Hindawi
International Journal of Telemedicine and Applications
Volume 2023, Article ID 9962100, 24 pages
https://doi.org/10.1155/2023/9962100

Research Article

Effective Utilization of Data for Predicting COVID-19 Dynamics: An Exploration through Machine Learning Models

Dmytro Chumachenko,[1] **Tetiana Dudkina**,[1] **Sergiy Yakovlev**,[1,2] **and Tetyana Chumachenko**[3]

[1]*Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute", 61070 Kharkiv, Ukraine*
[2]*Institute of Information Technology, Lodz University of Technology, 90-924 Lodz, Poland*
[3]*Epidemiology Department, Kharkiv National Medical University, 61000 Kharkiv, Ukraine*

Correspondence should be addressed to Dmytro Chumachenko; dichumachenko@gmail.com

Received 3 July 2023; Revised 20 September 2023; Accepted 9 December 2023; Published 20 December 2023

Academic Editor: Fei Hu

This study is centered around the COVID-19 pandemic which has posed a global health concern for over three years. It emphasizes the importance of effectively utilizing epidemic simulation models for informed decision-making concerning epidemic control. The challenge lies in appropriately choosing, adapting, and interpreting these models. The research constructs three statistical machine learning models to predict the spread of COVID-19 in specific regions and evaluates their performance using real COVID-19 incidence data. The paper presents short-term (3, 7, 14, 21, and 30 days) forecasts of COVID-19 morbidity and mortality for Germany, Japan, South Korea, and Ukraine. The precision of each model was scrutinized based on the type of input data used. Recommendations are provided on how various data sources can enhance the interpretation quality of machine learning models predicting infectious disease dynamics. The initial findings suggest the need for the comprehensive utilization of all available data, favoring cumulative data during holiday-rich periods and daily data otherwise. To minimize the absolute error, databases should be compiled using daily morbidity and mortality rates.

1. Introduction

The COVID-19 pandemic, caused by the spread of the SARS-CoV-2 coronavirus, has been a threat to global public health for almost three years. At the end of 2022, more than 640 million cases were registered worldwide, of which more than 6.6 million were fatal [1].

The global crisis caused by the pandemic has shown the critical role of information technology. The world has accelerated the digitalization of most areas of activity, including healthcare systems [2]. Research related to data-driven medicine is aimed at solving such problems as automated diagnostics [3], analysis of medical [4] and nonmedical interventions [5] to reduce the dynamics of morbidity, analysis of medical images [6], analysis of medical data [7], and modeling the dynamics of the epidemic process [8].

One of the essential tools for controlling the COVID-19 pandemic and other infectious diseases is modeling its dynamics, including forecasting. Forecasting the epidemic process dynamics allows us to predict how the incidence will develop and to conduct experimental studies to evaluate the effectiveness of various preventive measures.

Therefore, this study is aimed at building three statistical machine learning models for predicting the dynamics of COVID-19 in certain areas and at studying the performance of these models using experiments with actual COVID-19 incidence data.

To achieve the goal, the following tasks were formulated:

(1) To analyze models and methods for modeling the epidemic process of COVID-19

(2) To develop a predictive model for COVID-19 dynamics based on the logistic regression method

(3) To develop a predictive model of COVID-19 dynamics based on the decision tree method

(4) To develop a predictive model for COVID-19 dynamics based on the support vector regression method

(5) To evaluate the results of predicting the dynamics of COVID-19 using the developed models for data in various territories

(6) To compare the accuracy and adequacy of the developed models performed with the databases of different countries

(7) To analyze the performance of the developed models

The promising contribution of the research is twofold. Firstly, developing predictive models based on statistical machine learning methods will make it possible to analyze their effectiveness for modeling the epidemic process of COVID-19 and other infectious diseases. Secondly, developing predictive models based on statistical machine learning methods will make it possible to use them in public health practice in resource-limited settings to support decision-making on control measures to contain the dynamics of the COVID-19 pandemic. Thirdly, the analysis of models in terms of input data on morbidity will allow future research to be adjusted to model epidemic processes and apply models more effectively.

The further structure of the paper is the following: Section 2 provides an overview of models and methods of COVID-19 epidemic process simulation. Section 3 describes three regression approaches to COVID-19 morbidity forecasting, logistic regression, decision tree, and support vector regression, and describes the metrics used for models' performance evaluation. Section 4 describes the results of models' performance, estimation of developed models' adequacy, and forecasting accuracy. Section 5 discusses the perspective use of models and their limitations and analyzes the effectiveness of using different input data for forecasting. The conclusion describes the outcomes of the research.

Research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [9].

Preliminary research has been done for other statistical machine learning methods for modeling COVID-19: linear regression, lasso regression, ridge regression [10], random forest, K-nearest neighbors, and gradient boosting [11]. This study also explores the problem of input data in modeling epidemic processes.

2. Current Research Analysis

Epidemic process models have been used for over a century to control infectious disease dynamics, study disease behavior, and develop effective interventions to prevent epidemics. The global COVID-19 pandemic has stimulated a new round of research in this direction.

Compartmental models of the dynamics of the new coronavirus remain the most popular for practical application.

The authors of [12] study the theoretical foundations of the simplest SIR (susceptible-infected-recovered) model for modeling a new coronavirus. The authors explore the temporal evolution of different populations and track various significant parameters of the spread of the disease in different communities. However, the forecasts obtained in work are not sufficiently accurate. The work [13] presents a model for early prediction of COVID-19 based on the SIR structure, which allows predicting the situation for 700 days. The authors model the outbreak and possible scenarios for its termination with various types of control measures. The forecast presented by the authors can be retrospectively assessed as unreliable. However, the authors are investigating a scenario with a specialized treatment for COVID-19 that does not exist to date.

The study [14] is devoted to modeling COVID-19 in Canada using various models, including the SIR model. The constructed model does not assume the presence of asymptomatic cases, which is not valid and is an important characteristic that stimulates the spread of infectious diseases. The work [15] presents the SIR model for modeling the dynamics of COVID-19. The study calculates disease-free and endemic equilibrium, with global persistence calculated using the construction of the Lyapunov function and local persistence determined using the Jacobian matrix. The authors conclude that the nature of COVID-19 coincides with SARS, which is not valid.

The article's authors [16] explore the dynamics of the classical SIR model concerning COVID-19. The model considers the nonlinear removal rate, which depends on the number of hospital bed ratio. The authors conclude that the epidemic declines when the value of the basic reproductive number is less than one, but this is an epidemiological rule. In the study [17], the authors apply a modified SIR model to study the spread of COVID-19 in China. As a result, the authors argue that the increase in the number of control measures by the state has a positive effect on reducing the dynamics of COVID-19. However, the presented model does not allow us to draw such conclusions since social factors and the impact of such state control on other external factors that influence the development of the disease are not investigated.

In [18], the authors apply an implicit time-discrete SIR model that tracks transmission and recovery rates to predict the dynamics of COVID-19 in Fiji. The model does not take into account many factors that play an important role in the dynamics of infectious diseases, including the incubation period, the impact of control measures already taken, and the heterogeneity and openness of the population, as well as the difference between registered and real cases of the

disease. The authors of the article [19] extend the standard SIR model with the global dynamics of the COVID-19 pandemic. The proposed model was parameterized using a two-stage model fitting algorithm on data from six randomly selected US cities. Despite the increase in the accuracy of the model compared to the classical SIR model, it does not consider many factors that affect the dynamics of morbidity.

The study [20] discusses the numerical solution of the SIR model of the spread of COVID-19 using the Taylor matrix and the collocation method for Turkey. The model does not consider the dynamics of external factors, so the solutions obtained using the model are difficult to update to a changing situation. The paper [21] proposes an extension of the classical SIR model, the adaptive susceptible-infected-removed-vaccinated model with time-dependent transmission and removal rates. The authors propose a numerical solution to the inverse problem using the variational embedding method, which reduces the inverse problem to the problem of minimizing a well-formed functional to obtain the desired values. The model and its numerical solution are complex, making it difficult to introduce actual changes in disease dynamics into the model, such as changes in virulence and control measures.

Some researchers have extended the classic SIR model for modeling COVID-19 by adding new compartments. The authors of [22] extend the model with the exposed compartment. The model was applied to the early phase of the pandemic in Italy and was analyzed for sensitivity to determine the most critical parameters that have the most significant impact on the basic reproduction number. The article [23] describes the SEIAR model of COVID-19 with five compartments (susceptible-exposed-symptomatic-asymptomatic-recovered/removed). As a result, the authors conclude that the virus is highly contagious for people after the age of 45 years and has low susceptibility to the virus up to 14 years of age. The authors of [24] expand the SEIR model by introducing the characteristics of age groups, symptomatic and asymptomatic disease development, and vaccinated and unvaccinated population. The results show that, despite the high level of detail, the model cannot predict changes in epidemic dynamics caused by the emergence of new strains or the introduction of new control measures.

The work [25] proposes an extended specialized SEIR model for COVID-19 modeling called SEAHIR (susceptible-exposed-asymptomatic-hospitalized-isolated-removed). In the proposed model, the "infected" compartment is divided into "asymptomatic," "isolated," and "hospitalized." The model also considers the impact of nonpharmaceutical interventions such as physical distancing and different testing strategies. The paper [26] presents a hybrid compartmental model for studying the evolution of the COVID-19 pandemic in Italy. The model proposed by SEIRDV includes six compartments, considering the vaccinated population. At the same time, the representation of infection is presented both as a linear and as an exponential piecewise continuous function. The results show that different levels of vaccination give similar infection curves.

All the models described using the compartmental approach have several disadvantages, including modeling for homogeneous and closed populations, the impossibility of taking into account all the factors influencing the dynamics of the epidemic process and the complexity of systems of differential equations describing the system, and the difficulty of making changes to the model, adapting it to reality. These shortcomings affect the adequacy and accuracy of the model, which does not simulate the actual situation with the incidence of COVID-19 effectively.

Higher accuracy is shown by predictive models based on machine learning methods.

The paper [27] presents a predictive model for COVID-19 in India based on an artificial neural network with a long short-term memory (LSTM) architecture. The model predicts the total number of cases, recoveries, and deaths of COVID-19 over 80 days. The model showed an accuracy of 95.46%. The authors of [28] propose models based on recurrent neural networks such as LSTM, bidirectional LSTM, and encoder-decoder LSTM models for multistep (short term) COVID-19 infection forecasting. Using the presented models, a forecast for two months ahead is built based on data on the first and second waves of incidence. However, the authors note the difficulties with modeling associated with the unreliability of the data and the difficulty of considering factors such as population density, logistics, social aspects, and lifestyle of the studied population.

The authors of [29] proposed a deep learning approach that includes recurrent neural networks and LSTM networks for predicting the probable numbers of COVID-19 cases. For a pilot study, data from the European Centre for Disease Prevention and Control on the incidence of COVID-19 in Malaysia, Morocco, and Saudi Arabia were used. The results showed an accuracy of 98.58% for the LSTM model and 93.45% for the RNN model in predicting new COVID-19 cases over seven days.

The authors of [30] employed Bayesian optimization to tune the Gaussian process regression (GPR) hyperparameters to develop an efficient GPR-based model for forecasting the recovered and confirmed COVID-19 cases. The authors show the superiority of the proposed approach in comparison with other time series forecasting models. However, only one dataset was used for forecasting, so the model's performance may differ depending on the area where the simulation is carried out. The authors of [31] propose three deep learning models, including CNN, LSTM, and the CNN-LSTM, to predict the dynamics of COVID-19 in Brazil, India, and Russia. The authors note that various socioeconomic, geographic, and political reasons may influence public policy in implementing control measures to contain the epidemic dynamics.

Despite the high accuracy of COVID-19 predictive models based on deep learning, they cannot always be applied in resource-limited settings. The requirements for high computing power that such models impose are difficult to meet directly in public health institutions. Therefore, this paper proposes an analysis of COVID-19 predictive models based on statistical machine learning methods.

3. Materials and Methods

As part of this study, three machine learning models were built to predict the dynamics of COVID-19. The models

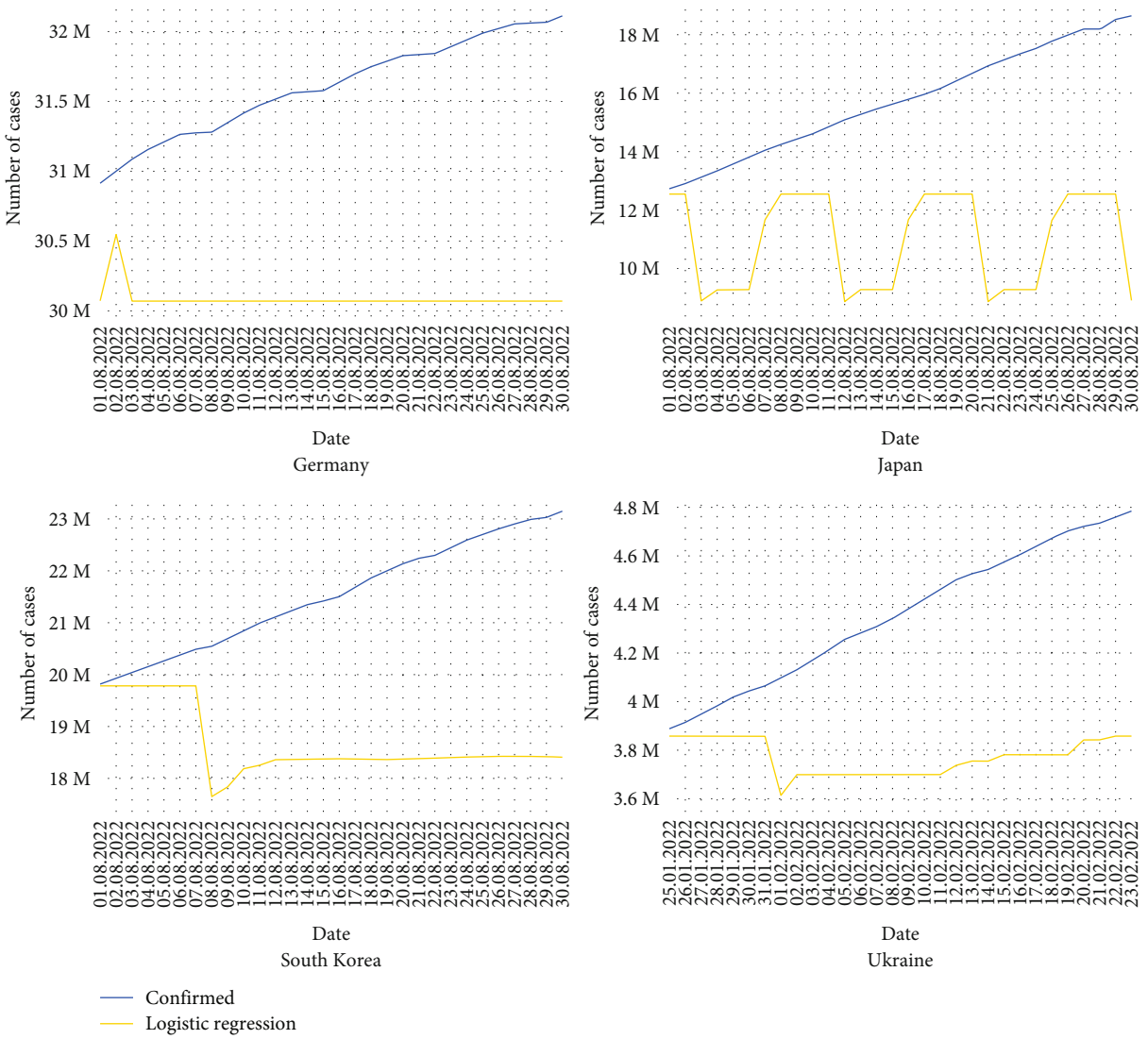

Figure 1: Forecasting of COVID-19 cumulative new cases by logistic regression model.

are based on regression methods: logistic regression, decision tree, and support vector regression.

