A BIM-data mining integrated digital twin framework for advanced project management. Automation in Construction, 124, 103564.

Huy, M. (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, Article 102680.

Huppert, A., & Katriel, G. (2013). 'Mathematical modelling and prediction in infectious disease epidemiology.'. Clinical microbiology and infection, 19(11), 999-1005.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN 95 International Conference on Neural Networks. IEEE.

Lim, J. M. (2021). 'Population anxiety and positive behaviour change during the COVID-19 epidemic: Cross-sectional surveys in Singapore, China and Italy.'. Influenza and Other Respiratory Viruses, 15(1), 45–55.

Mansour, S. (2021). "Sociodemographic determinants of COVID-19 incidence rates in Oman: Geospatial modelling using multiscale geographically weighted regression (MGWR).". Sustainable Cities and Society, 65, Article 102627.

Pan, Y. (2020). 'Improved Fuzzy Bayesian Network-Based Risk Analysis With Interval-Valued Fuzzy Sets and D-S Evidence Theory.'. IEEE Transactions on Fuzzy Systems, 28(9), 2063–2077.

Pan, Y., & Zhang, L. (2021). 'Roles of artificial intelligence in construction engineering and management: A critical review and future trends.'. Automation in Construction, 122, Article 103517.

Pan, Yue, & Zhang, Limao (2021). A BIM-data mining integrated digital twin framework for advanced project management. Automation in Construction, 124, 103564.

Rahman, M., A. (2020). 'Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices.'. Sustainable Cities and Society, 62, Article 102372.

Sambigari, S. (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, Article 102418.

Siegenfeld, A. F., & Bar-Yam, Y. (2020). 'The impact of travel and timing in eliminating COVID-19.'. Communications Physics, 3(1), 1-8.

Thakur, N.V. (2020). Coronavirus outbreak: Multi-objective prediction and optimization. intelligent systems and methods to combat Covid-19, Springer; 19-28.

Sun, Chantian, & Zhai, Zhiqiang (2020). The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission. Sustainable Cities and Society, 62, 102390.

Yan, Z. (2018). "On the design of sparse but efficient structures in operations.". Management Science, 64(7), 3421–3445.

Yang, L., & Shami, A. (2020). "On hyperparameter optimization of machine learning algorithms: Theory and practice.". Neurocomputing, 415, 295–316.

Yarikhanot, C. M. (2020). "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm.". Chaos, Solitons & Fractals, 140, Article 110210.

Yusoff, Y. (2011). "Overview of NSGA-II for optimizing machining process parameters.". Procedia Engineering, 15, 3978–3983.

Zaitchik, B. F. (2020). "A framework for research linking weather, climate and COVID-19.". Nature communications, 11(1), 1–3.

Zhan, C. (2021). 'Random-Forest-Bagging Broad Learning System with Applications for COVID-19 Pandemic.'. IEEE Internet of Things Journal, 8(7), 10051-10059.

Zipkovic, M. (2021). "COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach.". Sustainable Cities and Society, 66, Article 102669.