Nonlinear Dynamics and Systems Theory, **21** (5) (2021) 494–509

COVID-19 Outbreak Prediction in Indonesia Based on Machine Learning and SIRD-Based Hybrid Methods

E. R. M. Putri [1], M. Iqbal [1,*], M. L. Shahab [1], H. N. Fadhilah [1,2], I. Mukhlash [1], D. K. Arif [1], E. Apriliani [1] and H. Susanto [3,4]

[1] *Department of Mathematics, Sepuluh Nopember of Institute Technology Kampus ITS Sukolilo-Surabaya 60111, Indonesia.*
[2] *Department of Data Science, Telkom Institute of Technology Surabaya, Jl. Ketintang No.156, Ketintang, Kec. Gayungan, Kota SBY, Jawa Timur 60231.*
[3] *Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, United Kingdom, CO4 3SQ.*
[4] *Department of Mathematics, Khalifa University, Abu Dhabi Campus - PO Box 127788, Abu Dhabi, United Arab Emirates.*

Received: July 7, 2021; Revised: September 29, 2021

Abstract: This paper aims to forecast and analyze the spread of COVID-19 outbreak in Indonesia by applying machine learning and hybrid approaches. We show the performance of each method, an ensemble-support vector regression (ensemble-SVR), a genetic algorithm and an SIRD model (GA-SIRD) and an extended Kalman filter, a genetic algorithm and an extended Kalman filter (EKF-GA-SIRD), in obtaining the prediction of the outbreak. The GA-SIRD model is built based on the data availability and is enhanced by employing an extended Kalman filter to better predict the spread of the outbreak. Without considering the epidemic model, the ensemble SVR can provide a higher accuracy compare to the two hybrid approaches in the case of short-term forecasting. Furthermore, the EKF-GA-SIRD can better adapt to the extreme change and shows a better performance than the GA-SIRD.

Keywords: *pandemic; SIRD model; Kalman filter; machine learning.*

Mathematics Subject Classification (2010): 70-08, 93A30.

* Corresponding author: mailto:iqbal@matematika.its.ac.id

1 Introduction

The pandemic of the novel coronavirus or COVID-19 started in Wuhan, China, where the first case was reported on January 22nd, 2020, and the pandemic has spread worldwide in more than 200 countries. As China has passed through its first pandemic peak, some other countries such as the US, India, and Indonesia are still struggling to control the spread of the virus. The spread of COVID-19 in Indonesia was reported for the first time on March 2nd, 2020, in Jakarta, and currently, it reaches almost all provinces of Indonesia in less than two months. Statistical data of the outbreak in Indonesia is officially collected from `https://covid19.go.id/`.

The spread of the pandemic of novel coronavirus COVID-19 can be described mathematically in the so-called mathematics of epidemiology. There are some common models, a SIR model (Susceptible, Infected, Recovered) or a SIRD (Susceptible, Infected, Recovered, Dead), which are used by some researchers to describe the spread pattern of COVID-19. A modification of the SIR model, which includes the death variable of observation, is called a SIRD model. Some researchers have done studies about the spread of the pandemic of COVID-19 based on the SIRD model. Fanelli *et al.* [1] used the SIRD model to predict the spread of COVID-19 in China, Italy, and Iran. Parameters are estimated using stochastic differential evolution. However, inadequate accuracy is shown in the study [1] when the peak prediction is compared to the latest data. As an improvement, Susanto [2] suggested that a careful fitting of reported data to the SIR model should be done due to its sensitivity to the time-series information. Salgotra *et al.* [3] used a genetic-based algorithm for estimating the parameters. They have shown that the algorithm is highly reliable for predicting COVID-19 cases. Abdul Rahman [4] discussed machine learning to simulate the spread of COVID-19 based on the SIRD model. An error analysis and detail flow charts of the process are presented in his paper. The previous studies pointed out that the epidemic model is improved by machine learning in the parameters estimation process. Shortly, we may call the combination of the SIRD model and machine learning, a hybrid epidemic model.

In machine learning, regression models can offer promising forecasting by learning the given data set. Parbat *et al.* [5] applied *a support vector regression* to predict current and future COVID-19 cases in India without comparison to other regression models. Still with the cases in India, Sujath *et al.* [6] investigated *a linear regression* (LR), *a vector autoregression* (VA) and *a multilayer perceptron* (MLP) prediction. As to the results, the MLP showed better ones than the VA and LR. In addition, Tuli *et al.* [7] developed a real-time framework for the COVID-19 infected number prediction over the world by integrating cloud computing and machine learning. They adjusted iterative weighting on the generalized inverse Weibull distribution to have higher accuracy in the data-driven environment responding to the epidemic actively. Both the hybrid and the machine learning methods exhibit a relatively satisfying performance in predicting the spread of the COVID-19 outbreak. The use of the GA in estimating parameters of mathematical models give better performance than conventional methods [8].

Therefore, we attempt to propose new hybrid methods by integrating a genetic algorithm and a SIRD model (GA-SIRD), an extended Kalman filter (GA-EKF-SIRD) which provide a one-step updating process for predicting the spread of the outbreak in Indonesia. Furthermore, we proposed an ensemble-SVR method to forecast the COVID-19 cases without considering the SIRD model. The ensemble-SVR is a method that combines two different models under the SVR approach to tackle a limitation data on the decreasing

number of infected cases for the first time. In this study, we combine different COVID-19 case models of two countries that share similar distributions. More specifically, we focus on the COVID-19 cases in Indonesia by combining them with similar COVID-19 cases from another country, which has been through the first wave. Lastly, a comparison of the three methods is presented in this paper.

This paper is organized as follows. Section 2 and Section 3 discuss the hybrid of a genetic algorithm and a SIRD model, and accordingly, a genetic algorithm is incorporated into a hybrid of the extended Kalman filter and the SIRD model. We propose the use of the ensemble-SVR model in Section 4, respectively. Simulation and discussion are presented in Section 5 and conclusion is given in Section 6.

