Mathematical Biosciences and Engineering

MBE, 21(5): 6150–6166.
DOI: 10.3934/mbe.2024270
Received: 14 January 2024
Revised: 21 May 2024
Accepted: 29 May 2024
Published: 03 June 2024

http://www.aimspress.com/journal/MBE

Research article

Predicting the transmission trends of COVID-19: an interpretable machine learning approach based on daily, death, and imported cases

Hyeonjeong Ahn and Hyojung Lee*

Department of Statistics, Kyungpook National University, Daegu 41566, Republic of Korea

* **Correspondence:** Email: hjlee@knu.ac.kr.

Abstract: COVID-19 is caused by the SARS-CoV-2 virus, which has produced variants and increasing concerns about a potential resurgence since the pandemic outbreak in 2019. Predicting infectious disease outbreaks is crucial for effective prevention and control. This study aims to predict the transmission patterns of COVID-19 using machine learning, such as support vector machine, random forest, and XGBoost, using confirmed cases, death cases, and imported cases, respectively. The study categorizes the transmission trends into the three groups: L0 (decrease), L1 (maintain), and L2 (increase). We develop the risk index function to quantify changes in the transmission trends, which is applied to the classification of machine learning. A high accuracy is achieved when estimating the transmission trends for the confirmed cases (91.5–95.5%), death cases (85.6–91.8%), and imported cases (77.7–89.4%). Notably, the confirmed cases exhibit a higher level of accuracy compared to the data on the deaths and imported cases. L2 predictions outperformed L0 and L1 in all cases. Predicting L2 is important because it can lead to new outbreaks. Thus, this robust L2 prediction is crucial for the timely implementation of control policies for the management of transmission dynamics.

Keywords: COVID-19; machine learning; transmission; classification; prediction

1. Introduction

Worldwide, several threats to health and human wellness in the 2000s have been experienced due to pandemics. Outbreaks of infectious diseases included SARS in 2002, a subtype of influenza A (H1N1) in 2009, and MERS in 2015 [1,2]. Since the emergence of the novel disease COVID-19 in late 2019, the virus has given rise to several variants, including Delta and Omicron. Notably, ongoing mutations in the Omicron variant continue to occur, which raises concerns about the potential of a resurgence of COVID-19 [3,4].

Furthermore, a possibility exists of novel viruses or existing infectious diseases that can cause diseases in the near future. Thus, anticipating and predicting potential outbreaks of infectious diseases is currently essential to effectively prevent their spread [5,6]. The implementation of effective measures in advance can be beneficial to minimize societal impacts by reducing the cases of infections and fatalities.

Mathematical modeling and machine learning methods are widely used to predict or forecast the transmission dynamics and to analyze effective control interventions [7–12]. Mathematical modeling requires the development of a sophisticated model that considers numerous parameters to accurately represent real-world factors, including vaccines, variants, and policies [13,14]. Machine learning has emerged as an alternative method to predict infectious diseases because it can consider relatively many features with ease [15,16].

Several studies predicted the number of the confirmed or death cases to investigate the transmission trends of COVID-19 data using machine learning [17–21]. Chimmula et al. [17] represented the number of predictions for confirmed cases in Canada with those of the United States and Italy using long short-term memory. Sardar et al. [18] predicted the number of infections for various countries in South Asia using XGBoost (XGB). Lastly, Gothai et al. [19] predicted the spread of COVID-19 using confirmed cases, deaths, and recovery cases in multiple countries through linear regression (LR) and support vector machine (SVM). Many prior studies exclusively employed machine learning methods on COVID-19 data and solely focused on either confirmed cases or deaths to forecast the transmission dynamics over time [7,12,15,17,18,21,22].

We specifically examine the confirmed cases, death cases, and imported cases of COVID-19 data in the Republic of Korea by utilizing interpretable machine learning methods. There is a main reason to consider the deaths data. An increase in the number of severe patients or deaths can lead to serious consequences in the event of an infectious disease outbreak. Therefore, predicting severe patients and deaths can help to utilize medical resources. Moulaei et al. [22] predicted the mortality rate of COVID-19 by classifying the death and recovery rates based on data for hospitalized patients using machine learning methods including random forest. Ramazi et al. [21] forecasted the number of confirmed cases and deaths in the USA in the long range based on the daily confirmed cases.

There have already been numerous previous studies to predict the number of patients or deaths using simple machine learning methods, which represent an excellent performance in prediction [12,22,23]. However, early warnings which estimate the start time of new outbreak become more important for the novel infectious diseases. To do so, it is also important to catch an increasing trend within the disease dynamics. Therefore, our primary objective is to classify changes in the trends rather than to predict the specific case counts. Furthermore, we aim to develop a machine learning-based prediction method that can consistently achieve a high accuracy in forecasting the transmission trends, independent of the dataset employed. The current study proposes a universally applicable prediction approach.

2. Materials and methods

Machine learning has proved to be an effective tool to generate predictions by analyzing extensive datasets [24,25]. Meanwhile, methods that address the interpretability of results are lacking. To overcome this disadvantage, Cho et al. [6] put forward a novel approach to enhance the interpretability in machine learning predictions. The authors introduced a mathematical criterion called the *risk index*, which plays a role similar to an activation function in classifying groups for the machine learning method, to estimate the early detection of new outbreaks. Figure 1 shows the outline of the methodology. The present study extended the previously developed method as follows: (i) while previous studies focused on predicting the confirmed cases for the early detection of new outbreaks, the current study extends its application to include predictions for confirmed, death, and imported cases; and (ii) we developed several risk indexes to classify the groups of transmission trends as increase, maintain, and decrease.

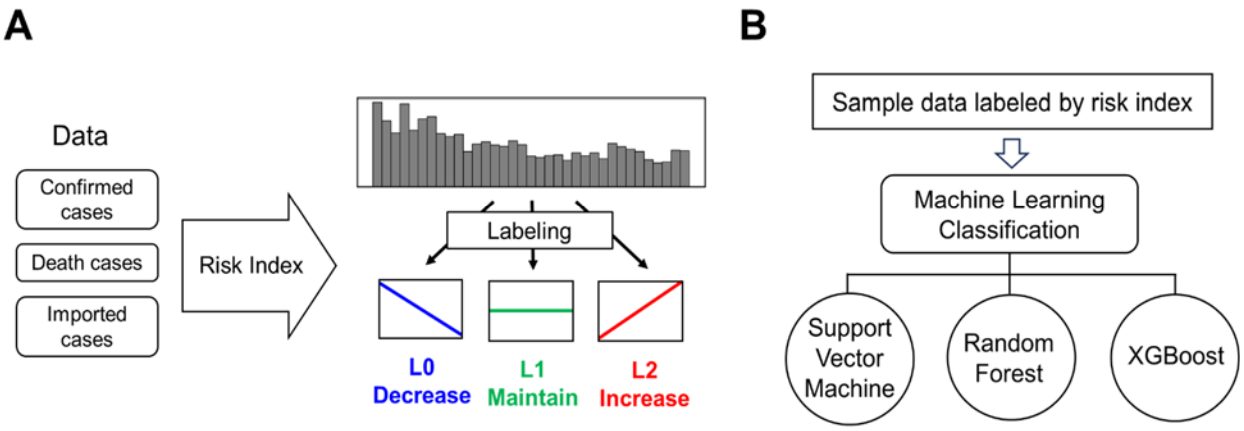

Figure 1. Outline of the study. A. Confirmed cases, deaths, and imported cases data due to COVID-19 are labeled for transmission patterns by risk index. A transmission pattern is divided into three groups, namely, decrease, maintain, and increase. Risk indexes are developed to quantify changes in transmission patterns. B. A transmission trend is estimated using machine learning methods based on sample data labeled according to value of risk index.