Regression analysis is an analytical method of statistical machine learning that calculates the estimated relationship between a dependent variable and one or more independent variables [32]. Regression analysis finds the model relationships between selected variables and model-based predictive values. Regression analysis uses the chosen estimation method, the dependent variable, and one or more independent variables to create an equation that estimates the values of the dependent variable.

3.1. Logistic Regression. Logistic regression is a data analysis technique that allows finding the relationship between two data factors [33]. This relationship is used to predict the value of one of these factors based on the other. To do this, a dependent variable y is introduced, which takes the values 0 and 1, and a set of independent variables $x_1, \cdots, x_n$. Based on these values, it is necessary to calculate the probability of accepting one or another value of the dependent variable.

Let objects be defined by n numerical features:

$$f_i : X \longrightarrow R, j = 1..n, \tag{1}$$

and the space for feature descriptions, in this case

$$X = R^n. \tag{2}$$

Y is a finite set of class labels, and a training set of "object-factor" pairs is given as follows:

$$X^m = \{(x_1, y_1), \cdots, (x_m, y_m)\}. \tag{3}$$

Consider the case of two classes: $Y = \{-1,+1\}$. In logistic regression, a linear classification algorithm is built:

$$\alpha : X \longrightarrow Y. \tag{4}$$

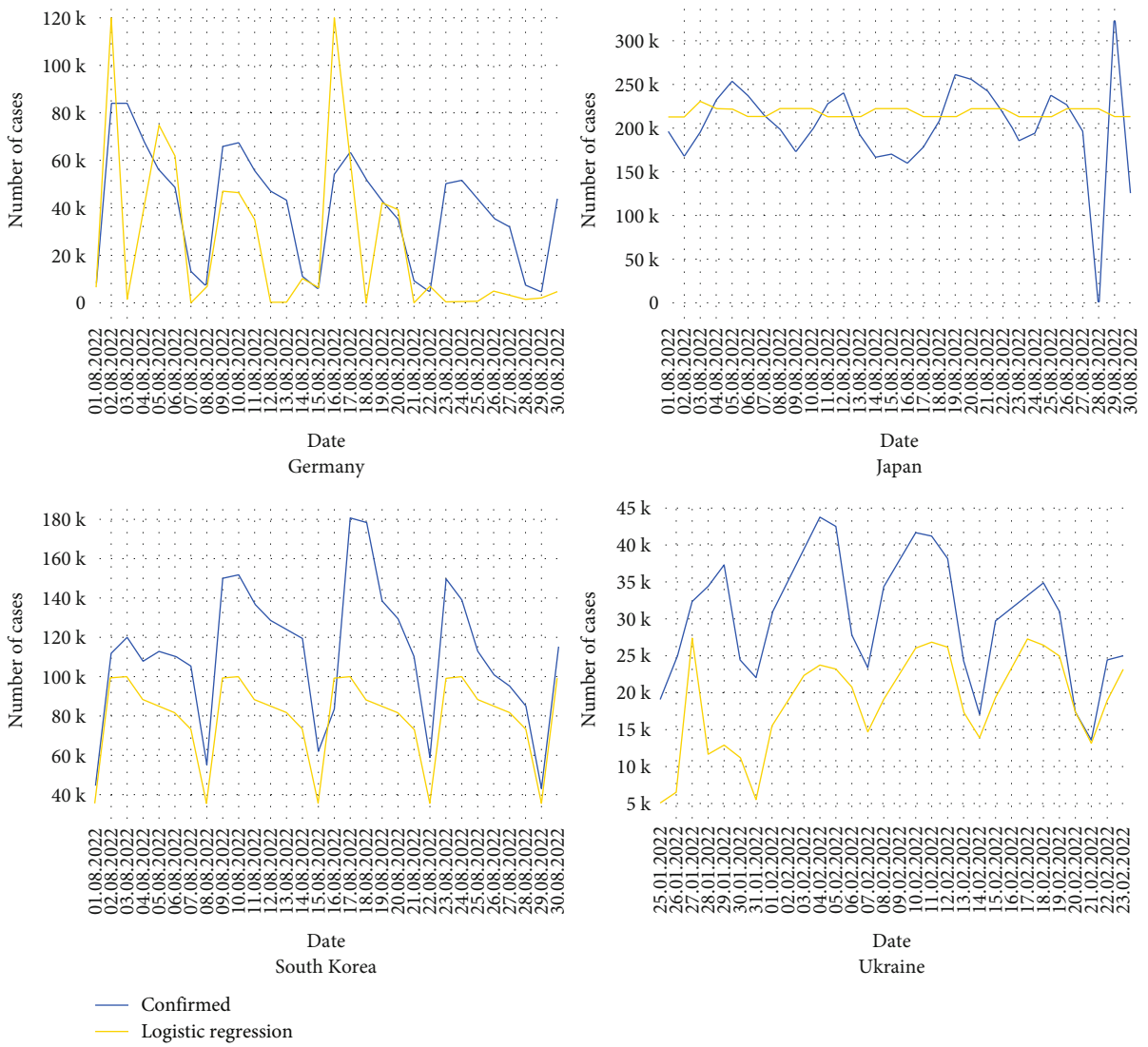

Figure 2: Forecasting of COVID-19 daily new cases by logistic regression model.

The following kind is shown:

$$a(x, w) = \text{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \text{sign}\ \langle x, w\rangle, \quad (5)$$

where w_j is the weight of the j^{th} feature, w_0 is the decision threshold, $w = (w_0, \cdots, w_n)$ is the weight vector, and $\langle x, w\rangle$ is the scalar product of the feature description of the object and the weight vector. At the same time, it is assumed that a zero sign is artificially introduced:

$$f_0(x) = -1. \quad (6)$$

The task of training a linear classifier is to adjust the weight vector w based on the X^m sample. In logistic regression, for this, the problem of minimizing empirical risk with a loss function of a special type is solved:

$$Q(w) = \sum_{i=1}^{m} \ln(1 + \exp\ (-y_i\langle x_i, w\rangle)) \longrightarrow \min_w. \quad (7)$$

After finding the solution w, it becomes possible to estimate the posterior probabilities of its belonging to the classes:

$$\mathbb{P}\{y|x\} = \sigma(y\langle x, w\rangle), y \in Y, \quad (8)$$

where

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \quad (9)$$

The following are the advantages of the logistic regression method:

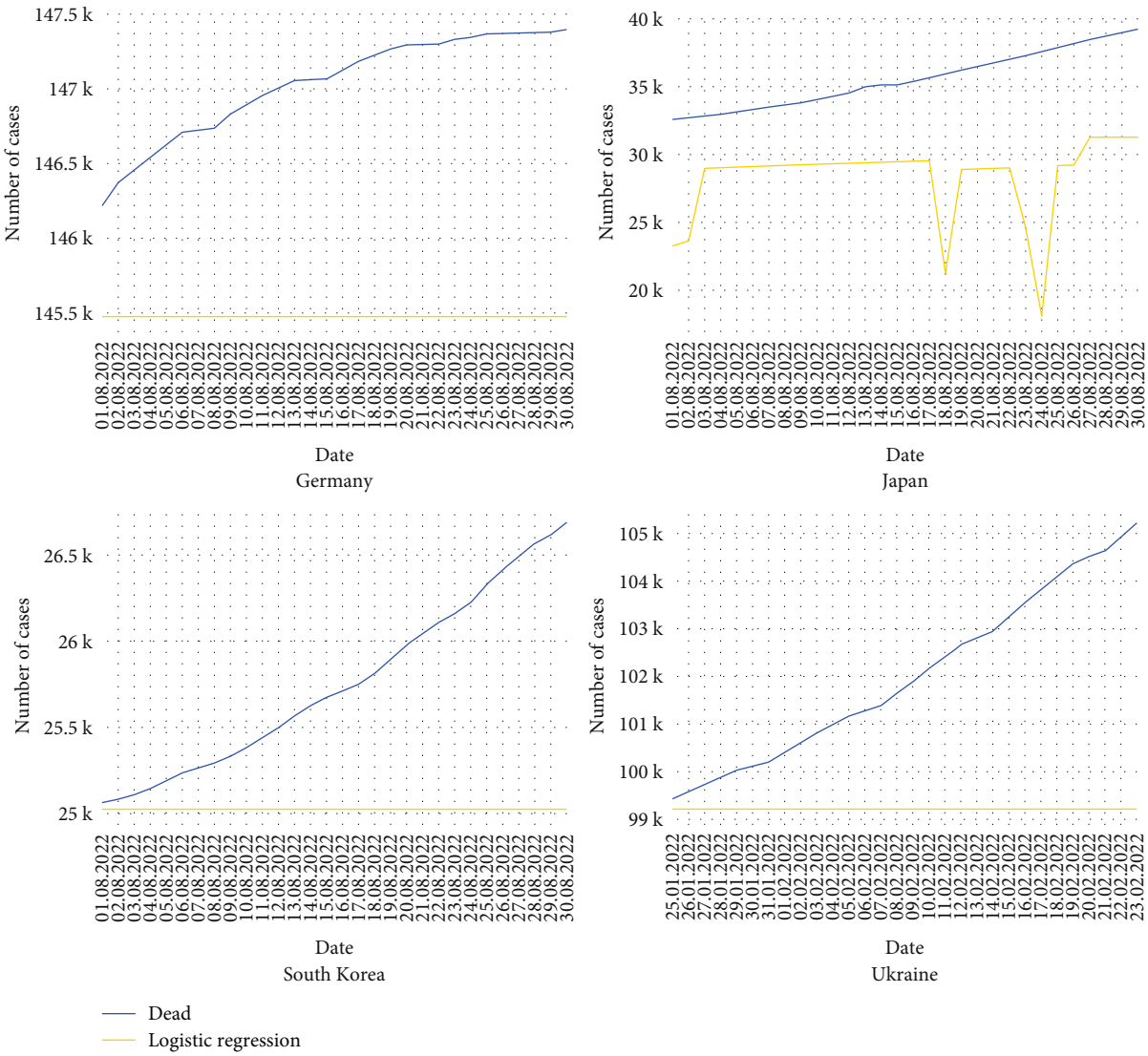

Figure 3: Forecasting of COVID-19 cumulative fatal cases by logistic regression model.

(i) Logistic regression models are mathematically less complex than other machine learning methods. It also makes troubleshooting easier

(ii) Logistic regression models allow developers to better understand internal processes than other machine learning methods

(iii) Logistic regression models can process large amounts of data at high speed because they require less computing power

The following are the disadvantages of the logistic regression method:

(i) The model handles a large number of categorical variables poorly

(ii) For the model to work, it is necessary to transform nonlinear functions

3.2. Decision Tree. Decision trees are a nonparametric supervised learning method used for classification and regression [34]. The goal of the method is to create a model that predicts the value of the target variable by learning simple decision rules derived from the characteristics of the data. If the target variable has continuous values, decision trees allow for establishing the dependence of the target variable on independent variables.

A decision tree is a hierarchical tree structure consisting of "if-then" decision rules that can be formulated in natural language.

The method recursively divides the original dataset into subsets that become more and more homogeneous concerning certain features, resulting in a tree-like hierarchical structure. The division is carried out based on traditional logical rules in the form "If A, then B", where A is some logical condition and B is the procedure for dividing the subset into two parts, for one of which condition A is true, and for the other, it is false.