2 A Modified Extended Kalman Filter-SIRD Model

A SIRD model describes the evolution of an individual into classes: susceptible, infected, recovered, and dead. It is assumed that individuals in the same class have the same characteristics and the movement of individuals in the same class can be described. The infected individuals can recover without the possibility of being reinfected. A referenced total population is assumed to be constant, which means that the population's birth rate and death rate are the same.

A differential equation of the SIRD model, which describes the movement of individuals from one class to another class, is written as

$$\begin{aligned} \dot{S}(t) &= -rS(t)I(t), \\ \dot{I}(t) &= rS(t)I(t) - (a+d)I(t), \\ \dot{R}(t) &= aI(t), \\ \dot{D}(t) &= dI(t), \end{aligned} \tag{1}$$

where $S(t)$ describes the individuals who are at a high risk of infection, $I(t)$ describes the number of infected individuals, $R(t)$ describes the recovered individuals after being infected, and $D(t)$ describes the number of dead individuals. Then r, a, d are the rate of infection, recovery, and death, respectively. The SIRD model is used in this study due to data availability in the resource website such as `https://www.worldometers.info/coronavirus/#countries`. The dynamic of the SIRD model is estimated using an extended Kalman filter method in a discrete scheme.

An extended Kalman filter (EKF) method is a method for estimating a weakly non-linear stochastic dynamic system [9] such as the epidemic SIRD model. The EKF method has three main stages: an initialization of system and measurement model including the initialization of state variables values, time updates (prediction), and measurement updates (correction). In the non-linear system, the updates equations are intractable, so that an approximation to time update and measurement update equations are required to provide a computationally viable algorithm to apply to the filter.

The EKF method provides a one-step prediction based on the SIRD model, so that it is necessary to modify the method to get a longer prediction range based on limited measurement data. The modification is to generate new measurement data, expand the limited data, and add noise to get a longer prediction. This modification can be considered as our contribution. Briefly, the basic algorithm of a modified EKF can be seen in Algorithm 2.1.

Algorithm 2.1 Modified Ensemble Kalman Filter

System model : $\dot{x} = f(x,u,t) + G(t)w$
Measurement model : $z_n = h\left[x(t_n), n\right] + v_n$, with $x(0) \sim (\bar{x}, P_0), w(t) \sim (0,Q), v_n \sim (0,R)$

Initialization : $x(0) = \bar{x}_0,\ \hat{P}(0) = P_0,\ z_n^- = z_k$

Time Update

Estimate : $\dot{x} = f(x,u,t)$
Error covariance : $\dot{P} = A(\hat{x},t)P + PA^T(\hat{x},t) + GQG^T$
Jacobian : $A(x,t) = \frac{\partial f(x,u,t)}{\partial x}$

Measurement Update

Kalman gain : $K_n = P^-(t_n)H^T(\hat{x}_n^-)\left[H(\hat{x}_n^-)P^-(t_n)H^T(\hat{x}_n^-) + R\right]^{-1}$
Error covariance : $P(t_n) = \left[I - K_nH(\hat{x}_n^-)\right]P^-(t_n)$
Generate measurement : $z_n = z_n^- + v_n$
Estimate : $\hat{x}_n = \hat{x}_n^- + K_n\left[z_n - h(\hat{x}_n^-, n)\right]$
Jacobian : $H(x) = \frac{\partial h(x,n)}{\partial x}$

The number of infected, recovered, and dead individuals are predicted using the modified EKF, which is applied to the SIRD model. The Jacobian matrix, which is considered as the value of the coefficient matrix of state variable A, and is based on the SIRD model, is obtained in (1) with the equilibrium point $(S,I,R,D) = (\frac{a+d}{r}, 0, 0, 0)$.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -(a+d) & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & a & 0 & 0 \\ 0 & 0 & 0 & 0 & d & 0 & 0 \end{bmatrix}. \tag{2}$$

A discretization of the non-linear model (1) results in the following:

$$\begin{aligned} S_k &= -rS_k^-I_k^-dt + S_k^-, \\ I_k &= \left[rS_k^-I_k^- - (a+d)\,I_k^-\right]dt + I_k^-, \\ R_k &= aI_k^-dt + R_k^-, \\ D_k &= dI_k^-dt + D_k^-. \end{aligned} \tag{3}$$

Suppose we have a system model $\dot{\hat{x}}$ and a measurement model z_k, the estimation result $\hat{x}_k$ can be obtained using the previous EKF algorithm in [9]. For the prediction stage using the modified EKF, we define $n = k + i$, where the time step is $i = 1, 2, ...$ Subsequently, the initial parameter values r_0, a_0, d_0, and S_0 are estimated using the genetic algorithm. The initial real data for I_0, R_0, D_0, are used to give the initial values in applying the modified EKF.

3 A Genetic Algorithm-SIRD

3.1 A genetic algorithm

A genetic algorithm (GA) is an algorithm that mimics a natural evolutionary model by using a genetic inheritance [10]. Chromosomes and fitness functions should be made before applying the algorithm. The chromosomes will be a solution to the problem addressed, and the fitness functions will be a tool to measure the value of a chromosome.

The genetic algorithm begins with creating an initial population containing several chromosomes as the solution to the problem. Usually, these allegations are chosen randomly from the points scattered within the search space. Then the genetic algorithm uses crossover and mutation operators to process the chromosomes in the population until it converges or finds the best results [11].

The crossover operator allows a merging of information from two or more chromosomes to form a new chromosome. The mutation operator is used to explore the search space even further in hopes of obtaining better chromosomes. A new population will be formed after the crossover and mutation have been applied to chromosomes in the initial population. Accordingly, a generation of the new population will increase by one level. The crossover and mutation processes continue until a certain number of iterations is exceeded or the termination criteria are met [11].