2.1. Data description

We analyzed data on the confirmed, death, and imported cases of COVID-19 in South Korea from February 24, 2020, to August 31, 2023. The cumulative numbers of the three cases are 34,491,410 confirmed cases, 35,931 deaths, and 79,907 imported cases. Data was extracted from an open source provided by the Korea Disease Control and Prevention Agency (KDCA) [26]. Additionally, we used the proportions of the Delta and Omicron variants described in [26,27].

2.2. Generation of sample data and linear regression analysis

First, we normalized the number of confirmed cases, deaths, and imported cases of COVID-19 using the min–max normalization (Figure 2A). The sample data were used as the input for the machine learning classification.

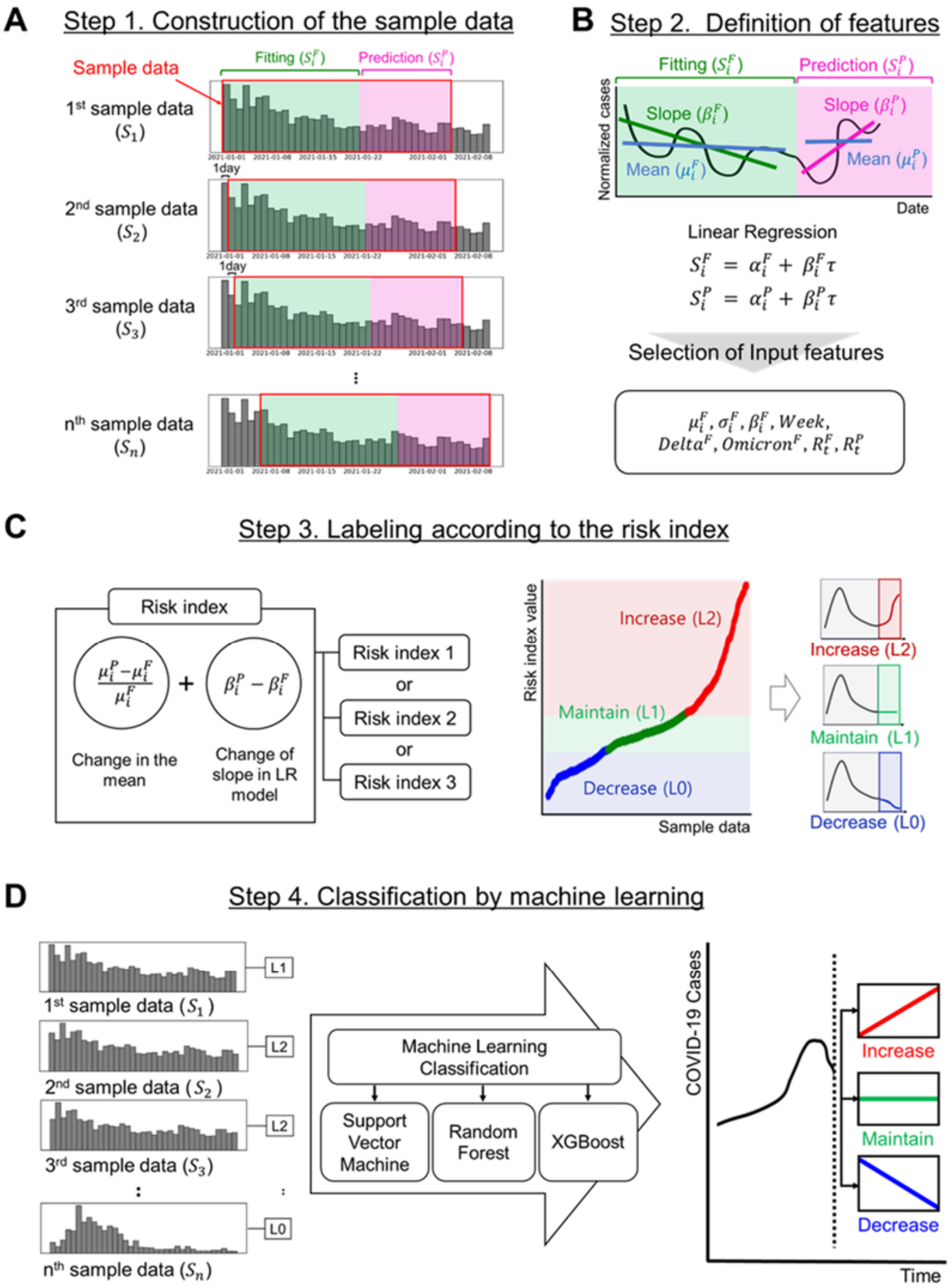

Figure 2. Construction of the sample data. A. The sample data consists of the fitting and prediction periods. B. Eight features were selected as input feature from several statistical values on sample data derived from normalized cases. Some features obtained from results of linear regression model applied for both periods. C. Sample data was labeled by risk indexes that are three functions developed from the linear regression model. The transmission trend is categorized into three groups, namely, decrease (L0), maintain (L1), and increase (L2). D. Transmission patterns are predicted by three machine learning classification methods.

Second, we divided the time period into two, namely the fitting and prediction periods. We considered the initial 21 and subsequent 14 days as the fitting and prediction periods, respectively, for the sample data. We initially considered a 21-day fitting period followed by a 14-day prediction period (i.e., the 21-day fitting and 14-day prediction period). This approach is based on the experience of the South Korean government, which has either maintained or strengthened social distancing measures for a period of 2 weeks during the COVID-19 pandemic described in [6,28,29]. $d_i(t)$ indicates normalized cases on day t for the i-th sample data, which is defined as $S_i = \{d_i(i), d_i(1+i), \dots, d_i(34+i)\}$ for $i \in \{1, \dots, n\}$. In other words, S_i consists of two subgroups (i.e., $S_i = S_i^F \cup S_i^P$). S_i^F and S_i^P represent the sample data for both periods, respectively.

$$S_i^F = \{d_i(i), d_i(1+i), \dots, d_i(19+i), d_i(20+i)\},\ S_i^P = \{d_i(21+i), \dots, d_i(34+i)\}.$$

We obtained the total number of sample data as $n = 1251$ (i.e., $\boldsymbol{S} = \{S_1, S_2, \dots, S_{1251}\}$). The time interval for each i-th sample data is defined as $T_i = T_i^F \cup T_i^P$, where the fitting $T_i^F = \{i, 1+i, \dots, 20+i\}$ and prediction $T_i^P = \{21+i, \dots, 34+i\}$ periods are described. Thus, S_i^F and S_i^P can be rewritten as $S_i^F = \{d_i(\tau)\},\ \tau \in T_i^F$ and $S_i^P = \{d_i(\tau)\},\ \tau \in T_i^P$.