To construct a tree, it is necessary to set the quality functional based on which the sample is split at each step. Let R_m

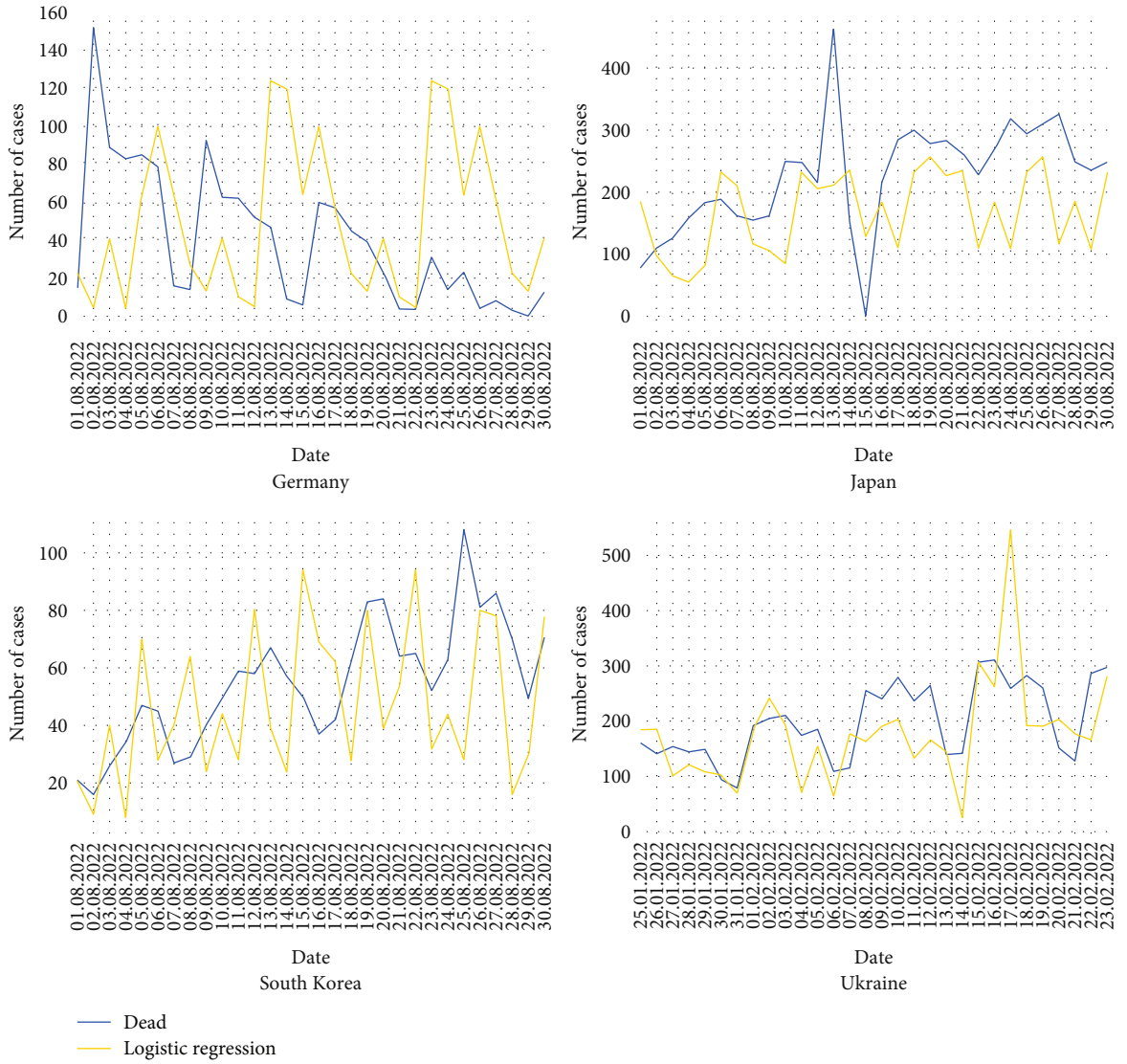

Figure 4: Forecasting of COVID-19 daily fatal by logistic regression model.

be the set of objects that fall into the vertex split at this step; R_t and R_r are the objects that fall into the left and right subtrees for a given predicate. Then, we will use the following functionals:

$$Q(R_m, j, s) = H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_r), \quad (10)$$

where $H(R)$ is the informativeness criterion that evaluates the distribution quality of the target variable among the objects of the set R. The smaller the diversity of the target variable, the lower the value of the informativeness criterion should be.

In each leaf, the tree will produce a real number. Based on this, it is possible to evaluate the quality of the set of R topic objects:

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c), \quad (11)$$

where $L(y, c)$ is some loss function.

To build a regression model, we choose the square of the deviation as a loss function. In this case, the informativeness criterion will look like this:

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - c)^2. \quad (12)$$

To build a regression model, we choose the square of the deviation as a loss function. In this case, the informativeness criterion will look like this:

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_j, y_j) \in R} y_j \right)^2. \quad (13)$$

The following are the advantages of the decision tree method:

(i) Easy interpretability and visualization capability

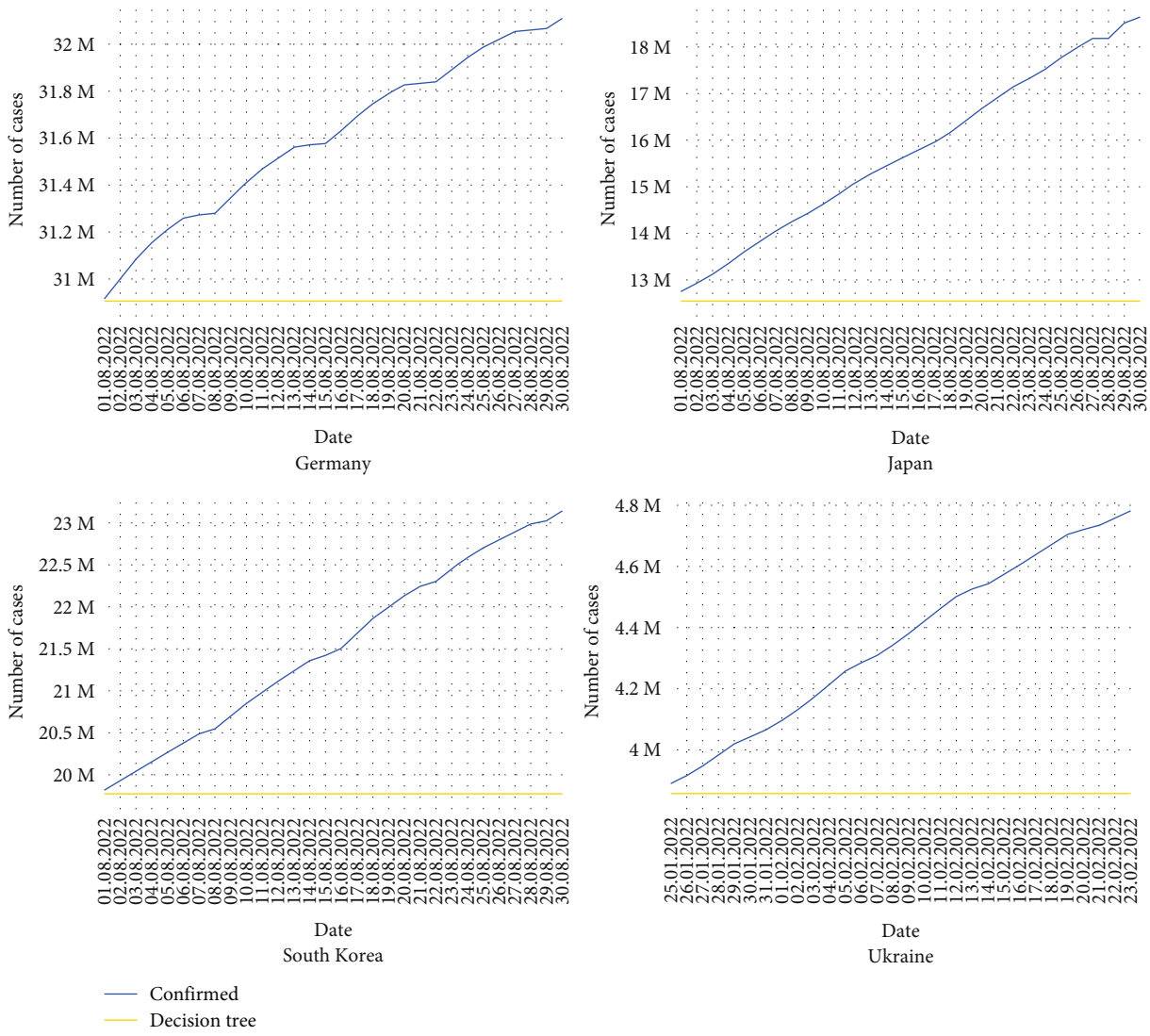

Figure 5: Forecasting of COVID-19 cumulative new cases by decision tree model.

(ii) Only a little preparation is required

(iii) The cost of using a tree is logarithmic in the number of data points used to train the tree

(iv) The decision tree model can handle both numerical and categorical data

The following are the disadvantages of the decision tree method:

(i) The model can create overly complex trees that do not generalize well

(ii) Decision trees can be unstable, and small changes in the data can lead to a completely different tree

(iii) The optimal decision tree learning problem is NP-complete in terms of several aspects of optimality and even for simple concepts. Therefore, practical algorithms for learning decision trees are based on heuristic algorithms, such as the greedy algorithm, in which locally optimal decisions are made at each node

(iv) It is recommended that the dataset be balanced before fitting to the decision tree

3.3. Support Vector Regression. The basis of the support vector machine for regression problems is the search for a hyperplane, in which the risk in a multidimensional space will be minimal [35].

The support vector machine estimates the coefficients by minimizing the quadratic loss. If the predicted value falls within the hyperplane region, then the loss is zero. Otherwise, the losses equal the difference between the predicted and actual values.

In the support vector machine for the regression problem, it is necessary to evaluate the functional dependence of the dependent variable y on the set of independent

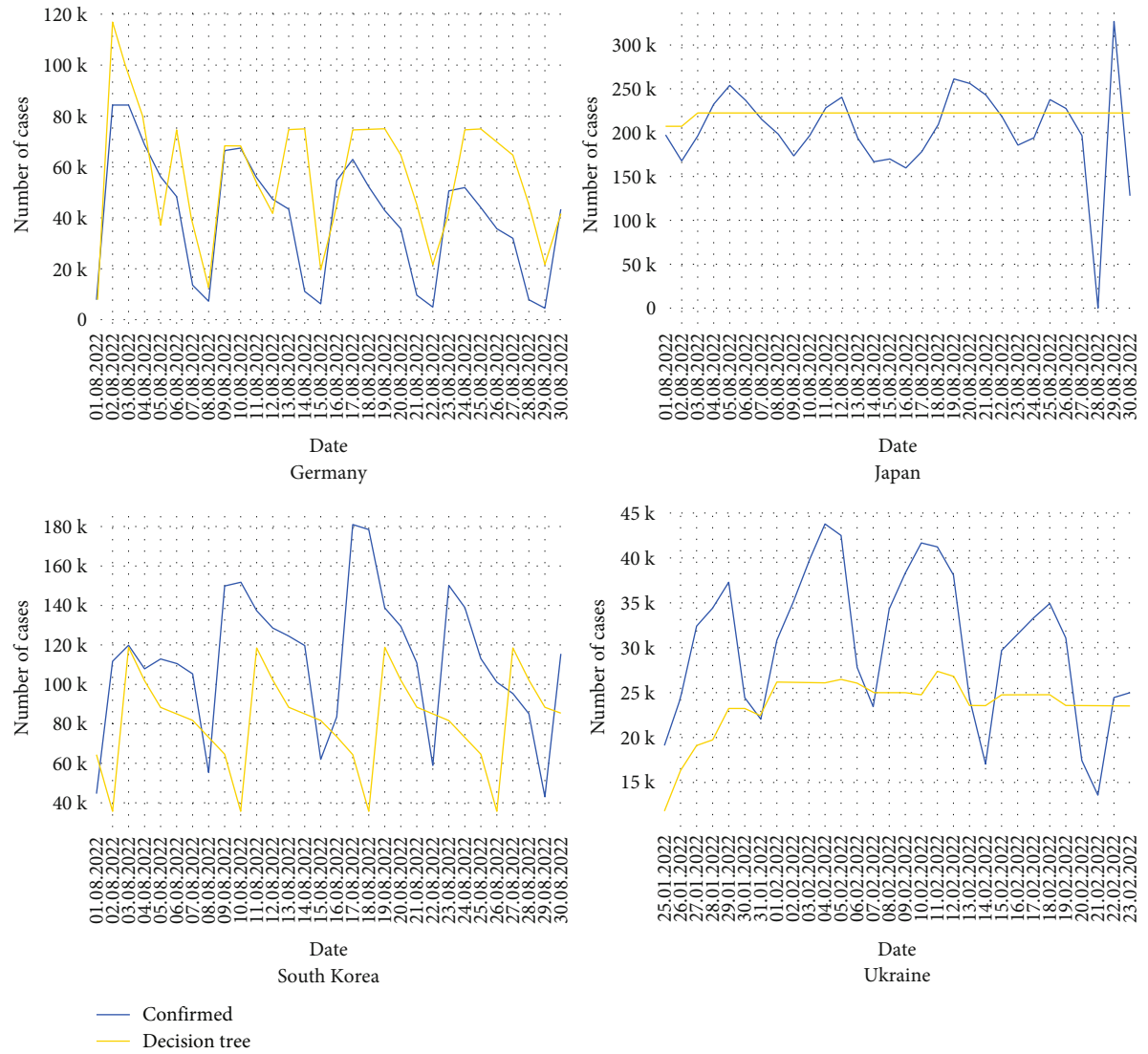

Figure 6: Forecasting of COVID-19 daily new cases by decision tree model.

variables x. To do this, the relationship between the independent and dependent variables is determined by a deterministic function and the addition of additive noise:

$$y = f(x) + \text{noise}. \tag{14}$$

In this case, it is necessary to find a functional form for f that can correctly predict new values. Functional dependence is sought by training the model on a sample population. In this study, we determined the error function by the formula:

$$\frac{1}{2}w^T w - C\left(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)\right). \tag{15}$$

The function is minimized under the condition:

$$\left(w^T\varphi(x_i) + b\right) - y_i \le \varepsilon + \xi_i, \tag{16}$$

$$y_i - \left(w^T\varphi(x_i) + b_i\right) \le \varepsilon + \xi_i^*, \tag{17}$$

$$\xi_i, \xi_i^* \ge 0, i = 1, \cdots, N, \varepsilon \ge 0. \tag{18}$$

Advantages of the support vector machine are as follows:

(i) The principle of the optimal separating hyperplane leads to the maximization of the width of the separating strip

(ii) The support vector machine is equivalent to a two-layer neural network in which the number of neurons in the hidden layer is determined automatically as the number of support vectors

(iii) The convex quadratic programming problem is well-studied and has a unique solution

Disadvantages of the support vector machine are as follows:

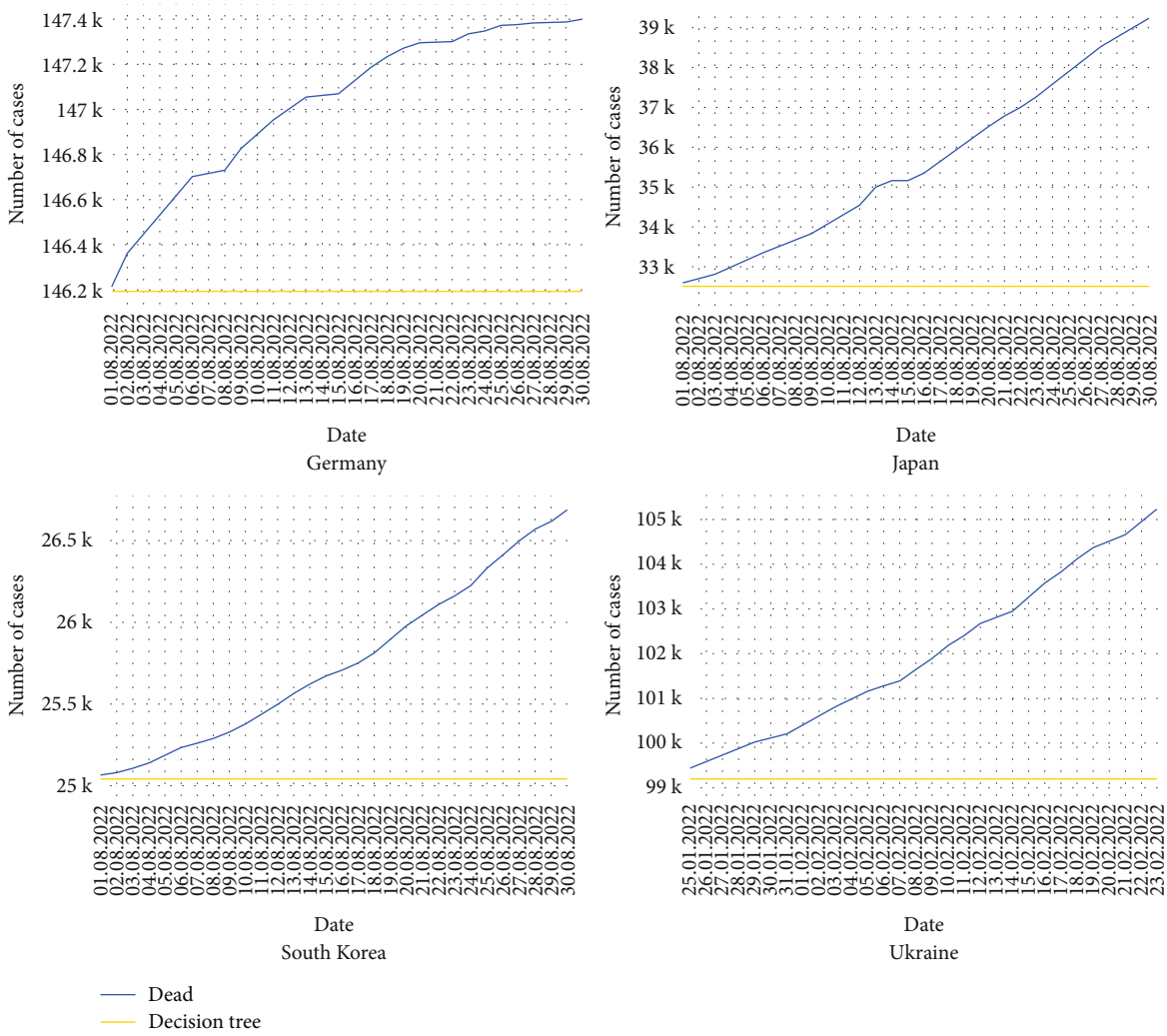

Figure 7: Forecasting of COVID-19 cumulative fatal cases by decision tree model.

(i) There is no feature selection in the method

(ii) The constant must be selected using cross-validation

(iii) Outliers in the initial data become reference objects-violators and directly affect the construction of the separating hyperplane

3.4. Models' Performance Evaluation Metrics. We used the following metrics to evaluate models' performance.

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \tag{19}$$

where y_i is the predicted value, x_i is the observed value, and n is the number of observations.

Relative absolute error (RAE) is expressed as a ratio, comparing a mean error to errors produced by a trivial or naïve model:

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{\sum_{i=1}^{n} |\overline{y_i} - x_i|}, \tag{20}$$

where y_i is the predicted value, x_i is the observed value, $\overline{y_i}$ is the average of the predicted values, and n is the number of observations.

Mean absolute percentage error is a measure of prediction accuracy, which expresses the accuracy as a ratio defined by formula:

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right|, \tag{21}$$

where y_i is the predicted value, x_i is the observed value, and n is the number of observations.

As an accuracy metric, we used a difference of MAPE from 100%:

$$\mathrm{Accuracy} = 100\% - \mathrm{MAPE}. \tag{22}$$

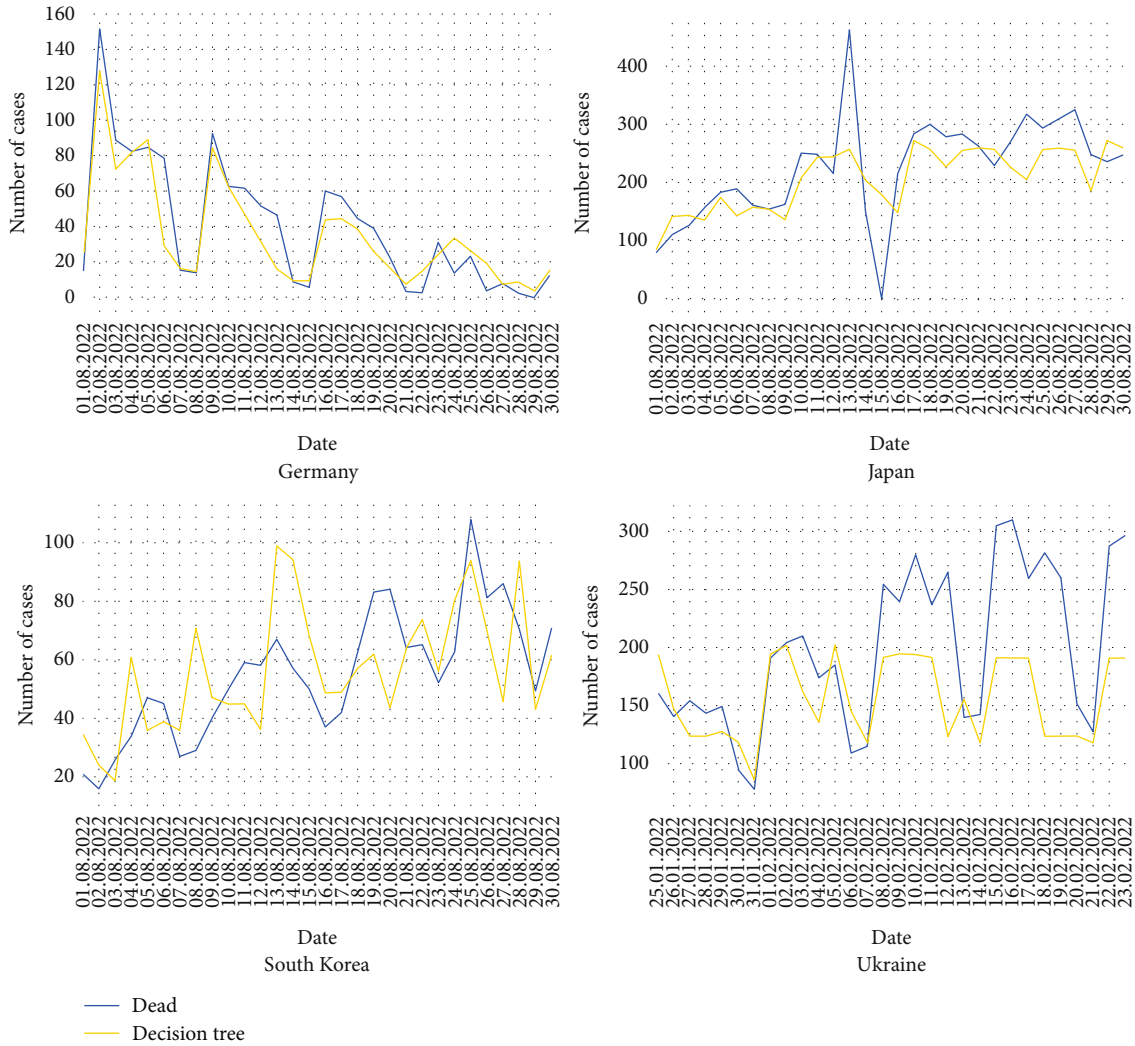

Figure 8: Forecasting of COVID-19 daily fatal cases by decision tree model.

4. Results

The program realization of the COVID-19 models was performed using Python programming language. An experimental investigation of the models was carried out on four types of data provided by World Health Organization Coronavirus Dashboard [35]: daily new cases, daily fatal cases, cumulative new cases, and cumulative fatal cases. We used data for Germany, Japan, South Korea, and Ukraine. These countries were selected due to the different nature of the dynamics of the epidemic process, various control measures implemented by governments, and various social factors influencing the dynamics of COVID-19. The forecast was calculated for 3, 7, 14, 21, and 30 days. For Germany, Japan, and South Korea, the forecasting period was from August 1, 2022, to August 30, 2022, and for Ukraine, from January 25, 2022, to February 23, 2022. This is due to the full-scale Russian invasion of Ukraine, which affected the dynamics of the epidemic process of infectious diseases, including COVID-19.

Historical morbidity and mortality data available before the forecast periods were utilized to train the machine learning models. This ensured that the models were well-acquainted with the past trends and patterns of the disease's spread in the respective countries. The forecast period, on the other hand, was exclusively reserved for testing the models' predictions. This approach maintained a clear demarcation between training and testing data, ensuring the integrity and validity of the model evaluation process.

The forecasting results show the retrospective forecasted dynamics of COVID-19 epidemic process dynamics in the selected area.

4.1. Forecasting Results Using a Logistic Regression Model. Figure 1 shows the results of forecasting of cumulative new cases of COVID-19 in the selected areas with logistic regression model. Figure 2 shows the results of forecasting of daily new cases of COVID-19 in the selected areas with logistic regression model. Figure 3 shows the results of forecasting of cumulative fatal cases of COVID-19 in the selected areas with logistic regression model. Figure 4 shows the results of forecasting of daily fatal cases of COVID-19 in the selected areas with logistic regression model.

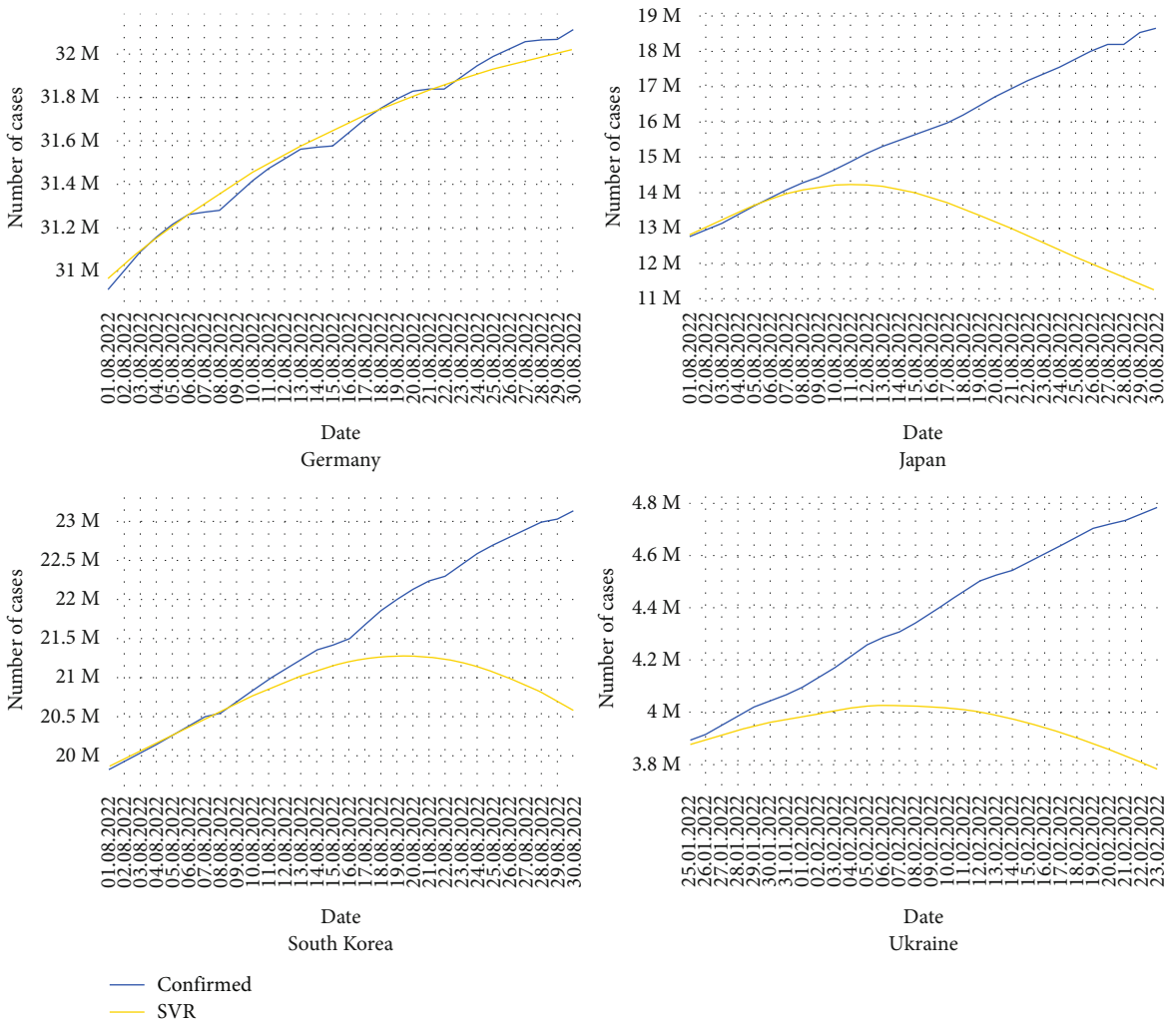

Figure 9: Forecasting of COVID-19 cumulative new cases by support vector regression model.