A genetic algorithm - SIRD (GA-SIRD) uses the SIRD model on daily data classified as the infected individual I, recovered individual R, and deceased individual D, which is solved by the genetic algorithm. In this study, we used the Indonesia daily data from March 2, 2020 until August 25, 2020 to obtain the initial values of S_0, I_0, R_0, and D_0, and the parameter values of r, a, and d.

Stages of the genetic algorithm - SIRD consist of:

1. Input
   The input used in the genetic algorithm is the daily data I, R, and D from Indonesia, which starts from March 2, 2020 to August 25, 2020 (176 days). Next, the values will be called actual I_k, R_k, D_k with $k = 1, 2, \ldots, n$ and $n = 176$. Finally, these values will be used in the process of calculating the fitness value of a chromosome.

2. Chromosome
   In this study, we used chromosomes in the form $x = (x_1, x_2, \ldots, x_7) \in \mathbb{R}^7$. Each element of x represents one parameter of the SIRD model, which is $x_1 = S_0$, $x_2 = I_0$, $x_3 = R_0$, $x_4 = D_0$, $x_5 = r$, $x_6 = a$, and $x_7 = d$. In this genetic algorithm, we use an initial population with 100 chromosomes that are made randomly over a certain range.

3. Fitness Function
   The purpose of the genetic algorithm is to find a chromosome that minimizes the difference between the actual I_k, R_k, D_k and predicted I_k, R_k, D_k. To obtain the predicted I_k, R_k, D_k, we use the discretization model (3) for $k = 1, \ldots, n-1$. For $k = 0$, we use $S_1 = S_0$, $I_1 = I_0$, $R_1 = R_0$, $D_1 = D_0$. In this case, because the data used is daily data, we use $\Delta t = 1$. Suppose $x = (x_1, x_2, \ldots, x_7)$ is the chromosome for which the fitness function is calculated. Using $S_0 = x_1$, $I_0 = x_2$, $R_0 = x_3$, $D_0 = x_4$, $r = x_5$, $a = x_6$, and $d = x_7$, we can calculate the predicted S_k, I_k, R_k, D_k for $k = 1, 2, \ldots, n$. After we get the prediction of I_k, R_k, D_k for $k = 1, 2, \ldots, n$,

we calculate the difference with the actual I_k, R_k, D_k using the RMSE (Root Mean Square Error). Suppose that the predicted I_k, R_k, D_k are symbolized by $\hat{I}_k$, $\hat{R}_k$, $\hat{D}_k$, then the fitness function of x is

$$RMSE(x) = \sqrt{\frac{\sum_{k=1}^{n}(\hat{I}_k - I_k)^2 + (\hat{R}_k - R_k)^2 + (\hat{D}_k - D_k)^2}{3n}}.$$

The final solution of the genetic algorithm is a chromosome in the population that has the smallest RMSE (fitness function) value.

4. Crossover
   The purpose of this kind of crossover is to take all the profits and get rid of all the losses. With this step, it can guarantee that the new chromosome is definitely better than or the same as the previous chromosome.

5. Mutation
   A mutation is performed on chromosomes in the population with a pm chance. For example, $x = (x_1, x_2, \ldots, x_7) \in \mathbb{R}^7$ is the chromosome to be mutated. First, we calculate the fitness function of x. Then x_1 at x is replaced with $x_1 + \epsilon$, where $-\frac{1}{2(u+1)} < \epsilon < \frac{1}{2(u+1)}$ and u is the current generation of the genetic algorithm. Then the fitness function from the new x is calculated. If the fitness function is better, then $x_1 + \epsilon$ is used to replace x_1. If the fitness function is worse, then $x_1 + \epsilon$ is replaced again with x1. Next, the same operation/step is performed on x_2, x_3, $\ldots$, x_n. This step can be guaranteed that the new chromosome is better than or the same as the previous chromosome.

3.2 SIRD optimal parameters based on the genetic algorithm

The genetic algorithm has several main parameters, for example, the number of chromosomes, the number of iterations, and the chance of mutations. These parameters can vary depending on the complexity of the problem. Selection of the interval in the determination of search space (domain) also dramatically affects the final results of the genetic algorithm. The wider the search space created, the more difficult the genetic algorithm for converging. Conversely, a too-small search space often results in genetic algorithms not converging to the optimum solution.

In this study, we initialize the number of chromosomes is 100, the number of iterations is 10000, and the minimum value of each element in a chromosome is 0. Moreover, the maximum value of each element can be seen in Table 1. Those intervals are obtained by

Table 1: The maximum value of each element in a chromosome.

Parameter	Maximum Value
x_1	1000000
x_2	1000
x_3	100
x_4	1000
x_5	10^{-5}
x_6	10^{-1}
x_7	10^{-2}

conducting several trials. In addition, we use $p_m = 1$ since there is a guarantee that the new chromosome obtained from the mutation is better than or the same as the previous chromosome.

Based on the steps explained before, we have found the best parameters for the SIRD model that fit actual data of the infected I, recovered R, and deceased D individuals. The results obtained from the genetic algorithm are shown in Table 2. With those values, we will get a good enough SIRD model, where RMSE is equal to 836.0047.

Table 2: The parameters found by the genetic algorithm.

Parameter	Value
S_0	221269.8187
I_0	305.9698464
R_0	0
D_0	433.6840396
r	3.57028E-07
a	0.033376181
d	0.0024362

4 An Ensemble Model-SVR

Machine learning has three major learning problems such as: (1) supervised learning, (2) unsupervised learning and (3) reinforcement learning. In supervised learning, input data will be mapped to a particular output value. Supervised learning comprises two tasks: classification for categorical values and regression for continuous values based on the output domain. Following the nature of this application, we focus on regression. A bunch of regression models in machine learning has succeeded in solving many real-world problems, e.g., electricity consumption forecasting [12], electric load forecasting [13], a bus passenger forecasting [14], and high-frequency stock return forecasting [15].