We conducted a sensitivity analysis to compare the prediction performance based on various lengths of the fitting and prediction periods. The other three settings, which were compared to the baseline of a 21-day fitting and a 14-day prediction period, are represented as follows: (i) 28-day fitting and 14-day prediction period (28/14), (ii) 21-day fitting and 28-day prediction period (21/28), and (iii) 21-day fitting and 21-day prediction period (21/21).

Third, we applied an LR model to both periods of the sample data to quantify changes in the transmission trends. β_i^F and β_i^P indicate the slope and α_i^F, α_i^P describe the y-intercept of the LR for both periods, respectively, of the i-th sample data (Figure 2B). The LR models are described as follows:

$$S_i^F = \alpha_i^F + \beta_i^F \tau \ \text{ if } \tau \in T_i^F$$

$$S_i^P = \alpha_i^P + \beta_i^P \tau \ \text{ if } \tau \in T_i^P$$

where τ denotes the days of the time period of the i-th sample data.

2.3. Definition of risk index

A risk index plays the role of a criterion used to determine which of the three labels each sample data corresponds to. The index provides a pivotal factor to categorize the transmission pattern into three distinct groups, similar to the role of an activation function in machine learning classification. It consists of two functions, namely f and g, which indicate trends in the transmission patterns. The f function stands for a change in the average of the normalized cases (i.e., S_i^F and S_i^P), while g denotes a change between the slopes in the LR model (i.e., β_i^F and β_i^P). For each sample data, the value of the risk index is allocated on the last day of the fitting period (T_i^F), represented by δ_i (i.e., $\delta_i = 20 + i$):

$$Risk\ index\ (RI) = f(X_1)g(X_2).$$

We considered three types of risk indexes: Risk index 1 (RI_1), Risk index 2 (RI_2), and Risk index 3 (RI_3), whose functions consist of X_1 and X_2. When X_1 and X_2 are defined as $X_1 = \frac{\mu_i^P - \mu_i^F}{\mu_i^F}$

and $X_2 = \beta_i^P - \beta_i^F$, respectively, we defined RI_1 as follows:

$$Risk\ index\ 1\ (RI_1)\ = f_1(X_1)g_1(X_2) = sinh\left(c_1\left(\frac{\mu_i^P - \mu_i^F}{\mu_i^F}\right)\right)\times e^{c_2(\beta_i^P - \beta_i^F)} \tag{1}$$

where μ_i^F and μ_i^P indicate the average of the normalized cases of COVID-19 for S_i^F and S_i^P, respectively. RI_1 was developed by Cho et al. [6] and is expressed as a combination of hyperbolic and logistic functions. f_1 represents the hyperbolic function that stands for a sharp increase within the cases, while g_1 pertains to the exponential function that denotes a rapidly increasing trend. The c_1 and c_2 indicate the scaling parameters of the functions (f and g) consisting of the risk index.

The two functions can express a rapid increase in the number of cases. Moreover, f_2 and g_2 represent the logistic and exponential functions for RI_2, respectively.

$$Risk\ index\ 2\ (RI_2) = f_2(X_1)g_2(X_2)\ = \left(\frac{1}{1+e^{-c_1\left(\frac{\mu_i^P-\mu_i^F}{\mu_i^F}\right)}}\right)\times e^{c_2(\beta_i^P-\beta_i^F)} \tag{2}$$

RI_3 is expressed by referring to the safety function used by Usherwood et al. [30], who proposed the form of RI_3 to quantify the efficacy of COVID-19 vaccines by examining the impact of human behavior. We developed RI_3 by reconstructing the equation of the safety function and applying the caution and safety factors as X_1 and X_2, respectively. The two functions of f and g for the three types of risk indexes are compared by varying the input variables, X_1 and X_2 (Figure S1).

$$\begin{aligned} Risk\ index\ 3\ (RI_3) &= f_3(X_1) + \left(1 - f_3(X_1)\right)g_3(X_2) \\ &= \frac{1}{e^{-c_1\left(\frac{\mu_i^P-\mu_i^F}{\mu_i^F}\right)}} + \left(1 - \frac{1}{e^{-c_1\left(\frac{\mu_i^P-\mu_i^F}{\mu_i^F}\right)}}\right)\times(-c_2(\beta_i^P - \beta_i^F)) \end{aligned} \tag{3}$$

2.4. Definition of features and classification of transmission trends for machine learning

We used the eight features to classify the transmission patterns into three groups for the fitting and prediction periods. We calculated the effective reproduction number, which is defined as the average number of secondary infected people by the first infected persons as described in Cori et al., by adhering to the gamma distribution of the serial interval (mean: 4.8; SD: 2.3) [31].

Table 1 summarizes the eight features: the average (μ_i^F) and SD (σ_i^F) for the fitting period, the slope (β_i^F) of the LR applied to the fitting period, the average of the ratios of the Delta and Omicron variants ($Delta^F$ and $Omicron^F$, respectively) for the fitting period, $Week$, and the effective reproduction numbers for the fitting and prediction periods.

Figure 2C depicts the labels for the transmission patterns, which were divided into three groups (i.e., L0, L1, and L2). The label for each sample data was determined using the values of the risk index according to the risk index function. The values of the risk indexes of all sample data were ranked from the minimum to the maximum values. Specifically, the sample data were labeled L0 and L2 when the values of the risk index were included in 30% of the smallest (decrease) and largest (increase) values of the risk index, respectively. The other sample data were labeled as L1 (maintain).

Figure 2D presents that transmission patterns that were predicted using the following machine learning methods: support vector machine (SVM), random forest (RF), and XGBoost (XGB). These methods are mainly used to predict trends of COVID-19 and exhibit high levels of performance [10,32]. The total sample data were composed of split training and testing data at a ratio of 7:3. The sizes of the training and testing data are 875 and 375, respectively. We used the respective sample data of the confirmed, death, and imported cases for the input data in machine learning. We predicted the transmission patterns for the fitting period of the sample data. The transmission patterns were designated into a group based on the values of the risk index. The performances of the machine learning methods in predicting transmission patterns are represented by the accuracy for the F1-score. The F1-score represents the harmonic mean of a model's precision and recall and provides a comprehensive evaluation of a model's performance. A higher F1-score indicates a better performance of the model. Additionally, the confusion matrices and the receiver operating characteristic (ROC) curves were compared. The hyperparameters were found through a grid search with a 10-fold cross validation. Table S1 provides the results of the grid search and the ranges of the hyperparameters.

Table 1. Features of machine learning classification.

Features	Description
μ_i^F	Average number of cases for the fitting period of the sample data
σ_i^F	Standard deviation of cases for the fitting period of the sample data
β_i^F	Slope obtained from the linear regression model applied to the fitting period of the sample data
$Week$	Day of the week at the first day of the fitting period of the sample data
$Delta^F$	Average number of ratios of the Delta variant for the fitting period of the sample data
$Omicron^F$	Average number of ratios of the Omicron variant for the fitting period of the sample data
R_t^F, R_t^P	Average number of the effective reproduction number for the fitting and prediction periods of the sample data

3. Results

3.1. Risk index used to classify the sample data into three labels

Figure 3 presents the numbers of the confirmed, death, and imported cases of COVID-19 from January 1, 2021, to August 31, 2023, in South Korea.

Figure 4 presents the range of the risk index using the three labels for three risk index functions using the confirmed cases. We determined the values of c_1 and c_2 based on achieving a high correlation between the risk index values and the labels. Table S2 presents the selected values of c_1 and c_2 that yielded a high correlation between the risk index and the labels, where c_1 and c_2 were predominantly set at 0.01.