4.2. Forecasting Results Using a Decision Tree Model. Figure 5 shows the results of forecasting of cumulative new cases of COVID-19 in the selected areas with decision tree model. Figure 6 shows the results of forecasting of daily new cases of COVID-19 in the selected areas with decision tree model. Figure 7 shows the results of forecasting of cumulative fatal cases of COVID-19 in the selected areas with decision tree model. Figure 8 shows the results of forecasting of daily fatal cases of COVID-19 in the selected areas with decision tree model.

4.3. Forecasting Results Using a Support Regression Model. Figure 9 shows the results of forecasting of cumulative new cases of COVID-19 in the selected areas with support vector regression model. Figure 10 shows the results of forecasting of daily new cases of COVID-19 in the selected areas with support vector regression model. Figure 11 shows the results of forecasting of cumulative fatal cases of COVID-19 in the selected areas with support vector regression model. Figure 12 shows the results of forecasting of daily fatal cases of COVID-19 in the selected areas with support vector regression model.

4.4. Performance of Logistic Regression Model. Table 1 shows MAE values of logistic regression models for confirmed cases of COVID-19 in selected territories.

Table 2 shows MAE values of logistic regression models for fatal cases of COVID-19 in selected territories.

Table 3 shows RAE values of logistic regression models for confirmed cases of COVID-19 in selected territories.

Table 4 shows RAE values of logistic regression models for fatal cases of COVID-19 in selected territories.

Table 5 shows MAPE values of logistic regression models for confirmed cases of COVID-19 in selected territories.

Table 6 shows MAPE values of logistic regression models for fatal cases of COVID-19 in selected territories.

4.5. Performance of Decision Tree Model. Table 7 shows MAE values of decision tree models for confirmed cases of COVID-19 in selected territories.

Table 8 shows MAE values of decision tree models for fatal cases of COVID-19 in selected territories.

Table 9 shows RAE values of decision tree models for confirmed cases of COVID-19 in selected territories.

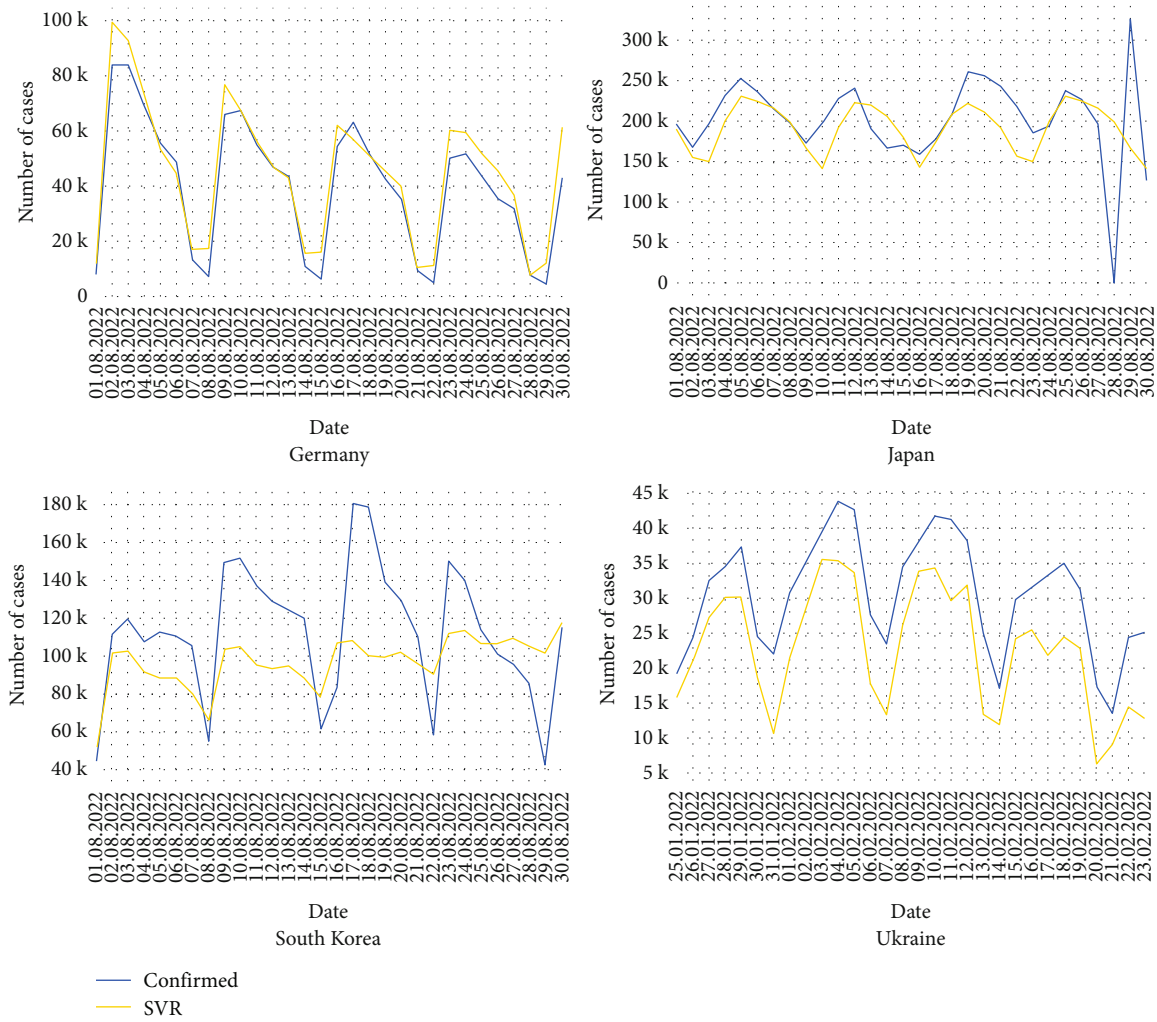

Figure 10: Forecasting of COVID-19 daily new cases by support vector regression model.

Table 10 shows RAE values of decision tree models for fatal cases of COVID-19 in selected territories.

Table 11 shows MAPE values of decision tree models for confirmed cases of COVID-19 in selected territories.

Table 12 shows MAPE values of decision tree models for fatal cases of COVID-19 in selected territories.

4.6. Performance of Support Vector Regression Model. Table 13 shows MAE values of support vector regression models for confirmed cases of COVID-19 in selected territories.

Table 14 shows MAE values of support vector regression models for fatal cases of COVID-19 in selected territories.

Table 15 shows RAE values of support vector regression models for confirmed cases of COVID-19 in selected territories.

Table 16 shows RAE values of support vector regression models for fatal cases of COVID-19 in selected territories.

Table 17 shows MAPE values of support vector regression models for confirmed cases of COVID-19 in selected territories.

Table 18 shows MAPE values of support vector regression models for fatal cases of COVID-19 in selected territories.

5. Discussion

The emerging virus SARS-CoV-2, which humanity learned about in December 2019, quickly spread around the globe, causing the COVID-19 pandemic. During the three years of the pandemic, the disease has claimed more than 6.5 million lives, and more than 663 billion cases have been registered worldwide. Each country has chosen its tactics in the fight against COVID-19. The measures included isolation and treatment of patients, wearing masks in crowded places, and physical distancing. Effective vaccines produced using various technologies were developed and implemented relatively quickly and began to be implemented. However, despite this, vaccination coverage against COVID-19 among the population of different countries is still needed to reach the required level. It did not allow for stopping the circulation of the pathogen among the population. The proportion of vaccines vaccinated with one, two, and three doses differ

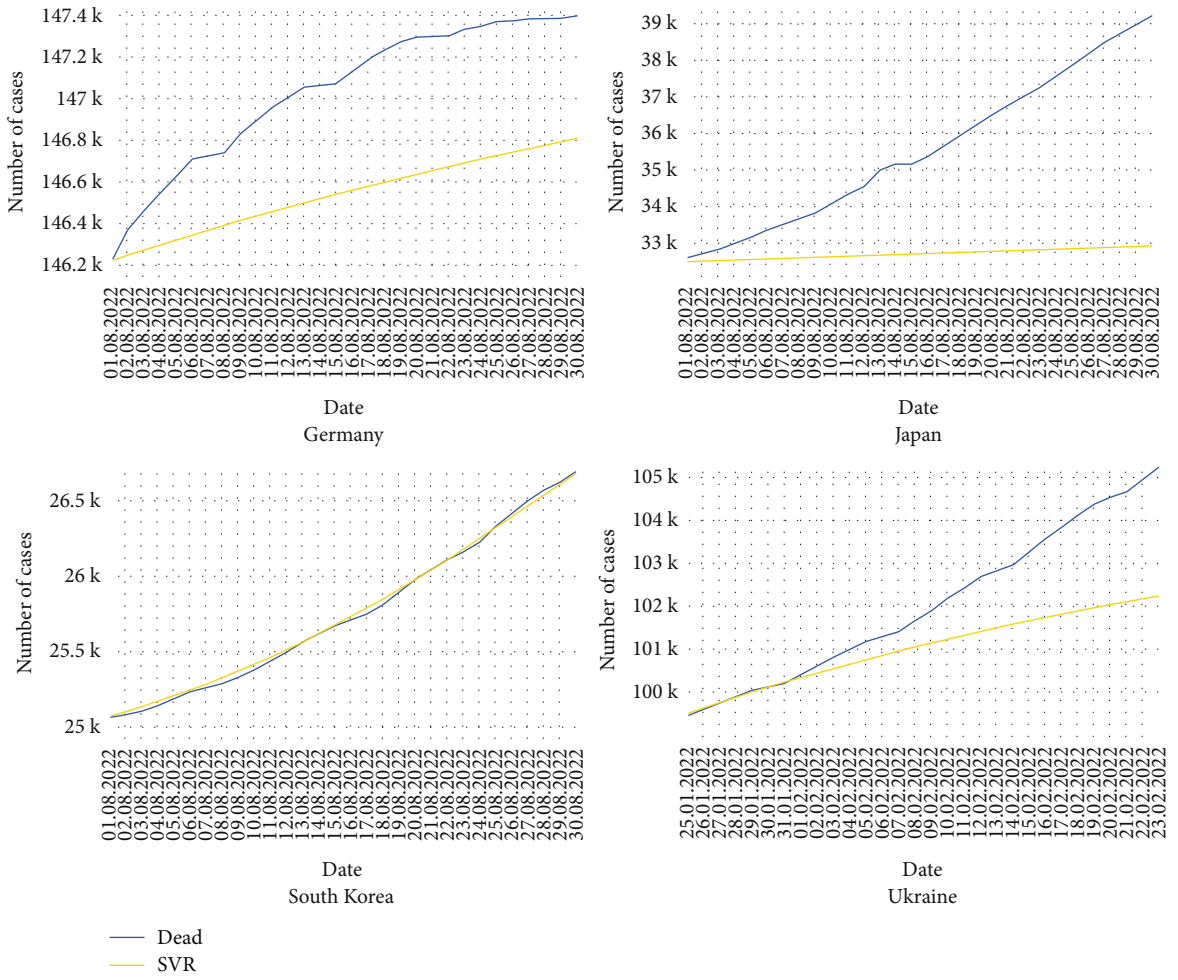

Figure 11: Forecasting of COVID-19 cumulative fatal cases by support vector regression model.

significantly in different countries; low vaccination coverage creates conditions for selecting new strains of the virus with new mutations and makes it difficult to fight infection [36].

Mathematical models have become a good tool for predicting the development of the COVID-19 pandemic and helping to make adequate management decisions to contain the pandemic. Various models have been developed [37–39]. However, each of them had some drawbacks, such as the impossibility of taking into account all the factors that negatively or positively affect the development of the COVID-19 epidemic process, that does not take into account the heterogeneity of the human population and differences in the structure of the population in different territories, etc.

We have built three models based on machine learning to predict the dynamics of the spread of COVID-19—logistic regression, decision tree, and support vector regression. The forecast was calculated for 3, 7, 14, 21, and 30 days. The timing of the forecast was not chosen by chance. It is clear that if there is a sharp deterioration in the epidemic situation, an increase in morbidity and mortality from COVID-19 is predicted on day 30; it is necessary to correct preventive measures as soon as possible. Intermediate forecasts for 3, 7, 14, and 21 days make it possible to control the adequacy of the tactics for preventing the incidence and containing the pathogen's spread. In addition, the weekly interval makes it possible to smooth out fluctuations in the number of registered cases associated with a lower population seeking medical care on weekends and holidays and a sharp increase in case registration immediately after the weekend. Forecasting for 3 days will show the trend in the dynamics of the epidemic process but will not reflect the changes associated with introducing additional preventive measures.

For the analysis of models, countries with different cultures, medical care organization, surveillance, chosen tactics to combat the COVID-19 pandemic, and other factors influencing the development of the pandemic were selected. We chose four countries—Germany, Japan, South Korea, and Ukraine. For the first three countries, the forecast was built for the period from August 1, 2022, to August 30, 2022, and for Ukraine, from January 25, 2022, to February 23, 2022, because it is impossible to check the accuracy of the forecast for August because full-scale Russian invasion of Ukraine led to the destruction and destruction of hospitals, a decrease in the number of medical personnel, limited access to medical care for the population, and part of the territory of Ukraine was occupied, which did not allow registering the incidence in these territories.