In general, we may have difficulty stating the best regression model, which depends on the domain of applications. Hence, a study on model comparison is an essential step when dealing with a new problem. For instance, particular five regression models reviewed on electricity consumptions showing that a Linear Regression (LR) has better accuracy [12] and selected six supervised methods compared for a residential energy consumption prediction indicating the accurate one is a Gradient Boosting [16]. Moreover, one of the well-known regression models is a Support Vector Regression (SVR). The method was introduced by Harris Drucker *et al.* in 1996 [17]. The SVR is an extended version of a Support Vector Machine (SVMs) [1]. Hence, the procedure in SVR is similar to the SVM, with the main difference on the target function called the *regressor*.

Let $\mathsf{X} = \{\mathsf{x}_1, \ldots, \mathsf{x}_{|\mathsf{X}|}\}$ be an input set that consists of $\mathsf{x} = (i_n, r_n, d_n)$ and $1 \leq n \leq t$ and $\mathsf{Y} = \{\mathsf{y}_1, \ldots, \mathsf{y}_{|\mathsf{Y}|}\}$ be an output set that contains prediction results, $\mathsf{y} = (i_m, r_m, d_m), t+1 \leq m \leq T$. In this study, i_n and i_m represent the number of infected cases, r_n and r_m represent the number of recovered cases, and d_n and d_m represent the number of death cases. We want to find three functions f_i, f_r and f_d for each COVID-19 case. For simplification, we will map each feature of x to a certain value of y, $\mathbf{f} : \mathsf{X} \to \mathsf{Y}$. Considering on linear problems, we have $\mathbf{f} = w \cdot \mathsf{x} + b$ with w and b being the weight and

[1] A SVM is one of the classification models in machine learning.

bias parameters, respectively. In the SVR, we want to obtain support vectors which are close to $\mathbf{f}$. To do so, we create two margins that are close enough to $\mathbf{f}$ by minimizing the norm-value of $w \cdot w^T$. The problem can be formulated as a convex optimization below:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} w \cdot w^T = \min \frac{1}{2} \|w\| \\ \text{subject to} \quad & \mathsf{y} - w \cdot \mathsf{x} - b \leq \epsilon, \\ & w \cdot \mathsf{x} + b - \mathsf{y} \leq \epsilon. \end{aligned} \tag{4}$$

To deal with infeasible constraints, we add slack variables ζ and ζ' for each point called *the soft margin*. As a result, (4) can be written as a primal formula as follows:

$$\begin{aligned} \text{minimize} \quad & \mathcal{J}(w) = \frac{1}{2} w \cdot w^T + C \sum (\zeta + \zeta'); \\ \text{subject to} \quad & \mathsf{y} - w \cdot \mathsf{x} - b \leq \epsilon + \zeta, \\ & w \cdot \mathsf{x} + b - \mathsf{y} \leq \epsilon + \zeta', \\ & \zeta, \zeta' \geq 0, \end{aligned} \tag{5}$$

where C is a positive numeric value that assists in avoiding overfitting. Furthermore, we use the Lagrange dual formulation to save the computational time on solving the problem in (5). Let α and α' be non-negative multipliers. The dual formulation of (5) is described as follows:

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j') \mathsf{x}_i \mathsf{x}_j' + \epsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i') + \sum_{i=1}^{N} \mathsf{y}_i (\alpha_i - \alpha_i'); \\ \text{subject to} \quad & \sum (\alpha - \alpha') = 0, \\ & 0 \leq \alpha \leq C, \\ & 0 \leq \alpha' \leq C. \end{aligned} \tag{6}$$

However, this study problem is considered a nonlinear one. In this case, we replace the dot product of $\mathsf{x} \cdot \mathsf{x}'$ with a kernel function $K(\mathsf{x}, \mathsf{x}')$ that transforms x to high-dimensional space. There are several kernel functions: linear, Gaussian (or radial basis function), and polynomial.

This part shows the COVID-19 cases predictions using several regression methods. Based on the comparison results, we use the best method to predict COVID-19 cases in Indonesia for the long term by explaining the scenario, and the result will be shown in Section 4. The data set was collected starting from March 2, 2020, until August 25, 2020. Furthermore, the data set is split into a training set $\mathcal{D}$ from the first day until the 141^{th} day and testing set T from the 142^{st} day to the 176^{th} day. In this study, we specified the parameters of the SVR $(C, \alpha, \epsilon) = (1, 0.1, 0.1)$. We use a radial basis function as the kernel parameter. Furthermore, the SVR will be compared with a Decision Tree (DT), a K-Nearest Neighbor (KNN), a Linear Regression (LR), a Gaussian Process Regression (GPR), and a Long Short Term Memory (LSTM) with ten-time steps as the mini-batch size. We utilized a mean absolute percentage error (MAPE) below for evaluating the performance

$$MAPE = \frac{1}{|\mathsf{X}|} \sum_{k=1}^{|T|} \left| \frac{\mathsf{x}_k - y_k}{\mathsf{x}_k} \right|. \tag{7}$$