When using RI_1 and RI_3, the value of the risk index for the sample data that exhibited a sharp increase were identified and categorized as L2, where L2 encompassed values within a wider range compared to the values from the other labels. In the case of RI_2, the study observed a consistent upward trend for the values of the risk index, which indicated that the three labels were classified based on the characteristics of the risk index functions. We applied a similar calculation to the data on the death and imported cases. Figures S2 and S3 illustrate the resulting ranges of the risk indexes for the data on the death and imported cases, respectively.

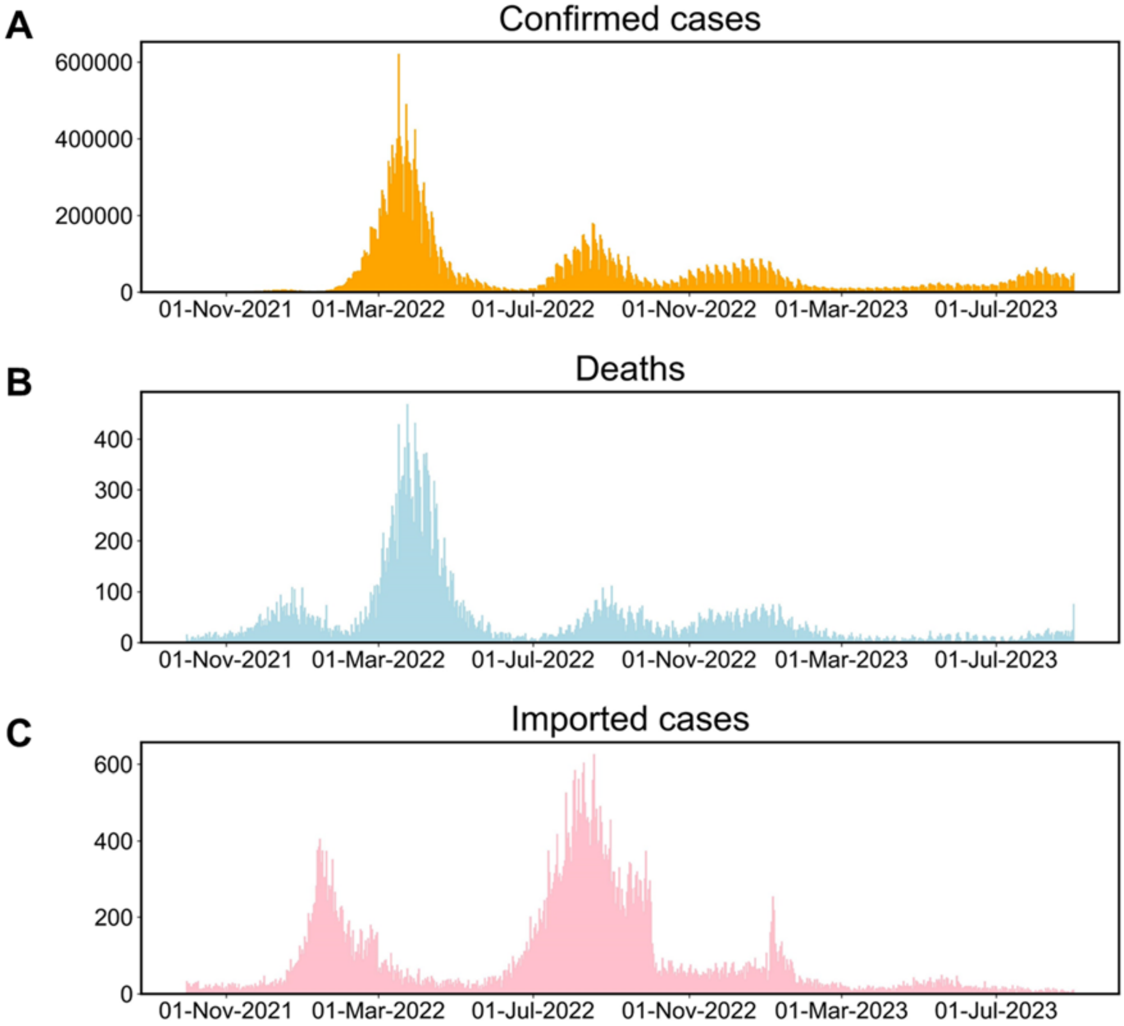

Figure 3. COVID-19 data for confirmed, death, and imported cases. The three types of COVID-19 data from February 24, 2020, to August 31, 2023, in South Korea was used for predicting transmission patterns.

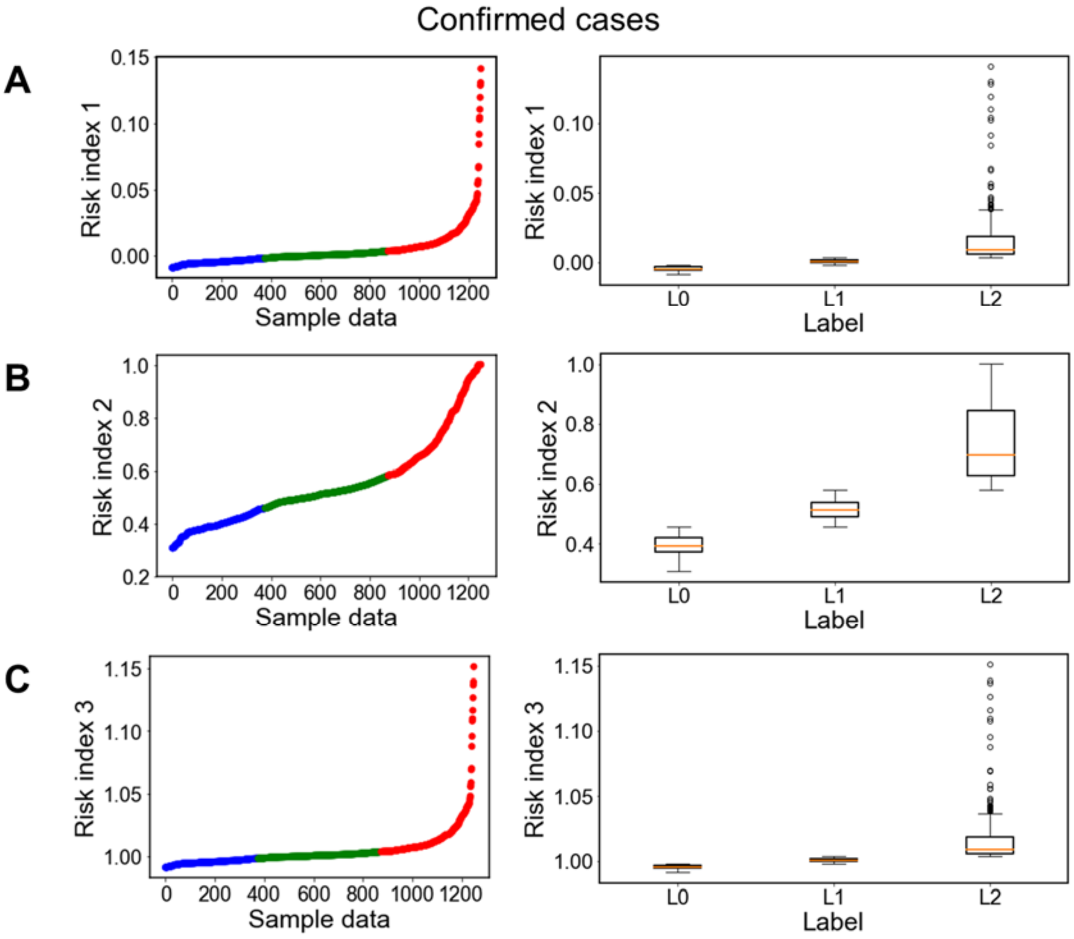

Figure 4. Range and distribution of the values of the three types of risk index for confirmed cases. The distribution of values (left panel) refers to the values of risk index labeled L0 (blue), L1 (green), and L2 (red). The box plot (right panel) depicts the range of the values of risk index according to label.