In our analysis, a noteworthy discrepancy emerged in the accuracy of the forecast data for Japan when juxtaposed with

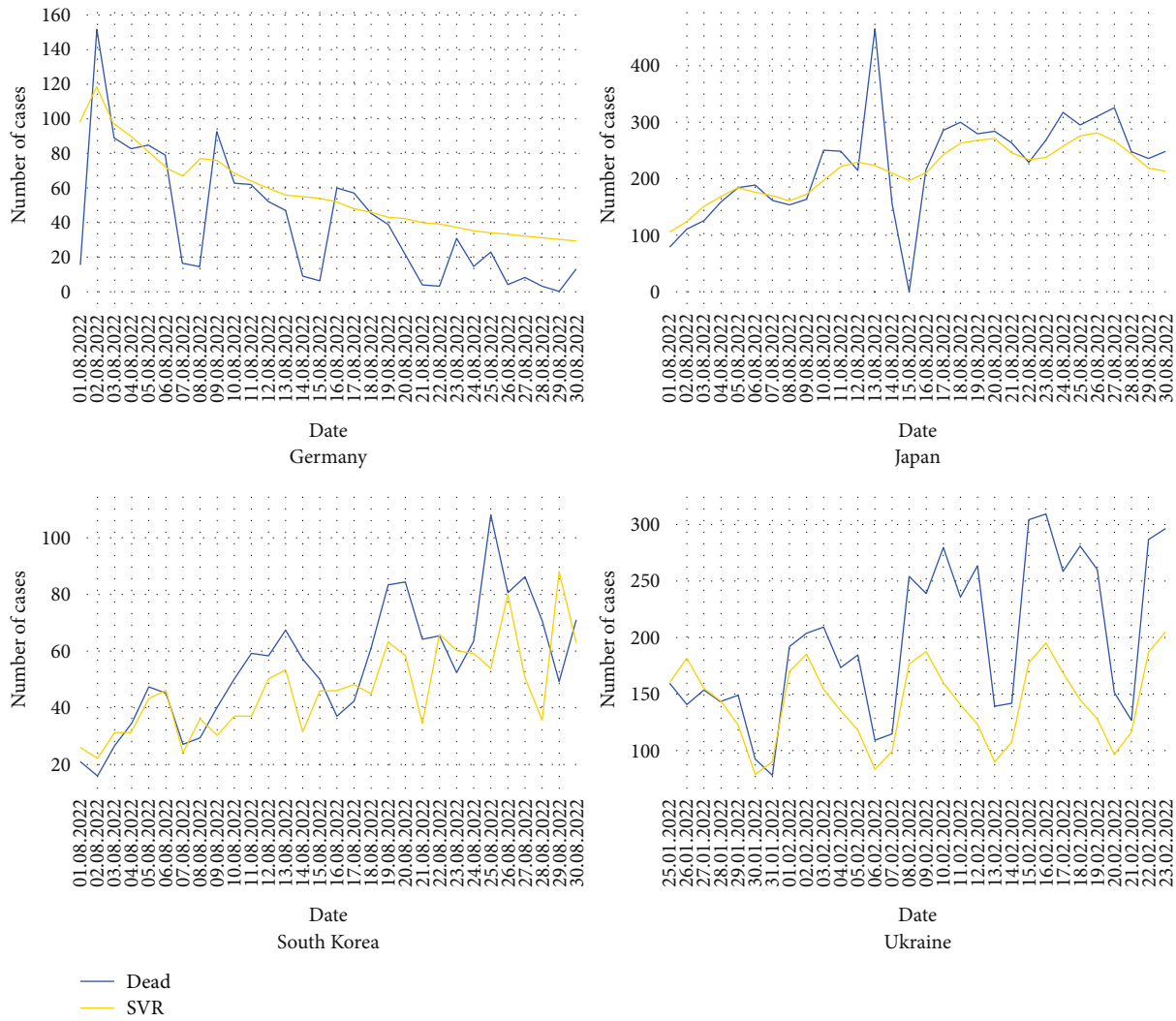

Figure 12: Forecasting of COVID-19 daily fatal cases by support vector regression model.

Table 1: MAE values of logistic regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	857718.6	286153.96	108432	55505.78	48276.21	18819.15	77411.48	5451.58
Test 3	772078.67	1587192	159047.67	33767.33	39960.33	31894.67	13655	14126
Train 7	1162616.09	445712.26	213582.58	136885.78	48782.19	20648.93	65870.27	5232.94
Test 7	988841.57	2865681.57	382687.57	94103.43	28426.29	23223.57	21234.43	15376.14
Train 14	1444834.85	551208.48	546825.8	241053.19	50805	22077.31	66760.47	6004.26
Test 14	1183490.57	3308088.5	1601407.64	279010.07	24988.43	27008.71	32216.71	15638.57
Train 21	1971303.35	607461.92	798955.03	291836.94	53295.33	23092.32	74409.74	6541.27
Test 21	1341054.1	3797773.29	2219878.71	421748.81	23132.67	30210.1	38164.95	14651.62
Train 30	1952622.63	688047.48	1134894.41	354217.3	53247.71	18295.2	74542.03	6966.61
Test 30	1534048	4897131.16	2920608.16	565430.74	24795.03	39843.06	32511.26	11527.39

that of Germany and South Korea. Several factors underpin this observed variation.

Firstly, the healthcare infrastructure and reporting mechanisms differ across countries. Germany and South Korea have been globally recognized for their robust healthcare systems and efficient disease surveillance mechanisms. Their rapid response to the pandemic, extensive testing, and meticulous data recording contributed to a more

Table 2: MAE values of logistic regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	10594.46	4389.47	779.36	2137.49	93.7	37.72	43.5	186.8
Test 3	875	7401	60.33	221.33	68	59.33	7	22.67
Train 7	14431.7	5036.62	1455.71	4339.38	96.37	41.63	41.45	186.37
Test 7	1046.57	5552.86	129.57	511.43	53.29	67.86	14.43	27.29
Train 14	17803.87	6076.29	2126.05	6719.15	102.1	46	41.38	187.03
Test 14	1252.86	5299.07	275.21	1036.36	55.43	78.07	19.43	31.79
Train 21	18747.84	6258.6	2283.06	8983.12	109.16	49.14	43.62	183.2
Test 21	1413.81	6149.38	453.9	1665.52	45.05	76.19	21.86	44.38
Train 30	20143.5	6774.49	2284.55	11098.24	115.11	48.62	45.96	185.85
Test 30	1567.74	7479.45	762.58	2716.65	45.1	83.74	24.1	57.29

Table 3: RAE values of logistic regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.099456	0.096796	0.016089	0.097866	0.863515	0.728601	1.374757	1.029529
Test 3	13.738489	12.770641	2.036247	2.242518	1.178802	2.506698	0.431483	3.609915
Train 7	0.134811	0.150769	0.03169	0.241352	0.872565	0.799443	1.169796	0.98824
Test 7	9.055444	7.346437	1.996374	1.767298	1.159893	1.000194	1.29776	2.164758
Train 14	0.167535	0.186455	0.081135	0.425016	0.908747	0.854744	1.185605	1.133903
Test 14	7.280291	4.480851	3.934655	2.515279	1.140845	1.144672	1.592753	2.101503
Train 21	0.228582	0.205483	0.118545	0.514556	0.953292	0.894041	1.321449	1.235317
Test 21	5.813652	3.54859	3.493077	2.35823	1.10771	1.08358	1.494972	1.998267
Train 30	0.226416	0.232743	0.16839	0.624543	0.95244	0.708316	1.323798	1.315643
Test 30	5.064591	3.096156	3.140794	2.235595	1.251722	1.077815	1.263328	1.588201

Table 4: RAE values of logistic regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.558983	0.748235	0.100892	0.130431	0.982956	0.946141	0.79988	1.233276
Test 3	10.019084	92.770195	3.992647	2.151188	1.450237	3.423077	2.1	0.607143
Train 7	0.761442	0.858548	0.188449	0.264792	1.010998	1.044404	0.76232	1.230431
Test 7	7.11657	19.883806	2.05869	2.056796	1.590134	2.022506	1.510684	0.964646
Train 14	0.939364	1.035772	0.275228	0.410008	1.071022	1.153797	0.760876	1.234803
Test 14	6.073407	7.771993	1.870907	1.948638	1.803453	1.303851	1.431579	0.887464
Train 21	0.98917	1.066849	0.295554	0.548156	1.145123	1.232673	0.802145	1.209536
Test 21	5.33746	5.418942	1.735014	1.822661	1.563513	1.045426	1.450346	0.879166
Train 30	1.062807	1.154789	0.295746	0.677223	1.207509	1.219678	0.84517	1.22699
Test 30	5.473092	4.012622	1.674963	1.726235	1.50333	1.227554	1.405158	0.911704

consistent and comprehensive dataset, facilitating more accurate forecasting.

Conversely, Japan, while having a sophisticated healthcare system, faced challenges in its initial response to the pandemic. The country experienced periods of underreporting, potentially due to limited testing capacities in the early stages and specific administrative bottlenecks. Such inconsistencies in data collection can introduce noise into the dataset, making it more challenging for machine learning models to discern underlying patterns and make accurate predictions.

Furthermore, sociocultural factors played a role. Japan's unique societal norms, including its densely populated urban centers and specific public health communication strategies, influenced the dynamics of the disease's spread

Table 5: MAPE values of logistic regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.085625	0.082203	0.032775	0.022732	1.094825	2.063729	4.719072	3.102245
Test 3	0.0249	0.12132	0.007956	0.008657	0.514282	0.175681	0.157631	0.761792
Train 7	0.109693	0.118866	0.097656	0.055901	0.975737	1.726743	3.959241	3.220471
Test 7	0.031743	0.211901	0.018863	0.023588	0.51773	0.114646	0.20854	0.612907
Train 14	0.135223	0.140357	0.186165	0.095634	0.8972	1.279357	3.639863	3.812503
Test 14	0.037776	0.230585	0.076686	0.067183	0.474387	0.13922	0.279167	0.542086
Train 21	0.170498	0.158015	0.233285	0.113069	1.149452	1.549936	3.975037	4.335126
Test 21	0.042581	0.251713	0.10385	0.098062	0.479235	0.155378	0.312001	0.483003
Train 30	0.166786	0.190446	0.283616	0.133828	1.136176	1.408250	4.407955	4.867122
Test 30	0.048385	0.2997	0.132413	0.126256	0.587006	3.224258	0.276267	0.383016

Table 6: MAPE values of logistic regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.077908	0.216183	0.28478	0.031407	4.302802	3.566545	4.780891	3.266033
Test 3	0.005978	0.226541	0.002404	0.002224	0.682114	0.650904	0.325321	0.156613
Train 7	0.107693	0.265915	0.397845	0.070797	8.605604	4.503600	9.561783	3.305931
Test 7	0.007141	0.168689	0.005143	0.00512	0.930147	0.528263	0.446618	0.188684
Train 14	0.135295	0.336248	0.479926	0.108925	8.605604	5.239857	4.780891	3.621599
Test 14	0.008536	0.157663	0.010831	0.0103	1.745331	0.437303	0.482487	0.210358
Train 21	0.143851	0.346251	0.481132	0.142261	7.649426	6.760180	4.780891	3.279957
Test 21	0.00962	0.177791	0.017679	0.016401	1.81934	2.702160	0.487855	0.246182
Train 30	0.153738	0.358279	0.460946	0.176627	1.147414	5.603205	4.780891	2.916246
Test 30	0.010656	0.207111	0.029169	0.026349	1.888606	1.830495	0.456813	0.291251

Table 7: MAE values of decision tree models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	118326.82	50614.35	83292.72	14470.19	13966.14	6129.43	12921.25	1787.4
Test 3	92206.67	373864.33	159047.67	33767.33	15383.67	25373	32482	5467.67
Train 7	210181.83	98095.02	159424.29	31352.4	14121.04	8610.34	14078.32	1799.26
Test 7	219605	819688.29	382687.57	94103.43	17714.57	19880.29	25214.43	8544.14
Train 14	411432.76	176243.54	296562.74	56438.12	18259.02	11183.94	24367.5	2139.05
Test 14	380742.29	1557960.64	787013.79	205029.5	16782.86	24695.79	36543.5	8752.71
Train 21	609247.27	245358.77	431184.05	80469.2	22295.56	12715.18	34615.07	2389.48
Test 21	527135.24	2260445.24	1210456.86	324242.62	18504.93	29095.67	41458	9037.81
Train 30	914462.41	345052.67	629787.46	113795.01	22851.02	14905.65	44442.16	3629.26
Test 30	712922.32	3288235.39	1799969.39	482712.19	20317.9	39667.35	41360.94	8159.19

in ways that diverged from patterns observed in Germany and South Korea.

It is essential to consider the potential influence of viral strains. Different regions might have been affected by varying strains of the virus at other times, each with its transmission dynamics. This could have introduced additional variability into the predictions if Japan was predominantly affected by a strain with varying transmission characteristics during the forecast period.

Table 19 shows accuracy of all models of COVID-19 in selected territories for 30 days regarding the character of the input data.

The support vector regression model shows the highest accuracy for all datasets. At the same time, it can be noted

Table 8: MAE values of decision tree models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	254.52	98.9	98.22	421.51	34.8	21.26	18.22	52.5
Test 3	146	192.33	40.33	221.33	16.67	18.67	9.67	15
Train 7	466.39	186.43	193.21	899.28	36.83	19.97	17.17	55.82
Test 7	317.57	505.86	109.57	511.43	15.14	19.43	11.71	20
Train 14	909.3	354.55	366.16	1640.46	53.07	24.1	23.57	69.55
Test 14	523.86	1183.71	255.21	1036.36	12.79	35.43	17.21	20.86
Train 21	1348.09	519.24	538.23	2351.47	56.39	27.8	32.1	85.5
Test 21	684.81	1937.67	433.9	1665.52	11.38	42.1	16.38	32.95
Train 30	2024.18	763.51	779.33	3369.01	90.04	34.75	40.28	104.06
Test 30	838.74	3183.81	742.58	2716.65	10.32	46.52	15.65	50

Table 9: RAE values of decision tree models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.013721	0.017121	0.012359	0.025513	0.249812	0.237307	0.22947	0.337549
Test 3	1.64074	3.008135	2.036247	2.242518	0.453807	1.99414	1.026396	1.397268
Train 7	0.024372	0.033182	0.023655	0.055279	0.252583	0.333357	0.250018	0.339791
Test 7	2.011061	2.101346	1.996374	1.767298	0.722817	0.856205	1.541001	1.202903
Train 14	0.047708	0.059617	0.044003	0.09951	0.326598	0.432997	0.432744	0.403959
Test 14	2.342152	2.110279	1.933691	1.848344	0.76622	1.046646	1.806664	1.176185
Train 21	0.070645	0.082996	0.063977	0.14188	0.3988	0.49228	0.614732	0.451253
Test 21	2.285203	2.112131	1.904707	1.813019	0.88611	1.043608	1.623965	1.232625
Train 30	0.106036	0.116719	0.093445	0.200639	0.408735	0.577086	0.789252	0.685385
Test 30	2.353681	2.07895	1.93567	1.908543	1.025704	1.073062	1.60721	1.124144

Table 10: RAE values of decision tree models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.013429	0.016858	0.012715	0.025721	0.365092	0.533253	0.33507	0.346589
Test 3	1.671756	2.410864	2.669118	2.151188	0.35545	1.076923	2.9	0.401786
Train 7	0.024607	0.031779	0.025012	0.054875	0.386407	0.500952	0.315743	0.36852
Test 7	2.15945	1.811386	1.740921	2.056796	0.451888	0.579075	1.226496	0.707071
Train 14	0.047976	0.060438	0.047401	0.100102	0.556737	0.604513	0.433459	0.459151
Test 14	2.539474	1.736119	1.734947	1.948638	0.416003	0.591684	1.268421	0.582336
Train 21	0.071128	0.08851	0.069676	0.143488	0.591582	0.697237	0.590335	0.564496
Test 21	2.585315	1.707506	1.658565	1.822661	0.39501	0.577598	1.08697	0.652772
Train 30	0.1068	0.13015	0.100888	0.20558	0.944503	0.871579	0.740768	0.687041
Test 30	2.928104	1.708068	1.631034	1.726235	0.34411	0.681869	0.912318	0.795688

that all models for data from Germany and South Korea show the highest accuracy. This indicates a more complete testing and registration of a higher percentage of actual incidence than in Japan and Ukraine.