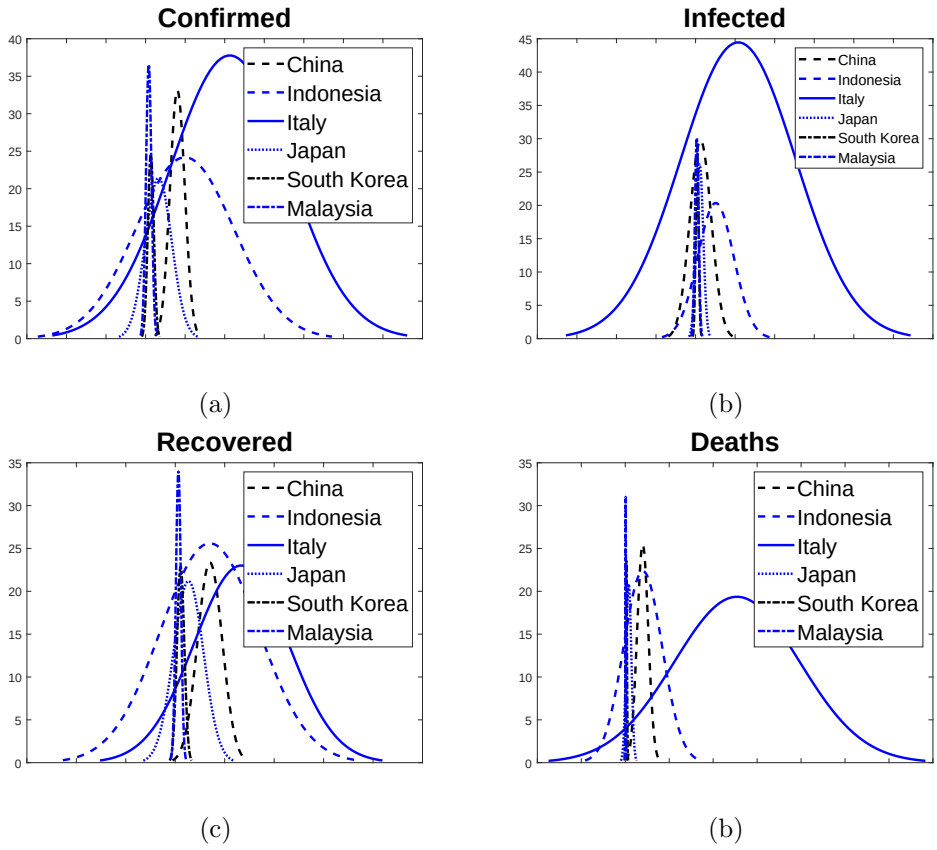

Figure 1: Data distribution of COVID-19 cases in several countries: (a) Confirmed cases, (b) Infected cases, (c) Recovered cases and (d) Death cases.

Overall, the GPR has better MAPE results than other models, as shown in Table 3. Both the DT and the LSTM did not display as the most accurate compared to the others. It is because the DT required a discretization step, leading to predicting the number of cases imprecisely, and the LSTM cannot update the learning parameters properly from insufficient information (only learn from small data). According to Table 3, a spatial factor is essential since some models showed different results on each country toward the cases. The GPR is dominant over others for South Korea. The SVR outperformed others for Indonesia and India. Therefore, this study further developed the SVR to forecast the spread of COVID-19 cases in Indonesia in Section 5.

In a naïve way, predicting long future data will take the output y_{k+1} at each one step ahead as the one to be foreseen y_{k+2}. However, it might be hard to use the SVR to forecast the spread of COVID-19 in Indonesia. To the best of our knowledge, the infected case number in Indonesia is still not reduced up to August 25, 2020; thus, the SVR may provide either a sudden fall dramatically or a rise up continuously. To overcome the issue, we have drawn long-term data (after August 25, 2020) for each data class of COVID-19 in Indonesia based on some countries data distributions that have passed the peak of the

infected cases. More specifically, three countries, i.e., China, South Korea, and Malaysia, are examined. The data distribution for each data class is depicted in Figure 1.

From Figure 1(a), we can state that the number of the infected cases in Indonesia is close enough to that in South Korea and Malaysia. South Korea has no massive restrictions (or lockdowns), yet high awareness of both government and citizens plays an important role. As a result, South Korea has already passed the peak after a rapid spread (hit around 10.000 infected cases). Malaysia chose to apply a massive lockdown that has already made them pass the peak faster (less than 6000 infected cases). In addition, China is being considered in the ensemble model where the COVID-19 was first found. China was also the first country where the massive lockdown was applied to suppress the infected cases, and the country has already passed the peak of infected cases. As the population number in China is more extensive than in Indonesia, this study considers scaling on the population number when ensemble with the China COVID-19 model. The details are described in Section 5.

5 Simulation and Discussion

Simulation of the spread of the novel coronavirus COVID-19 in Indonesia is conducted based on official data released by the government of Indonesia and collected from `https://covid19.go.id/`. Data presented include the number of confirmed cases $N(t)$, number of infected cases $I(t)$, number of recovered cases $R(t)$, and number of death cases $D(t)$. The data is collected from March 2, 2020 until August 25, 2020. The simulation employs three methods. Firstly, an Ensemble-SVR method will be applied to the data to predict the growth of the outbreak. Secondly, a modified EKF-SIRD method and GA-SIRD method are used to describe the outbreak's dynamics.

Since the methods are applied based on the SIRD model, the number of infected cases $I(t)$ is obtained by $N(t) - R(t) - D(t)$. Infected cases data will be used as the input in variable infected $I(t)$. The initial values of the parameters $I(0)$, $R(0)$, $D(0)$ are taken from the first time step of data and, in particular, are used by the modified EKF-SIRD method. The initial value of $S(0)$, parameters r, a, and d for the modified EKF-SIRD are estimated using the genetic algorithm. All parameters and the initial values are estimated using the genetic algorithm to put in the GA-SIRD method. For the Ensemble-SVR, the initial value is not required. The values are obtained directly from a random generator in the method.

The first results presented in Figure 2 are based on the Ensemble-SVR and show the prediction of COVID-19 spread for infected individuals ($I(t)$), recovered individuals ($R(t)$), and deceased individuals ($D(t)$). The simulation is conducted by, first, determining an ensemble model based on China, South Korea and Malaysia [2]. The three countries are chosen based on data characteristic analysis on the data distributions towards Indonesia data in Figure 1.

There are some differences of the simulation results based on the three countries, namely, China (see Figure 2(a)), South Korea (see Figure 2(b)), and Malaysia (see Figure 2(c)). The figures describe the possibilities of the dynamic of the outbreak as the data characteristics are similar to the Indonesia data.