3.2. Classification of transmission trends via machine learning methods

Figures S4–S6 display the results of the feature importance among the eight features to predict the labels of the transmission trends according to the different risk index functions. We compared the feature importance by computing for RF and XGB, where both methods produced high levels of accuracy (Figure 5). Figure S4 (confirmed cases) and Figure S5 (death cases) indicate that the SD for the fitting period (σ_i^F) and the slope of the LR applied to the fitting period (β_i^F) are important features. In addition, the effective reproduction number for the fitting and prediction periods (R_t^F and R_t^P) were considered important features. However, Figure S6 indicates that the standard deviation (σ_i^F), the average (μ_i^F), the average of the ratios of the Delta variant ($Delta^F$) for the fitting period, and the slope of the LR applied to the fitting period (β_i^F) are important features for the imported cases.

Given the data on the three types of cases, we aim to examine the extent to which it predicts the transmission trends as L0, L1, and L2 by employing the machine learning methods SVM, RF, and XGB. Figure 5 describes their accuracy in terms of the F1-score using the three functions of the risk indexes (i.e., RI_1, RI_2, and RI_3) for the data on confirmed (Figure 5A), death (Figure 5B), and imported cases (Figure 5C). The three risk indexes demonstrated high levels of accuracy in classifying the confirmed cases compared to those for the death and imported cases. The accuracy of predicting the death and imported cases ranged from 85.6% to 91.8% and from 77.7% to 89.4%, respectively. For the confirmed cases, the accuracy exceeds 91.5% regardless of the machine learning method used, which reached a maximum of 95.5% for all functions of the risk indexes.

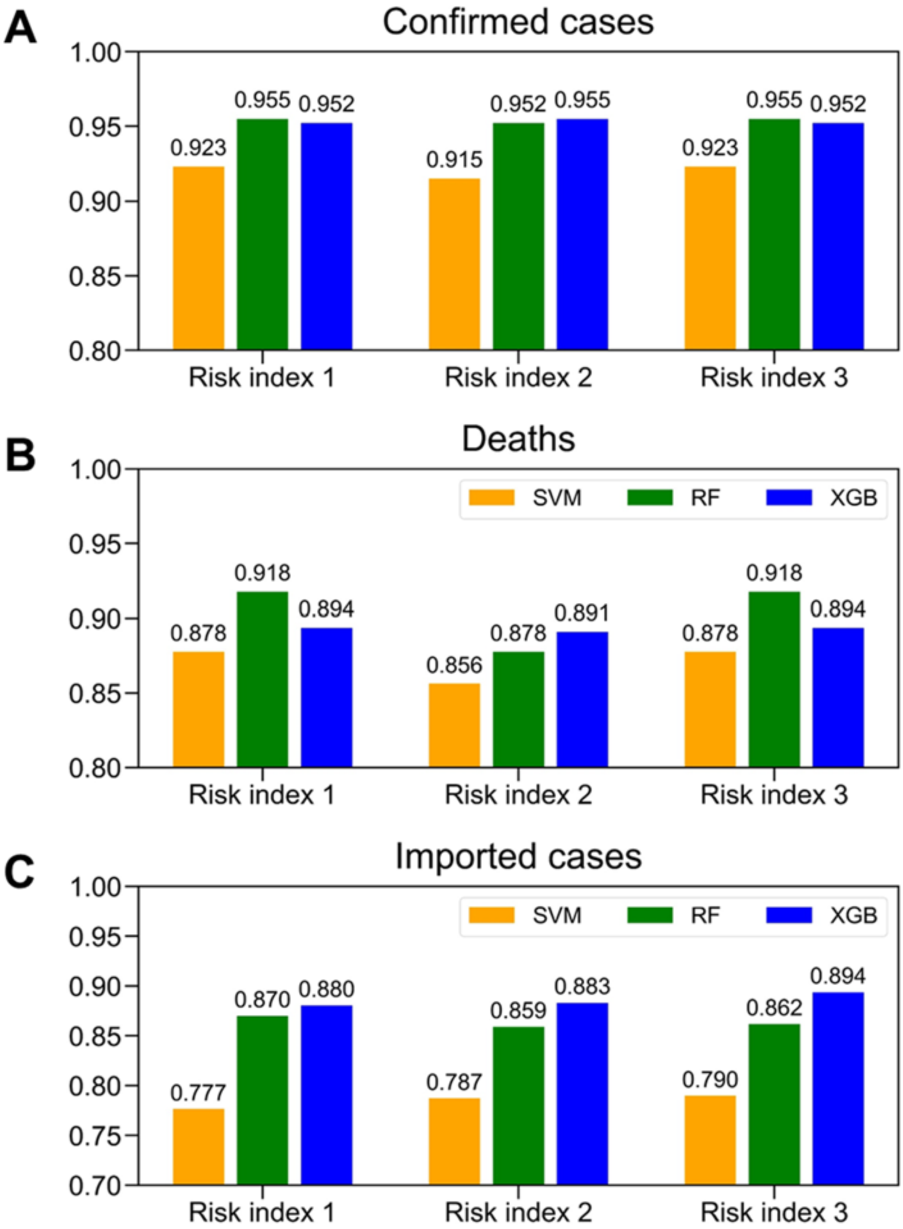

Figure 5. Accuracy according to risk index for confirmed, death, and imported cases.

In summary, the accuracy for the death cases exceeded 85.6% using SVM with RI_2 and reached a maximum at 91.8% with RI_1 or RI_3 using RF. Moreover, the accuracy for the imported case data exceeded 77.7% when obtained through the SVM with RI_1 and reached a maximum at 89.4% using XGB with RI_3 (Figure 5B,C). In addition, for the confirmed and death cases, machine learning classification techniques indicated that the highest accuracy for RI_1 and RI_3 was produced using RF, while the highest accuracy for RI_2 was obtained via XGB. For the imported cases, the highest accuracy was produced using XGB regardless of the three types of risk indexes. We observed a similar result of the accuracy for RI_1 and RI_3 even if they were differently defined functions. The distribution of RI_1 and RI_3 may be similar, as depicted in Figures 4, S2, and S3, regardless of the type of case.

We demonstrate the results of the ROC curve in Figures S7–S9 for the confirmed, death, and imported cases, respectively, to evaluate the performance of the prediction in terms of the classification and the accuracy. An ROC curve is frequently used in binary classifications and medical applications. The closer the curve is to the upper left, the higher the prediction accuracy. Using an ROC curve, we found that the performance of the prediction was higher for the confirmed cases compared to the death or imported cases.