The model analysis results showed that the use of cumulative case and death data as input increased the accuracy of the models, which at first glance is attractive and may lead to the misconception of not using data on daily new cases and deaths. However, the evaluation of models using MAE shows a much smaller absolute error. Based on the data obtained, it should be concluded that to build models. It is necessary to use the entire set of available data, both daily and cumulative, giving preference to cumulative data during periods full of holidays and weekends and daily data in other

Table 11: MAPE values of decision tree models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.00967	0.013254	0.021191	0.006507	0.356152	7.767514	0.28222	0.332541
Test 3	0.002969	0.028793	0.007956	0.008657	0.261771	0.140789	0.380797	0.265373
Train 7	0.017271	0.025857	0.039773	0.01404	0.380668	7.023082	0.318073	0.337635
Test 7	0.007039	0.060258	0.018863	0.023588	0.5067	0.097694	0.266316	0.2949
Train 14	0.033259	0.046418	0.070861	0.025124	0.550332	1.016977	0.477096	0.402219
Test 14	0.012129	0.10712	0.037762	0.049494	0.786634	0.128593	0.317757	0.261905
Train 21	0.048688	0.065957	0.098308	0.035699	0.753696	8.665087	0.740159	0.586648
Test 21	0.016697	0.146555	0.056531	0.075343	0.911067	0.151105	0.328267	0.258661
Train 30	0.071595	0.092106	0.13495	0.050189	1.020199	7.372852	1.175772	0.658837
Test 30	0.022424	0.196695	0.081206	0.107261	1.142982	3.224258	0.364767	0.263778

Table 12: MAPE values of decision tree models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.002234	0.005224	0.011059	0.008466	1.051796	5.287666	1.912357	0.398852
Test 3	0.000997	0.005873	0.001607	0.002224	0.354808	0.1715	0.478632	0.106842
Train 7	0.004101	0.009886	0.021461	0.01804	1.051796	5.526710	5.737070	0.419804
Test 7	0.002166	0.015232	0.004348	0.00512	0.259721	0.138082	0.418675	0.150069
Train 14	0.007975	0.018619	0.039748	0.032437	1.051796	6.683686	2.868535	0.582338
Test 14	0.003567	0.034589	0.01004	0.0103	0.236864	0.158846	0.456952	0.145728
Train 21	0.011804	0.027272	0.056557	0.046043	1.338650	6.989663	9.561783	0.783502
Test 21	0.004658	0.054933	0.016894	0.016401	0.293422	3.838783	0.383957	0.173759
Train 30	0.01766	0.039383	0.079718	0.065041	1.061358	7.850224	3.824713	0.698311
Test 30	0.005699	0.085935	0.028394	0.026349	0.581109	2.600466	0.321396	0.221548

Table 13: MAE values of support vector regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	177461.93	42413.24	345269.25	6168.6	21543.02	6972.67	12801.19	1318.78
Test 3	29971.42	72498.7	31604.26	14735.92	9307.22	20931.64	11677.88	2895.02
Train 7	265493.7	77343.36	611024.82	11372.04	23277.99	9217.39	15721.31	1711.46
Test 7	18319.71	67374.52	17284.68	40875.99	6040.81	18690.31	17388.82	4444.81
Train 14	445035.41	146217.19	873006.51	27015.18	29017.56	12723.74	19964.82	2773.76
Test 14	28997.77	377267.67	71368.74	107157.97	5043.77	22257.87	25912.74	6375.72
Train 21	667271.19	230578.59	1075023.62	43444.73	31731.82	10768.37	26608.89	2806.22
Test 21	28105.31	1147611.83	246277.95	207970.03	4936.3	22706.61	30160.83	7056.76
Train 30	941923.57	297501.25	1375049.59	86625.4	36162.97	12782.22	34437.58	3063.81
Test 30	39144.99	2740984.16	782585.62	395002.04	6565.9	33405.66	27506.52	7493.87

periods. To reduce the absolute error, it is necessary to form databases based on daily morbidity and mortality.

The intricate relationship between machine learning and the available data forms the bedrock of our research endeavors. At its core, machine learning thrives on data; the quality, granularity, and comprehensiveness of this data directly influence the efficacy of the predictive models [40]. In the context of our study, the data sourced from the World Health Organization Coronavirus Dashboard served as the empirical foundation upon which our models were trained, validated, and tested.

They are implementing a forecasting system for a phenomenon as dynamic and multifaceted as the COVID-19 pandemic presents a unique set of challenges distinct from

Table 14: MAE values of support vector regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	399.16	90.81	72.14	617.96	50.06	14.44	28.87	40.12
Test 3	102.69	180.45	20.51	64.21	41.55	21.76	5.51	21.52
Train 7	681.17	171.72	214.52	1067.67	50.84	15.75	30.78	44.9
Test 7	225.59	461.32	21.5	38.32	27.67	14.07	3.79	15.48
Train 14	1202.87	321.98	641.75	1778.12	60.34	19.37	33.74	58.43
Test 14	352.16	1085.13	22.18	136.92	24.66	36.1	9.06	24.67
Train 21	1723.38	476.29	1037.56	2673.69	68.13	24.05	39.41	67.3
Test 21	439.12	1787.47	20.5	394.72	22.38	39.28	11.32	42.76
Train 30	2471.49	743.6	1612.14	3814.64	77.89	27.76	41.99	89.43
Test 30	496.3	2963.69	19.63	985.68	22.33	37.67	13.84	57.79

Table 15: RAE values of support vector regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.020578	0.014347	0.051229	0.010876	0.385339	0.269954	0.227337	0.249052
Test 3	0.533316	0.583329	0.404621	0.978625	0.274556	1.645081	0.369008	0.739826
Train 7	0.030785	0.026163	0.090661	0.020051	0.416373	0.35686	0.279196	0.32321
Test 7	0.167765	0.172721	0.090169	0.767667	0.246486	0.804955	1.062733	0.62577
Train 14	0.051604	0.04946	0.129533	0.047632	0.519036	0.492612	0.354557	0.523824
Test 14	0.178381	0.511014	0.175353	0.96603	0.230273	0.943323	1.281093	0.856766
Train 21	0.077373	0.077997	0.159507	0.0766	0.567586	0.416908	0.472549	0.529954
Test 21	0.12184	1.072314	0.387529	1.162875	0.236375	0.814444	1.18144	0.962439
Train 30	0.10922	0.100634	0.204023	0.152735	0.646846	0.494876	0.61158	0.578599
Test 30	0.129235	1.732956	0.841585	1.561756	0.331465	0.903673	1.068853	1.032478

Table 16: RAE values of support vector regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.02106	0.01548	0.009339	0.037708	0.525147	0.362153	0.530935	0.264905
Test 3	1.175869	2.261846	1.357158	0.624051	0.88613	1.255252	1.652311	0.5763
Train 7	0.03594	0.029272	0.02777	0.06515	0.533342	0.395006	0.566022	0.29647
Test 7	1.534016	1.651911	0.341557	0.15409	0.825802	0.419491	0.396898	0.547401
Train 14	0.063465	0.054886	0.083077	0.108502	0.632977	0.486001	0.620467	0.385795
Test 14	1.707127	1.591522	0.150748	0.257448	0.802261	0.602924	0.667564	0.688777
Train 21	0.090928	0.081189	0.134317	0.16315	0.714705	0.603202	0.724743	0.444301
Test 21	1.65779	1.575147	0.078365	0.431956	0.776902	0.538983	0.751257	0.847105
Train 30	0.1304	0.126756	0.208699	0.232772	0.817074	0.696338	0.772135	0.590408
Test 30	1.732633	1.589979	0.043109	0.626331	0.744521	0.552149	0.806762	0.919628

conventional forecasting endeavors. Traditional forecasting models often rely on stable, predictable patterns. In contrast, the COVID-19 pandemic, influenced by many sociopolitical, environmental, and biological factors, exhibits a level of volatility that demands a more adaptive and nuanced modeling approach [41]. Our machine learning models, particularly the support vector regression, were designed to navigate this volatility, learning from the intricacies of the data to make robust predictions.

Looking ahead, the field of COVID-19 forecasting is poised to encounter several challenges. The emergence of new viral strains, changing vaccination rates, and evolving public health measures can introduce unforeseen complexities into the data. By emphasizing the importance of daily

Table 17: MAPE values of support vector regression models for confirmed cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.01081	0.00761	0.037698	0.002576	0.428747	0.834929	0.26526	0.207072
Test 3	0.000968	0.005581	0.001588	0.00378	0.266631	0.110273	0.136966	0.156381
Train 7	0.016709	0.014585	0.06877	0.004817	0.475477	0.706441	0.422481	0.271758
Test 7	0.00059	0.005033	0.000865	0.010244	0.179739	0.087962	0.172116	0.168697
Train 14	0.029249	0.029065	0.103574	0.011468	0.655775	0.829557	0.594183	0.414621
Test 14	0.000927	0.025298	0.003399	0.025782	0.239441	0.109741	0.220938	0.216823
Train 21	0.04544	0.052855	0.13259	0.018537	0.803028	0.870909	0.839054	0.526918
Test 21	0.000895	0.07172	0.011317	0.047973	0.260429	0.107378	0.242646	0.23299
Train 30	0.063497	0.067262	0.176482	0.036429	0.946648	0.784371	0.993964	0.631374
Test 30	0.001233	0.157201	0.034564	0.086736	0.342491	0.288011	0.261809	0.268311

Table 18: MAPE values of support vector regression models for fatal cases.

Forecast period (days)	Cumulative cases				Daily cases			
	Germany	Japan	South Korea	Ukraine	Germany	Japan	South Korea	Ukraine
Train 3	0.003765	0.004802	0.006072	0.010939	4.416103	4.064632	0.857073	0.30425
Test 3	0.000701	0.005511	0.000817	0.000646	1.948793	0.222187	0.276933	0.2145
Train 7	0.006419	0.009178	0.015502	0.019093	4.343917	4.159286	0.956759	0.317696
Test 7	0.001539	0.013893	0.000854	0.000385	1.320155	0.123105	0.15837	0.142423
Train 14	0.011256	0.016954	0.043637	0.031653	4.861689	3.829763	0.911865	0.401937
Test 14	0.002398	0.031704	0.000878	0.001358	1.392047	0.15902	0.21682	0.173629
Train 21	0.016029	0.024938	0.070958	0.046336	4.903038	4.657209	0.497613	0.517992
Test 21	0.002987	0.050664	0.000807	0.003872	1.797445	0.419907	0.227797	0.229247
Train 30	0.02277	0.038143	0.10976	0.063765	5.883873	3.226588	0.881677	0.654389
Test 30	0.003373	0.079958	0.000763	0.00951	4.348637	0.284453	0.239565	0.267526

Table 19: Accuracy of models (%).

Data	Germany	Japan	South Korea	Ukraine
Logistic regression				
Cumulative new cases	99.95162	99.70030	99.86759	99.87374
Daily new cases	99.41299	96.77574	99.72373	99.61698
Cumulative fatal cases	99.98934	99.79289	99.97083	99.97365
Daily fatal cases	98.11139	98.16951	99.54319	99.70875
Decision tree				
Cumulative new cases	99.97758	99.80331	99.91879	99.89274
Daily new cases	98.85702	96.77574	99.63523	99.73622
Cumulative fatal cases	99.99430	99.91407	99.97161	99.97365
Daily fatal cases	99.41889	97.39953	99.67860	99.77845
Support vector regression				
Cumulative new cases	99.99877	99.84280	99.96544	99.91326
Daily new cases	99.65751	99.71199	99.73819	99.73169
Cumulative fatal cases	99.99663	99.92004	99.99924	99.99049
Daily fatal cases	95.65136	99.71555	99.76044	99.73247

and cumulative data, our research offers a blueprint for addressing some of these challenges. By ensuring that our models are trained on comprehensive datasets that capture the full spectrum of the pandemic's dynamics, we enhance their adaptability and resilience against future uncertainties [42].

Our study contributes significantly to the broader discourse on the role of machine learning in healthcare systems. By demonstrating the potential of machine learning models to make accurate short-term forecasts in the context of a global pandemic, we underscore the transformative potential of these technologies in public health decision-making. As healthcare systems worldwide grapple with the challenges of the 21st century, from pandemics to chronic diseases, the integration of machine learning tools, as evidenced by our research, will be pivotal in driving innovation, efficiency, and improved patient outcomes [43].

The salient observation from our research underscores the differential impact of accumulated holiday data versus daily data during weekdays on the predictive accuracy of our models. While evident in our results, this distinction warrants a more in-depth exploration to elucidate the underlying mechanisms that contribute to this phenomenon.