The data reflects its conditions in China: many people at risk of infection and a total lockdown policy. Different policies are applied in South Korea. There is no lockdown policy but a high level of discipline in applying social distancing, mask-wearing, a vast

[2] Data is collected from `https://www.worldometers.info/coronavirus/#countries`.

Table 3: MAPE comparison of several regression models on COVID-19 cases prediction.

COUNTRY	INFECTED						DEATH						RECOVERED					
	KNN	GPR	LSTM	DT	LR	SVR	KNN	GPR	LSTM	DT	LR	SVR	KNN	GPR	LSTM	DT	LR	SVR
CHINA	0.1448	**0.1368**	0.2421	0.3515	25.8119	6.7098	**0**	0.0001	0.9985	0.00025	0.2083	0.0383	0.0113	**0.00098**	0.9736	0.002	0.0544	0.0544
SOUTH KOREA	0.1324	**0.0172**	0.8479	0.0949	1.562	0.2223	0.0358	**0.0053**	0.9885	0.0413	0.1233	0.0639	0.0419	**0.0011**	0.5792	0.0566	0.1384	0.0565
ITALY	0.1371	0.1256	0.9968	0.1592	1.0285	**0.0846**	0.0435	**0.0031**	0.9993	0.0544	0.0274	0.0551	0.1935	**0.0054**	0.9964	0.2408	0.1387	0.0215
MALAYSIA	0.1182	**0.0798**	0.9009	0.1174	0.3738	0.0946	0.0328	**0.0034**	0.9823	0.0403	0.1384	0.0481	0.1524	0.0536	0.0312	0.1658	0.0562	**0.0481**
INDONESIA	0.3106	0.0541	0.9938	0.3561	0.1146	**0.0305**	0.3062	0.0701	0.9877	0.3672	0.2746	**0.0222**	0.5379	0.7188	0.933	0.5945	0.1001	**0.008**
SPANYOL	0.0343	**0.0084**	0.9979	0.0305	0.3207	0.0091	0.0104	**0.0072**	0.9992	0.0186	0.0325	0.0471	**0**	0.0041	0.996	0.0227	0.2887	0.0144
US	0.113	**0.018**	0.9999	0.1358	0.1702	0.0407	0.1431	**0.0236**	0.9998	0.1906	0.3345	0.0308	0.3883	0.067	0.999	0.4117	0.0836	**0.0372**
UK	0.1491	0.1777	0.9995	0.1818	**0.1442**	0.1443	0.8012	**0.4838**	0.9061	1.0183	16.3876	1.2935	0.1145	0.1	0.9971	0.1316	**0.0323**	0.0324
INDIA	0.3127	0.0228	0.9993	0.3758	0.0888	**0.0162**	0.4081	0.3915	0.9887	0.4515	0.1268	**0.0153**	0.4027	0.6866	0.9994	0.4755	0.0107	**0.0102**

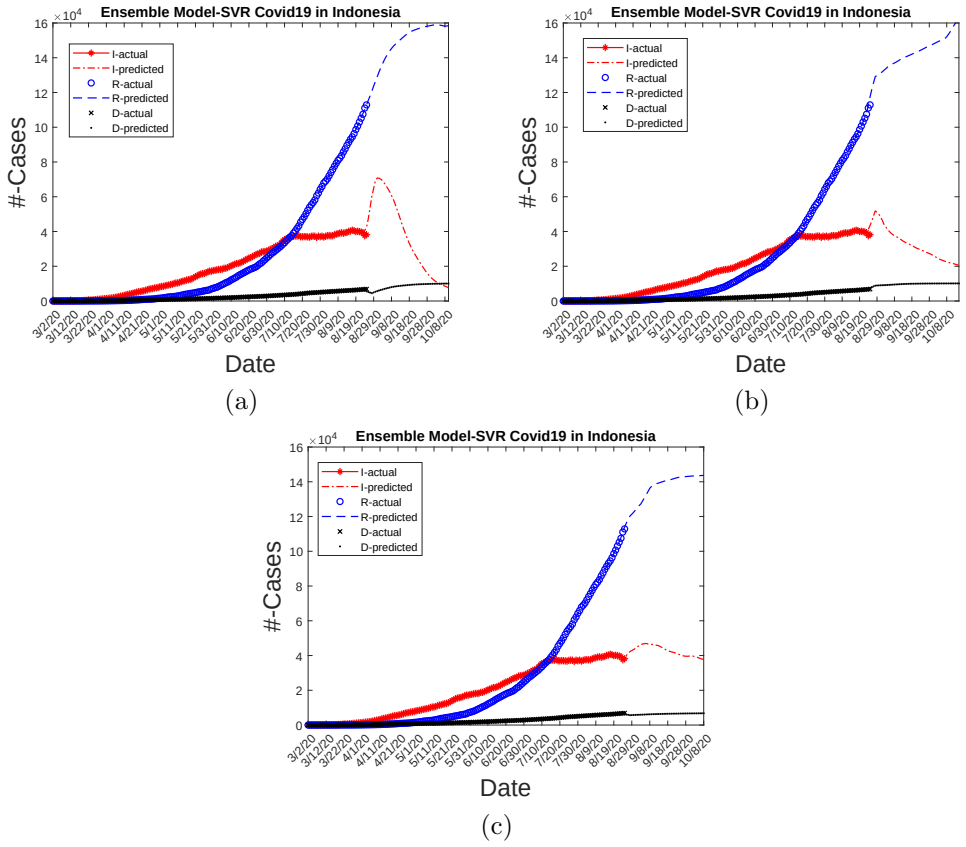

Figure 2: Ensemble Model-SVR prediction: (a) ensemble model with China, (b) ensemble model with South Korea and (c) ensemble model with Malaysia.

number of tests to its people, and readiness of its health system. As the nearest country to Indonesia, Malaysia has a total lockdown policy. The ensemble method tries to approximate the outbreak spread based on the data distribution of referenced countries. It implies that Indonesia is assumed to have similar conditions with the countries of reference. From Figure 2(a), Figure 2(b), and Figure 2(c), it can be seen that the accuracy of the method is satisfying with three possible models based on the countries of reference primarily for the next seven consecutive days.