We varied the period of fitting and prediction as a 28-day fitting and a 14-day prediction period, a 21-day fitting and a 28-day prediction period, and a 21-day fitting and a 21-day prediction period. Table S3 and Figure S10 present the results of the prediction accuracy across three risk indexes and different fitting and prediction periods. Here, we found that longer fitting or prediction periods improved the prediction performance for the deaths and imported cases compared to the 21-day fitting and 14-day prediction period. The prediction accuracy reached a maximum at 93.5% when RI_1 and RI_2 were used for the imported cases compared to a maximum accuracy of 86.8% for a 21-day fitting and a 14-day prediction period setting.

The highest accuracy for the death cases showed 93.8% and obtained the best performance in a 21-day fitting and a 28-day prediction period setting for all three risk indexes. For the confirmed cases, the prediction accuracy was higher than in the event of the deaths and imported cases, regardless of the period settings.

3.3. High accuracy in identifying the rising trend of transmission

Figure 6 provides a comparison of the accuracy according to the three labels (i.e., L0, L1, and L2) to compare the efficiency of the classification when using different machine learning methods and functions of the risk indexes. The three risk indexes displayed high levels of performance for the confirmed cases in terms of accuracy according to the label compared with the death and imported cases, which had the same interpretation as explained in Figure 5. In summary, observing that the accuracy was higher for L2 (increase) compared with other labels of transmission trends is easy. Figure 6A presents the results of the comparison of the accuracy according to RI_1 using confirmed cases. The results obtained through RF indicated that the accuracy for L0, L1, and L2 was obtained at 95.5%, 94.6%, and 96.8%, respectively.

Similarly, we obtained a high accuracy regardless of the data and machine learning methods for L2. It plays a crucial role in detecting early outbreaks because it can accurately capture a rapid increase in the trends. We investigated the accuracy according to label for the death and imported cases (Figure S11). All classification techniques represented that the accuracy for L2 was higher than those for L0 and L1 for the death and imported cases. In addition, RI_2 showed a higher performance accuracy of 96.9% for L0 for the confirmed cases compared to when using the other two risk indexes. Table S4 provides the specific values of the accuracy of prediction using the three labels and risk indexes for the COVID-19 data. We presented confusion matrices to represent the confirmed, death, and imported cases in Figures S12–S14, which directly demonstrate the number of errors of prediction according to the label.

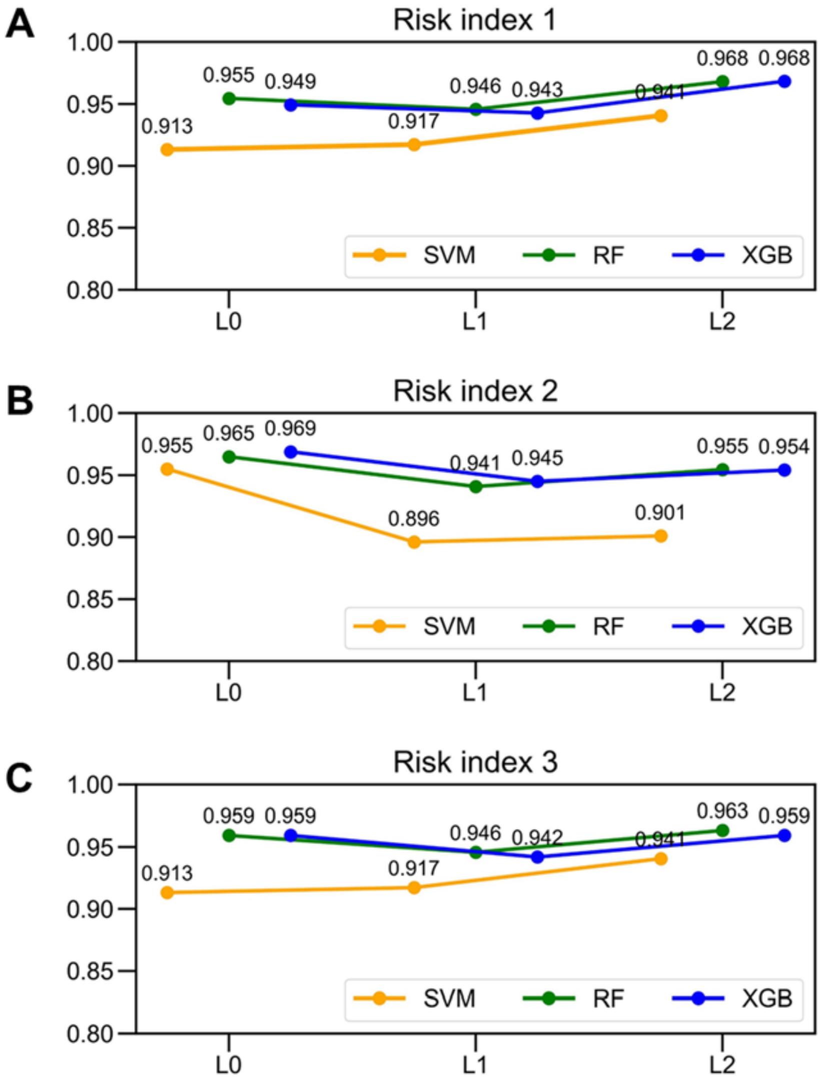

Figure 6. Accuracy of classification using machine learning methods to predict transmission trends according to the three labels and risk indexes for confirmed cases. The accuracy of prediction for L2 is higher than those for L0 and L1 for all types of risk index. A. For RI_1, RF well predicted all labels (L0: 95.5%, L1: 94.6%, L2: 96.8%). B. The accuracy of prediction according to label is highly represented in overall cases of RI_2 using XGB (L0: 96.9%, L1: 94.5%, L2: 95.4%). C. The accuracy of prediction according to label produced high values using RF (L0: 95.9%, L1: 94.6%, L2: 96.3%).

4. Discussion

The objective of the present study was to utilize machine learning techniques to estimate the transmission patterns of COVID-19. We developed a mathematical criterion called the risk index to quantitatively assess changes in the transmission trends. The risk index values were computed for each dataset, which encompasses the confirmed, death, and imported cases of COVID-19. We classified the labels of transmission trends into three categories, namely L0 (decrease), L1 (maintain), and L2 (increase), based on the risk index values of each sample.

The findings demonstrated a notable performance in predicting the transmission patterns of COVID-19. The accuracy for the confirmed, death, and imported cases ranged from 91.5% to 95.5%, from 85.6% to 91.8%, and from 77.7% to 89.4%, respectively. Notably, machine learning classification using the three risk indexes exhibited a higher accuracy for the confirmed cases compared with those for the death and imported cases.