One plausible hypothesis is that during holidays and weekends, there is a marked reduction in the number of individuals seeking medical attention, leading to the potential underreporting of cases. This underreporting can introduce noise into the data, making daily figures during these periods less reliable. By accumulating data over such periods, we might mitigate this noise's effects, thereby enhancing the model's predictive capabilities. Conversely, on regular weekdays, when medical facilities operate at their usual capacity and individuals are more likely to seek medical care, daily data provides a more granular and accurate representation of the disease's spread.

Furthermore, the implications of this observation extend beyond the realm of academic interest. In practical terms, understanding the nuances of data collection and its impact on model accuracy can significantly influence how healthcare systems approach data-driven decision-making. For instance, policymakers and healthcare administrators could prioritize the collection of cumulative data during holiday-rich periods and place greater emphasis on daily data during regular operational days. This tailored approach to data collection, driven by our findings, could potentially enhance the accuracy of future predictive models, leading to more informed and effective epidemic control measures.

Moreover, while our study has shed light on this particular aspect of data utilization, it also underscores the broader need for a holistic approach to model development in healthcare. It is not merely about selecting the right algorithm or having vast amounts of data; it is about understanding the intricacies of the data, the context in which it is collected, and the myriad factors that can influence its quality and reliability. Only by addressing these nuances can we truly harness the power of machine learning in the service of public health.

6. Conclusions

The article describes the results of an experimental study of three models for predicting the dynamics of COVID-19 based on statistical machine learning methods: logistic regression, decision tree, and support vector regression. For the experiments, data on the incidence and mortality of COVID-19 in Germany, Japan, South Korea, and Ukraine, provided by the World Health Organization COVID-19 Dashboard, were used.

All developed models have shown sufficient accuracy for use in healthcare practice for the development and implementation of control measures to curb the spread of infectious diseases.

The prediction accuracy of the logistic regression model ranged from 96.78% to 99.95% for morbidity and from 98.11% to 99.99% for fatal cases. The accuracy of the decision tree model ranged from 96.78% to 99.98% for morbidity and from 97.39% to 99.99% for lethal cases. The accuracy of the support vector regression model ranged from 99.65% to 99.99% for morbidity and from 95.65% to 99.99% for lethal cases.

At the same time, the analysis of model indicators for all data showed that the most accurate model is a model based on the support vector regression method. The results of the model analysis showed that the use of cumulative case and death data as input increased the accuracy of the models, which at first glance is attractive and may lead to the misconception of not using data on daily new cases and deaths. However, the evaluation of models using MAE shows a much smaller absolute error.

It should be concluded that to build models, it is necessary to use the entire set of available data, both daily and cumulative, giving preference to cumulative data during periods full of holidays and weekends and daily data in other periods. To reduce the absolute error, it is necessary to form databases based on daily morbidity and mortality.

The scientific novelty of the research lies in the development of COVID-19 predictive models based on statistical machine learning methods. Unlike other studies, the article analyzes the performance of the model depending on different forecasting periods. Unlike other studies, the article analyzes the use of various input data (cumulative and daily) for modeling. The results of the analysis will increase the effectiveness of the use of machine learning models of infectious diseases in healthcare systems.

Data Availability

The initial data used in this research is publicly available in World Health Organization (WHO) COVID-19 Dashboard (https://covid19.who.int/) (accessed on 25 May 2023).

Conflicts of Interest

The authors declare that they have no financial and nonfinancial competing interests.

Authors' Contributions

D.C. was responsible for the conceptualization. D.C. was responsible for the methodology. D.C. and T.D. were responsible for the software. D.C., T.D., S.Y., and T.C. were responsible for the validation. D.C. and T.C. were responsible for the formal analysis. D.C. and S.Y. were responsible for the investigation. D.C. was responsible for the resources. D.C. and T.D. were responsible for the data curation. D.C.,

T.D., and T.C. were responsible for the writing—original draft preparation. S.Y. was responsible for the writing—review and editing. D.C. and T.D. were responsible for the visualization. D.C. was responsible for the supervision. S.Y. was responsible for the project administration. D.C. was responsible for the funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We would like to thank the Armed Forces of Ukraine for providing security to perform this work. This work has become possible only because of the resilience and courage of the Ukrainian Army and people. The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic "Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management."

References

[1] Worldometer, "COVID-19 coronavirus pandemic," November 2022, https://www.worldometers.info/coronavirus/#countries.

[2] J. Amankwah-Amoah, Z. Khan, G. Wood, and G. Knight, "COVID-19 and digitalization: the great acceleration," *Journal of Business Research*, vol. 136, pp. 602–611, 2021.

[3] A. Sheremet, Y. Kondratenko, I. Sidenko, and G. Kondratenko, "Diagnosis of Lung Disease Based on Medical Images Using Artificial Neural Networks," in *2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 561–565, Lviv, Ukraine, 2021.

[4] K. Bazilevych, M. Butkevych, and H. Padalko, "Classification of Cardiovascular Disease Using AdaBoost Method," in *Smart Technologies in Urban Engineering. STUE 2022*, O. Arsenyeva, T. Romanova, M. Sukhonos, and Y. Tsegelnyk, Eds., vol. 536 of Lecture Notes in Networks and Systems, Springer, Cham, 2023.

[5] N. Davidich, I. Chumachenko, Y. Davidich, H. Taisiia, N. Artsybasheva, and M. Tatiana, "Advanced Traveller Information Systems to Optimizing Freight Driver Route Selection," in *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 111–115, Liverpool, United Kingdom, 2020.

[6] R. Radutniy, A. Nechyporenko, V. Alekseeva, G. Titova, D. Bibik, and V. V. Gargin, "Automated Measurement of Bone Thickness on SCT Sections and Other Images," in *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pp. 222–226, Lviv, Ukraine, 2020.

[7] I. Izonin, R. Tkachenko, I. Dronyuk, P. Tkachenko, M. Gregus, and M. Rashkevych, "Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method," *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2599–2613, 2021.

[8] Y. Xiang, Y. Jia, L. Chen, L. Guo, B. Shu, and E. Long, "COVID-19 epidemic prediction and the impact of public health interventions: a review of COVID-19 epidemic models," *Infectious Disease Modelling*, vol. 6, pp. 324–342, 2021.

[9] K. Bazilevych, D. Chumachenko, L. Hulianytskyi, I. Meniailov, and S. Yakovlev, "Intelligent decision-support system for epidemiological diagnostics. I. A concept of architecture design," *Cybernetics and System Analysis*, vol. 58, no. 3, pp. 343–353, 2022.

[10] D. Chumachenko, I. Meniailov, K. Bazilevych, and O. Chub, "On COVID-19 epidemic process simulation: three regression approaches investigations," *Radioelectronic and Computer Systems*, vol. 2022, no. 1, pp. 6–22, 2022.

[11] D. Chumachenko, I. Meniailov, K. Bazilevych, T. Chumachenko, and S. Yakovlev, "Investigation of statistical machine learning models for COVID-19 epidemic process simulation: random forest, k-nearest neighbors, gradient boosting," *Computation*, vol. 10, no. 6, p. 86, 2022.

[12] I. Cooper, A. Mondal, and C. G. Antonopoulos, "A SIR model assumption for the spread of COVID-19 in different communities," *Chaos, Solitons & Fractals*, vol. 139, article 110057, no. 139, 2020.

[13] S. A. Alanazi, M. M. Kamruzzaman, M. Alruwaili, N. Alshammari, S. A. Alqahtani, and A. Karime, "Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care," *Journal of Healthcare Engineering*, vol. 2020, Article ID 8857346, 12 pages, 2020.

[14] L.-P. Chen, Q. Zhang, G. Y. Yi, and W. He, "Model-based forecasting for Canadian COVID-19 data," *PLoS One*, vol. 16, no. 1, article e0244536, 2021.

[15] R. u. Din and E. A. Algehyne, "Mathematical analysis of COVID-19 by using SIR model with convex incidence rate," *Results in Physics*, vol. 23, article 103970, 2021.

[16] A. Ajbar, R. T. Alqahtani, and M. Boumaza, "Dynamics of an SIR-based COVID-19 model with linear incidence rate, nonlinear removal rate, and public awareness," *Frontiers in Physics*, vol. 9, 2021.

[17] J. Wang, Y. Liu, X. Liu, and K. Shen, "A modified SIR model for the COVID-19 epidemic in China," *Journal of Physics: Conference Series*, vol. 2148, no. 1, article 012002, 2022.

[18] R. A. Singh, R. Lal, and R. R. Kotti, "Time-discrete SIR model for COVID-19 in Fiji," *Epidemiology and Infection*, vol. 150, pp. 1–17, 2022.

[19] H. AlQadi and M. Bani-Yaghoub, "Incorporating global dynamics to improve the accuracy of disease models: example of a COVID-19 SIR model," *PLoS One*, vol. 17, no. 4, article e0265815, 2022.

[20] D. Uçar and E. Çelik, "Analysis of Covid 19 disease with SIR model and Taylor matrix method," *AIMS Mathematics*, vol. 7, no. 6, pp. 11188–11200, 2022.

[21] T. T. Marinov and R. S. Marinova, "Adaptive SIR model with vaccination: simultaneous identification of rates and functions illustrated with COVID-19," *Scientific Reports*, vol. 12, no. 1, article 15688, 2022.

[22] M. Kamrujjaman, P. Saha, M. S. Islam, and U. Ghosh, "Dynamics of SEIR model: a case study of COVID-19 in Italy," *Results in Control and Optimization*, vol. 7, article 100119, 2022.

[23] Z.-Y. Zhao, Y.-Z. Zhu, J.-W. Xu et al., "A five-compartment model of age-specific transmissibility of SARS-CoV-2," *Infectious Diseases of Poverty*, vol. 9, no. 1, p. 117, 2020.

[24] M. Fošnarič, T. Kamenšek, J. Žganec Gros, and J. Žibert, "Extended compartmental model for modeling COVID-19 epidemic in Slovenia," *Scientific Reports*, vol. 12, no. 1, article 16916, 2022.

[25] A. Leontitsis, A. Senok, A. Alsheikh-Ali, Y. Al Nasser, T. Loney, and A. Alshamsi, "SEAHIR: a specialized compartmental model for COVID-19," *International Journal of Environmental Research and Public Health*, vol. 18, no. 5, p. 2667, 2021.

[26] E. Antonelli, E. L. Piccolomini, and F. Zama, "Switched forced SEIRDV compartmental models to monitor COVID-19 spread and immunization in Italy," *Infectious Disease Modelling*, vol. 7, no. 1, pp. 1–15, 2022.

[27] S. Dash, S. Chakravarty, S. N. Mohanty, C. R. Pattanaik, and S. Jain, "A deep learning method to forecast COVID-19 outbreak," *New Generation Computing*, vol. 39, no. 3-4, pp. 515–539, 2021.

[28] R. Chandra, A. Jain, and D. Singh Chauhan, "Deep learning via LSTM models for COVID-19 infection forecasting in India," *PLoS One*, vol. 17, no. 1, article e0262708, 2022.

[29] M. O. Alassafi, M. Jarrah, and R. Alotaibi, "Time series predicting of COVID-19 based on deep learning," *Neurocomputing*, vol. 468, pp. 335–344, 2022.

[30] Y. Alali, F. Harrou, and Y. Sun, "A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models," *Scientific Reports*, vol. 12, no. 1, p. 2467, 2022.

[31] L. Xu, R. Magar, and A. Barati Farimani, "Forecasting COVID-19 new cases using deep learning methods," *Computers in Biology and Medicine*, vol. 144, article 105342, 2022.

[32] A. Almalki, B. Gokaraju, Y. Acquaah, and A. Turlapaty, "Regression analysis for COVID-19 infections and deaths based on food access and health issues," *Healthcare*, vol. 10, no. 2, p. 324, 2022.

[33] A. M. Almeshal, A. I. Almazrouee, M. R. Alenizi, and S. N. Alhajeri, "Forecasting the spread of COVID-19 in kuwait using compartmental and logistic regression models," *Applied Sciences*, vol. 10, no. 10, p. 3402, 2020.

[34] M. Giotta, P. Trerotoli, V. O. Palmieri et al., "Application of a decision tree model to predict the outcome of non-intensive inpatients hospitalized for COVID-19," *International Journal of Environmental Research and Public Health*, vol. 19, no. 20, article 13016, 2022.

[35] J. Zhang, X. Qiu, X. Li, Z. Huang, M. Wu, and Y. Dong, "Support vector machine weather prediction technology based on the improved quantum optimization algorithm," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6653659, 13 pages, 2021.

[36] Y.-T. Chen, "Effect of vaccination patterns and vaccination rates on the spread and mortality of the COVID-19 pandemic," *Health Policy and Technology*, vol. 12, no. 1, article 100699, 2023.

[37] M. Tanhaeean, N. Nazari, S. H. Iranmanesh, and M. Abdollahzade, "Analyzing factors contributing to COVID-19 mortality in the United States using artificial intelligence techniques," *Risk Analysis*, vol. 43, no. 1, pp. 19–43, 2023.

[38] A. V. R. Amaral, J. A. González, and P. Moraga, "Spatio-temporal modeling of infectious diseases by integrating compartment and point process models," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 4, pp. 1519–1533, 2023.

[39] B. Naffeti, S. Bourdin, W. Ben Aribi, A. Kebir, and S. Ben Miled, "Spatio-temporal evolution of the COVID-19 across African countries," *Frontiers in Public Health*, vol. 10, 2022.

[40] S. Uddin, S. Ong, and H. Lu, "Machine learning in project analytics: a data-driven framework and case study," *Scientific Reports*, vol. 12, no. 1, article 15252, 2022.

[41] J. Luo, "Forecasting COVID-19 pandemic: unknown unknowns and predictive monitoring," *Technological Forecasting and Social Change*, vol. 166, article 120602, 2021.

[42] F. Piccialli, V. S. di Cola, F. Giampaolo, and S. Cuomo, "The role of artificial intelligence in fighting the COVID-19 pandemic," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1467–1497, 2021.

[43] U. Kose, O. Deperlioglu, J. Alzubi et al., "Future of medical decision support systems," *Deep Learning for Medical Decision Support Systems*, vol. 909, pp. 157–171, 2021.