Next, a simulation using a modified extended Kalman filter based on the SIRD model (shortly, modified EKF-SIRD) is conducted. It should be noted that the use of the modified EKF-SIRD method requires an estimation of the initial value S_0 and parameters r, a, and d, which are obtained using a genetic algorithm based on the Indonesia data. The fitting of the modified EKF-SIRD to the actual data in Figure 3 and the accuracy is relatively high for a short time prediction, approximately for the next seven consecutive days (August 26th - September 1st, 2020). If the prediction time range is longer, the accuracy will decrease as the data dynamic can not be captured well by the model. Therefore, for a longer prediction time, the parameters of the modified EKF-SIRD model should be updated to represent the dynamic of the data.

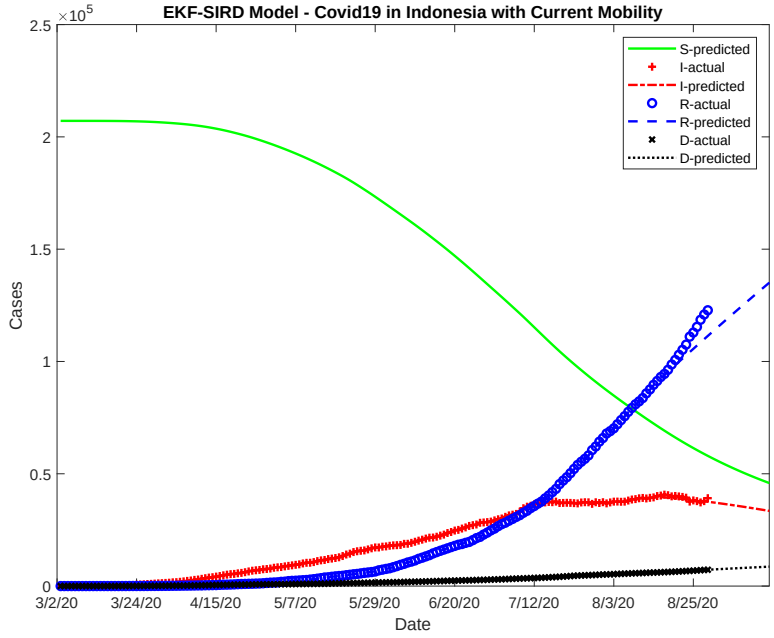

Figure 3: The dynamic of the number of infected, recovered, and deceased individual using a modified EKF-SIRD method in Indonesia.

The genetic algorithm is used not only for estimating the parameters of the modified EKF-SIRD model, but it is also used to modify the SIRD model itself. Then we name the model a GA-SIRD model. The dynamic of the SIRD model is trained by the GA algorithm based on the parameters found. However, in Figure 4, it seems that the GA-SIRD model is less capable of capturing the dynamic of COVID-19 data. The updated process of a GA-SIRD model based on the data can not perfectly follow the change of the actual data, mainly when there are extreme jumps in the actual data. This flaw results in lower accuracy in predicting the dynamics of the outbreak than the first method, the Ensemble-SVR.

This study also analyzed the three methods' performances based on the MAPE in (7), as shown in Table 4. We analyze the error trend for each class of compartments and each method as the MAPE shows different trends for those variations. In the infected compartment class, the modified EKF-SIRD and GA-SIRD methods show similar trends that the longer the time step for prediction, the higher the error. On the other hand, the MAPE of the Ensemble-SVR method does not have a trend, although the time step increases. For the recovered compartment class, both the modified EKF-SIRD and the GA-SIRD show trends that the longer the time step, the higher the errors. On the other hand, the Ensemble-SVR has a different trend than the other two. In the deceased compartment class, the modified EKF-SIRD has a more significant error when the number of time steps increases, but the GA-SIRD and the Ensemble-SVR do not show a trend of the errors.

In summary, the Ensemble-SVR method shows different behavior, as the prediction

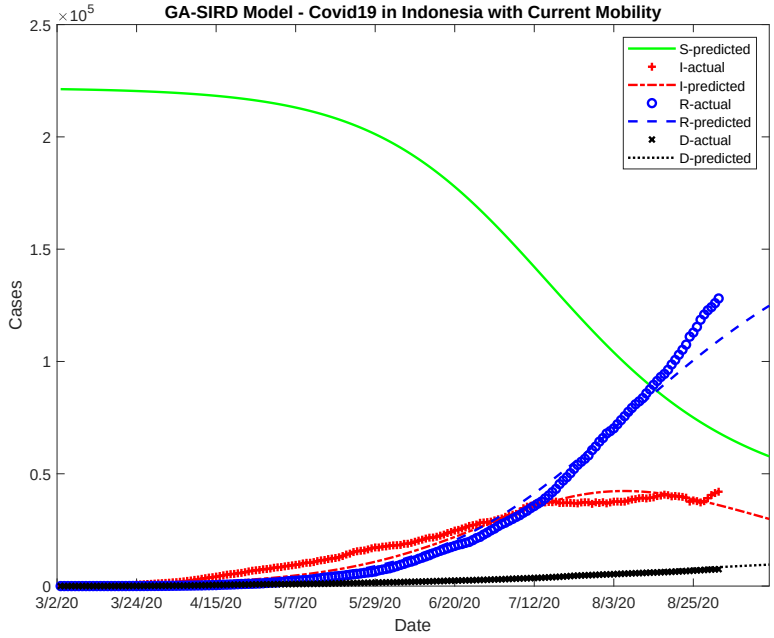

Figure 4: The dynamic of the number of infected, recovered, and deceased individual using a modified GA-SIRD method in Indonesia.

is made based on the data distribution and the results are long-term data randomly generated. Therefore, the method is not affected by the number of time steps. For the modified EKF-SIRD and GA-SIRD, predictions are likely to depend on the number of time steps as both methods have parameters used in the methods. The parameters do not change, although the number of time steps increases. The parameter changes only if the data set changes as the methods used herein are data-driven. Being considered time-dependent, the modified EKF-SIRD and GA-SIRD are less accurate if the number of time steps increases. That is, in general, the Ensemble-SVR outperforms the two other methods.