When assessing accuracy based on labels, the prediction accuracy for L2 surpassed that of L0 and L1 in all cases, except for accuracy obtained from RI_2 for the death cases. Among the three risk indexes, RI_2 represented the highest prediction accuracy of 96.9% for L0 for confirmed cases. Specifically, the accuracy of predicting L2 was impressive, which ranged from 95.4% to 96.8% for the confirmed cases, from 89.1% to 95.6% for the death cases, and from 83.0% to 93.6% for the imported cases. This robust performance in predicting L2 is particularly crucial in analyzing the transmission dynamics because the timely identification of new outbreaks enables the implementation of effective control policies to halt the spread of infection. Several studies have utilized classification methods to predict COVID-19 outcomes, such as the confirmed cases and deaths. Holanda et al. [12] used the various machine learning algorithms to forecast the number of hospitalized cases and deaths in Brazil, achieving an accuracy of 83% and an area under the curve (AUC) of 0.89. Keser et al. [32] focused on the number of deaths in Wuhan using gradient-boosting models, thereby incorporating patient history and demographic information, which resulted in an accuracy ranged from 74.07% to 84.39% and an AUC between 86.50% and 87.97%. Our study surpasses these in the predictive accuracy, with a confirmed case accuracy between 91.5% and 95.5% and death outcomes from 85.6% to 91.8%. Moreover, our results feature high AUC scores close to 1, demonstrating the effectiveness of using a risk index as a mathematical criterion for machine learning classification.

We found how differently our prediction method performs on each of the confirmed cases, imported cases, and deaths. In the present study, we newly suggested the RI_2 among the three risk indexes. It has several advantages with respect to the performance: (i) RI_2, in conjunction with the other risk indices, shows the highest predictive accuracy for the increase pattern (L2), achieving an accuracy of 96.8%; and (ii) in terms of predicting the decrease pattern (L0) in the confirmed cases, RI_2 outperforms the other risk indices, with a maximum prediction accuracy of 96.9%. We suggest the potential to predict the transmission patterns for data originating from other countries or emerging infectious diseases in the future.

We conducted a sensitivity analysis by varying the length of the fitting and prediction periods. The results were compared using a consistent setting of 21 days for the fitting period and 14 days for the prediction period. Our analysis, as shown in Table S3, indicates that altering the fitting and prediction periods can enhance the prediction performance for the deaths and imported cases. Specifically, the prediction accuracy for the deaths using the 21/28 days setting showed a higher performance compared to other period settings across all the three risk indexes. Therefore, extending the prediction period improves the prediction performance for the deaths.

This had several limitations. Our objective was to establish a robust methodology to categorize the transmission trends by devising multiple metrics, such as Risk index 1, Risk index 2, and Risk index 3, rather than emphasizing public health aspects. We analyzed epidemiological data solely in Korea. However, previous studies compared the results of predictions for various countries worldwide and predicted the trend of transmission globally [17,18,33]. Chumachenko et al. [33] predicted the number of confirmed cases in Japan, Germany, the Republic of Korea, and Ukraine, which implemented various control policies, using RF and XGBoost. Their analysis primarily focused on data of the confirmed cases. However, in the current study, we demonstrated effective predictions not only for the confirmed cases, but also for the death and imported cases with a high performance.

On the other hand, we didn't consider the deep learning method to predict the transmission patterns, although deep learning conducts prediction by performing complex calculations with large amounts of data. Most studies that used deep learning predicted the number of confirmed cases or deaths. Ardabili et al. [34] predicted the number of infections based on the confirmed cases data using artificial Neural Networks such as Multi-Layered Perceptron (MLP) and Adaptive Network-based Fuzzy Inference System (ANFIS) in five countries, including the USA and Germany. Shahid et al. [35] compared several deep learning forecast models to estimate the confirmed cases, deaths and recoveries, focusing on Long Short-Term Memory (LSTM). However, the current study had a different objective compared to previous studies using deep learning. We proposed a generally applicable prediction approach that captures the changes of the transmission trends. It does not require complex data and predicts the trend patterns with a high performance accuracy of over 95%, using only simple daily data such as confirmed cases, deaths, and imported cases. Additionally, our method has the potential to be applied regardless of the data set.

Moreover, this study is limited to the estimation of the classification of the transmission trends without the forecasting transmission trends compared with those of several studies [20,23,36,37]. Srivastava et al. [36] forecasted daily confirmed cases of COVID-19 by pandemic scenarios using machine learning and compared the performances of the United States, Italy, and Australia in forecasting. Rustam et al. [23] forecasted the number of confirmed, death, and recovered cases for the next 10 days in several countries using supervised machine learning methods. This study presented newly developed activation functions that could capture changes in the transmission trends using LR models in addition to using data for death and imported cases. We proposed a prediction method to analyze patterns in the transmission trends of COVID-19 using interpretable machine learning methods combined with a mathematical criterion. Thus, we predicted the transmission patterns of COVID-19, regardless of the dataset used, with a high accuracy using the simple machine learning methods for classification.

Therefore, appropriately implementing policies for Nonpharmaceutical interventions (NPIs) or vaccination and pharmaceutical treatments in a timely manner could be helpful for policy makers in the case of the resurgence of COVID-19 or the occurrence of new infectious diseases by novel viruses. As such, the spread of the disease can be predicted by analyzing data from various regions, as well as data for cases of the disease.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by Kyungpook National University Research Fund, 2021.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. E. D. Wit, N. V. Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: recent insights into emerging coronaviruses, *Nat. Rev. Microbiol.*, **14** (2016), 523–534. https://doi.org/10.1038/nrmicro.2016.81
2. H. Nishiura, C. Castillo-Chavez, M. Safan, G. Chowell, Transmission potential of the new influenza A (H1N1) virus and its age-specificity in Japan, *Eurosurveillance*, **14** (2009), 19227. https://doi.org/10.2807/ese.14.22.19227-en
3. D. V. Parums, Editorial: A rapid global increase in COVID-19 is due to the emergence of the EG.5 (Eris) subvariant of omicron SARS-CoV-2, *Med. Sci. Monit.*, **29** (2023), e942244. https://doi.org/10.12659/MSM.942244
4. C. Chakraborty, M. Bhattacharya, H. Chopra, M. A. Islam, G. Saikumar, K. Dhama, The SARS-CoV-2 Omicron recombinant subvariants XBB, XBB.1, and XBB.1.5 are expanding rapidly with unique mutations, antibody evasion, and immune escape properties—an alarming global threat of a surge in COVID-19 cases again?, *Int. J. Surg.*, **109** (2023), 1041–1043. https://doi.org/10.1097/JS9.0000000000000246
5. M. Coccia, Sources, diffusion and prediction in COVID-19 pandemic: lessons learned to face next health emergency, *AIMS Public Health*, **10** (2023), 145–168. https://doi.org/10.3934/publichealth.2023012
6. G. Cho, J. R. Park, Y. Choi, H. Ahn, H. Lee, Detection of COVID-19 epidemic outbreak using machine learning, *Front. Public Health*, **11** (2023), 1252357. https://doi.org/10.3389/fpubh.2023.1252357
7. A. Dairi, F. Harrou, A. Zeroual, M. M. Hittawe, Y. Sun, Comparative study of machine learning methods for COVID-19 transmission forecasting, *J. Biomed. Inf.*, **118** (2021), 103791. https://doi.org/10.1016/j.jbi.2021.103791
8. H. Kang, K. D. Min, S. Jeon, J. Y. Lee, S. I. Cho, A measure to estimate the risk of imported COVID-19 cases and its application for evaluating travel-related control measures, *Sci. Rep.*, **12** (2022), 9497. https://doi.org/10.1038/s41598-022-13775-0
9. W. C. Wang, T. Y. Lin, S. Y. Chiu, C. N. Chen, P. Sarakarn, M. Ibrahim, et al., Classification of community-acquired outbreaks for the global transmission of COVID-19: Machine learning and statistical model analysis, *J. Formosan Med. Assoc.*, **120** (2021), S26–S37. https://doi.org/10.1016/j.jfma.2021.05.010
10. S. G. Paul, A. Saha, A. A. Biswas, M. S. Zulfiker, M. S. Arefin, M. M. Rahman, et al., Combating COVID-19 using machine learning and deep learning: Applications, challenges, and future perspectives, *Array*, **17** (2023), 100271. https://doi.org/10.1016/j.array.2022.100271

11. G. Cho, Y. J. Kim, S. H. Seo, G. Jang, H. Lee, Cost-effectiveness analysis of COVID-19 variants effects in an age-structured model, *Sci. Rep.*, **13** (2023), 15844. https://doi.org/10.1038/s41598-023-41876-x
12. W. D. de Holanda, L. C. e Silva, Á. A. C. C. Sobrinho, Machine learning models for predicting hospitalization and mortality risks of COVID-19 patients, *Expert Syst. Appl.*, **240** (2024), 122670. https://doi.org/10.1016/j.eswa.2023.122670
13. S. Kim, Y. Ko, Y. J. Kim, E. Jung, The impact of social distancing and public behavior changes on COVID-19 transmission dynamics in the Republic of Korea, *PLoS One*, **15** (2020), e0238684. https://doi.org/10.1371/journal.pone.0238684
14. A. Olivares, E. Staffetti, Optimal control-based vaccination and testing strategies for COVID-19, *Comput. Methods Programs Biomed.*, **211** (2021), 106411. https://doi.org/10.1016/j.cmpb.2021.106411
15. S. Agrebi, A. Larbi, Use of artificial intelligence in infectious diseases, in *Artificial Intelligence in Precision Health*, (2020), 415–438. https://doi.org/10.1016/B978-0-12-817133-2.00018-5
16. F. Wong, C. de la Fuente-Nunez, J. J. Collins, Leveraging artificial intelligence in the fight against infectious diseases, *Science*, **381** (2023), 164–170. https://doi.org/10.1126/science.adh1114
17. V. K. R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, *Chaos Solitons Fractals*, **135** (2020), 109864. https://doi.org/10.1016/j.chaos.2020.109864
18. I. Sardar, M. A. Akbar, V. Leiva, A. Alsanad, P. Mishra, Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: methodology, evaluation, and case study in SAARC countries, *Stochastic Environ. Res. Risk Assess.*, **37** (2023), 345–359. https://doi.org/10.1007/s00477-022-02307-x
19. E. Gothai, R. Thamilselvan, R. R. Rajalaxmi, R. M. Sadana, A. Ragavi, R. Sakthivel, Prediction of COVID-19 growth and trend using machine learning approach, *Mater. Today Proc.*, **81** (2023), 597–601. https://doi.org/10.1016/j.matpr.2021.04.051
20. I. Heredia Cacha, J. Sainz-Pardo Diaz, M. Castrillo, A. Lopez Garcia, Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain's case study, *Sci. Rep.*, **13** (2023), 6750. https://doi.org/10.1038/s41598-023-33795-8
21. P. Ramazi, A. Haratian, M. Meghdadi, A. Mari Oriyad, M. A. Lewis, Z. Maleki, et al., Accurate long-range forecasting of COVID-19 mortality in the USA, *Sci. Rep.*, **11** (2021), 13822. https://doi.org/10.1038/s41598-021-91365-2
22. K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, H. Kazemi-Arpanahi, Comparing machine learning algorithms for predicting COVID-19 mortality, *BMC Med. Inf. Decis. Making*, **22** (2022), 2. https://doi.org/10.1186/s12911-021-01742-0
23. F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. W. On, W. Aslam, et al., COVID-19 future forecasting using supervised machine learning models, *IEEE Access*, **8** (2020), 101489–101499. https://doi.org/10.1109/access.2020.2997311
24. E. Y. Alqaissi, F. S. Alotaibi, M. S. Ramzan, Modern machine-learning predictive models for diagnosing infectious diseases, *Comput. Math. Methods Med.*, **2022** (2022), 6902321. https://doi.org/10.1155/2022/6902321
25. N. M. Tayarani, Applications of artificial intelligence in battling against COVID-19: A literature review, *Chaos Solitons Fractals*, **142** (2021), 110338. https://doi.org/10.1016/j.chaos.2020.110338

26. Korea Disease Control and Prevention (KDCA), *Open Source Data for COVID-19*, 2023. Available from: https://dportal.kdca.go.kr/pot/cv/trend/dmstc/selectMntrgSttus.do.
27. CoVariant, *Overview of Variants in Countries*, 2023. Available from: https://covariants.org/per-country.
28. H. Zhao, N. N. Merchant, A. McNulty, T. A. Radcliff, M. J. Cote, R. S. B. Fischer, et al., COVID-19: Short term prediction model using daily incidence data, *PLoS One*, **16** (2021), e0250110. https://doi.org/10.1371/journal.pone.0250110
29. H. Du, E. Dong, H. S. Badr, M. E. Petrone, N. D. Grubaugh, L. M. Gardner, Incorporating variant frequencies data into short-term forecasting for COVID-19 cases and deaths in the USA: a deep learning approach, *EBioMedicine*, **89** (2023), 104482. https://doi.org/10.1016/j.ebiom.2023.104482
30. T. Usherwood, Z. LaJoie, V. Srivastava, A model and predictions for COVID-19 considering population behavior and vaccination, *Sci. Rep.*, **11** (2021), 12051. https://doi.org/10.1038/s41598-021-91514-7
31. H. Lee, Y. Kim, E. Kim, S. Lee, Risk assessment of importation and local transmission of COVID-19 in South Korea: Statistical modeling approach, *JMIR Public Health Surveillance*, **7** (2021), e26784. https://doi.org/10.2196/26784
32. S. B. Keser, K. Keskin, A gradient boosting-based mortality prediction model for COVID-19 patients, *Neural Comput. Appl.*, **35** (2023), 23997–24013. https://doi.org/10.1007/s00521-023-08997-w
33. D. Chumachenko, I. Meniailov, K. Bazilevych, T. Chumachenko, S. Yakovlev, Investigation of statistical machine learning models for COVID-19 epidemic process simulation: Random forest, K-nearest neighbors, gradient boosting, *Computation*, **10** (2022), 86. https://doi.org/10.3390/computation10060086
34. S. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. Varkonyi-Koczy, U. Reuter, et al., COVID-19 outbreak prediction with machine learning, *Algorithms*, **13** (2020), 249. https://doi.org/10.3390/a13100249
35. F. Shahid, A. Zameer, M. Muneeb, Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM, *Chaos Solitons Fractals*, **140** (2020), 110212. https://doi.org/10.1016/j.chaos.2020.110212
36. A. K. Srivastava, S. M. Tripathi, S. Kumar, R. M. Elavarasan, S. Gangatharan, D. Kumar, et al., Machine learning approach for forecast analysis of novel COVID-19 scenarios in India, *IEEE Access*, **10** (2022), 95106–95124. https://doi.org/10.1109/access.2022.3204804
37. Y. Alali, F. Harrou, Y. Sun, A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models, *Sci. Rep.*, **12** (2022), 2467. https://doi.org/10.1038/s41598-022-06218-3