Table 4: MAPE comparison of the proposed models on forecasting the COVID-19 cases in Indonesia.

Time steps	INFECTED			DEATH			RECOVERED		
	EKF-SIRD	GA-SIRD	Ensemble SVR	EKF-SIRD	GA-SIRD	Ensemble SVR	EKF-SIRD	GA-SIRD	Ensemble SVR
$k = 1$	0.0029	0.0141	0.0143	0.0003	0.1325	0.0316	0.0118	0.1179	0.0203
$k = 2$	0.0062	0.0198	0.0225	0.0057	0.1265	0.0320	0.0261	0.1307	0.0186
$k = 3$	0.0144	0.0054	0.0222	0.0091	0.1229	0.0327	0.0328	0.1369	0.0190
$k = 4$	0.0526	0.0486	0.0219	0.0107	0.1213	0.0335	0.0359	0.1400	0.0189
$k = 5$	0.0901	0.0908	0.0210	0.0109	0.1211	0.0339	0.0349	0.1396	0.0190
$k = 6$	0.1147	0.1201	0.0209	0.0101	0.1221	0.0339	0.0370	0.1420	0.0210
$k = 7$	0.1322	0.1421	0.0231	0.0113	0.1207	0.0344	0.0415	0.1465	0.0200
Average	0.0590	0.0630	0.0196	0.0083	0.1240	0.0331	0.0314	0.1362	0.0195

6 Conclusion

This paper proposes three new hybrid methods named a modified EKF-SIRD, a GA-SIRD, and an Ensemble-SVR. We simulated the first two methods to present the dynamic of the COVID-19 outbreak for short-term predictions. As a result, these two methods exhibit their dependency on the number of time steps as the accuracy decreases when the prediction time window is wider. On the other hand, the Ensemble-SVR shows that prediction accuracy does not depend on the number of time steps. Therefore, the Ensemble-SVR is the best model amongst the other machine learning methods in terms of accuracy. The study results in the conclusion that the Ensemble-SVR method outperforms the modified EKF-SIRD and GA-SIRD.

As an extension, we will continue the study in predicting an effective reproduction number R_t and dispersion number K. R_t represents the growth rate of the infection from the infected individual to the healthy individual, and K represents the ability of the infected individual to trigger new cases in a very short time. The two measures are essential for the policymaking process.

Acknowledgment

It is a tribute to our colleague, Professor Erna Apriliani who passed away on July 6, 2021.

References

[1] F. Duccio and P. Francesco. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals.* **134** (2020) 109761.

[2] H. Susanto, V. R. Tjahjono, A. Hasan, M. F. Kasim, N. Nuraini, E. R. M. Putri, R. Kusdiantara and H. Kurniawan. How Many Can You Infect? Simple (and Naive) Methods of Estimating the Reproduction Number. *Communication in Biomathematical Sciences* **3** (2020) 28–36.

[3] R. Salgotra, M. Gandomi and A.H. Gandomi. Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. *Chaos, Solitons & Fractals.* (2020) 109945.

[4] I. K. Abdulrahman. SimCOVID: An Open-Source Simulation Program for the COVID-19 Outbreak. *MedRxiv* (2020).

[5] D. Parbat and M. Chakraborty. A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons & Fractals* **138** (2020) 109942.

[6] R. Sujath, J. M. Chatterjee and A. E. Hassanien. A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment* **34** (2020) 959–972.

[7] T. Shreshth, T. Shikhar, T. Rakesh and S. S. Gill. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things* **11** (2020) 100222.

[8] N. Tutkun. Parameter estimation in mathematical models using the real coded genetic algorithms. *Expert Systems with Applications* **36** (2009) 3342–3345.

[9] F. Lewis, X. Lihua and P. Dan. *Optimal and Robust Estimation: With an Introduction to Stochastic Control Theory.* CRC press, 2017.

[10] H. Nazif and L. S. Lee. Optimised crossover genetic algorithm for capacitated vehicle routing problem. *Applied Mathematical Modelling* **36** (2012) 2110–2117.

[11] D. A. Coley. *An Introduction to Genetic Algorithms for Scientists and Engineers*. World Scientific Publishing Company, 1999.

[12] A. Gonzlez-Briones, G. Hernndez, J. M. Corchado, S. Omatu and M. S. Mohamad. Machine Learning Models for Electricity Consumption Forecasting: A Review. In: *2019 2nd International Conference on Computer Applications Information Security (ICCAIS),* 2019, 1–6.

[13] Z. Zhang, W. Hong and J. Li. Electric Load Forecasting by Hybrid Self-Recurrent Support Vector Regression Model with Variational Mode Decomposition and Improved Cuckoo Search Algorithm. *IEEE Access* **8** (2020) 14642–14658.

[14] C. Li, X. Wang, Z. Cheng and Y. Bai. Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach. *IEEE Access* **8** (2020) 19717–19725.

[15] S. Kavitha, S. Varuna and R. Ramya. A comparative analysis on linear regression and support vector regression. In: *2016 Online International Conference on Green Engineering and Technologies (IC-GET),* 2016, 1–6.

[16] A. Gonzlez-Briones, G. Hernndez, T. Pinto, Z. Vale, and J. M. Corchado. A Review of the Main Machine Learning Methods for Predicting Residential Energy Consumption.. In: *2019 16th International Conference on the European Energy Market (EEM),* 2019, 1–6.

[17] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola and V. Vapnik. Support Vector Regression Machines. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems.* Denver, Colorado, 1996, 155–161.