*COMPLETE*
*Journal of Computer, Electronic, and Telecommunication*
*ISSN 2723-4371, E-ISSN 2723-5912*
*doi.org/10.52435/complete.v1i1.48*

# The Comparative Study for Predicting Disease Outbreak

**Anifatul Faricha [1]*, M. Achirul Nanda[2], Siti Maghfirotul Ulyah[3], Ni'matut Tamimah[4], Enny Indasyah[5], Robin Addwiyansyah Alvaro Samrat[6]**

[1] Electrical Engineering Department, Institut Teknologi Telkom Surabaya, 60231, Jawa Timur, Indonesia; faricha@ittelkom-sby.ac.id
[2] Doctoral Alumni of Agricultural Engineering Science Department, Faculty of Agricultural Engineering and Technology, IPB University, Bogor 16680, West Java, Indonesia; m.achirulnanda@gmail.com
[3] Departement of Mathematics, Universitas Airlangga, 60115, Surabaya, Jawa Timur, Indonesia; maghfirotul.ulyah@fst.unair.ac.id
[4] Teknik Permesinan Kapal, Politeknik Perkapalan Negeri Surabaya, 60111, Surabaya, Jawa Timur, Indonesia; nimatuttamimah@ppns.ac.id
[5] Departemen Teknik Elektro Otomasi, Fakultas Vokasi, Institut Teknologi Sepuluh Nopember, 60111, Surabaya, Jawa Timur, Indonesia; enny_indasyah@its.ac.id
[6] Electrical Engineering Department, Institut Teknologi Telkom Surabaya, 60231, Jawa Timur, Indonesia.
* Correspondence: faricha@ittelkom-sby.ac.id or anifatulfaricha@gmail.com

**Abstract:** As the global pandemic caused by coronavirus disease is spread massively in Indonesia, proper predictive modeling is required to represent the prediction of disease outbreak. This study presents the comparative predictive modeling for predicting disease outbreak using two models i.e., optimizable support vector machine (SVM) and optimizable gaussian process regression (GPR). The dataset used in this study contains three cases i.e., positive cases, recovered cases, and death cases. The dataset at each case is divided into training dataset for the training process and external validation dataset for the validation process. Based on the training process and validation process, the root mean square error (RMSE) at positive cases, recovered cases, and death cases using optimizable GPR is substantially more effective for prediction than the optimizable SVM. According to the result performance, by applying optimizable GPR, the training process has the average RMSE of 19.54 and the validation process has the average RMSE of 15.85.

**Keywords:** Dataset; Optimizable GPR; Optimizable SVM.

## 1. Introduction

The coronavirus pandemic or COVID-19 pandemic is the ongoing pandemic which caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first outbreak was identified in December 2019 in Wuhan, China [1]. The outbreak of COVID-19 affects many sectors such as economy, logistics, transportations, etc. In Indonesia, the first case of COVID-19 was confirmed on 2 March 2020 and by 9 April 2020 was spread to all 34 provinces [2-4].

According to the data of coronavirus cases released by the Indonesia Government, a predictive modeling is required to represent the data behavior of confirmed cases in Indonesia. This study presents the comparative predictive modeling using optimizable support vector machine (SVM) and optimizable gaussian process regression (GPR). The simulation was performed using Regression Learner Application in MATLAB 2020 (Trial version). The contents of this paper are organized as follows: section 2 discusses the materials and methods used in this study. Furthermore, section 3 demonstrates the results and verification analysis. Finally, we present our conclusions in section 4.

## 2. Materials and Methods

*2.1 Materials*

The dataset used in this study was taken from the confirmed coronavirus cases information in Indonesia presented by Indonesia Government and Worldometer [1-2]. The dataset includes three cases i.e., positive cases, recovered cases, and death cases. The positive cases are the cases of people actively infected by coronavirus. Furthermore, recovered cases include cases of recovered patients from coronavirus. Finally, the death cases are the total number of death cases due to coronavirus.

In this study, the dataset is divided into two parts i.e., training dataset for the training process and external validation dataset for the validation process. The training dataset includes all the positive cases, recovered cases, and death cases started from 2nd March 2020 to 30th April 2020. Whereas, the external validation dataset includes all the confirmed coronavirus cases in Indonesia started from 1st May 2020 to 12th May 2020.

*2.2 Methods*

To represent the model of confirmed coronavirus cases dataset in Indonesia, the predictive modeling is required. In this study, we used two models to predict the coronavirus disease outbreak in Indonesia i.e., SVM regression and gaussian process regression (GPR). The SVM Regression can be categorized as nonparametric technique because it depends on the kernel function selections [7]. Generally, SVM Regression can be written in Eq (1). Furthermore, the common equation of GPR is determined in Eq (2) [8-9].

$$f(x) = \sum_{n=1}^{N} (a_n - a_n^*) G(x_n, x) + b \tag{1}$$

Where: f(x) is response value (total cases), x is predictive value (number of day), N is the amount of dataset, a and $a^*$ are the lagrange multiplier, G is kernel function, b is betha parameter.

$$P(y_i \mid f(x_i), x_i) \sim N(y_i \mid h(x_i)^T \beta + f(x_i), \sigma^2) \tag{2}$$

Where: $P(y_i \mid f(x_i), x_i)$ is density function, β is betha parameter, $\sigma^2$ is noise variance, h is explicit basis function.

In this study, the simulation was performed using Regression Learner Application in MATLAB 2020 (Trial version). In the Regression Learner Application in MATLAB, the optimizable SVM contains the selection of several kernel functions such as linear, quadratic, cubic, fine Gaussian, medium Gaussian, coarse Gaussian, etc [5]. Furthermore, the optimizable GPR includes selection of several kernel functions such as rational quadratic, squared exponential, mattern 5/2, exponential, etc [6]. Both the optimizable SVM and optimizable GPR use two processes to build the optimal predictive model i.e., training process and external validation process. The root mean square error (RMSE) and average RMSE are applied to evaluate the optimal predictive model performance [6].

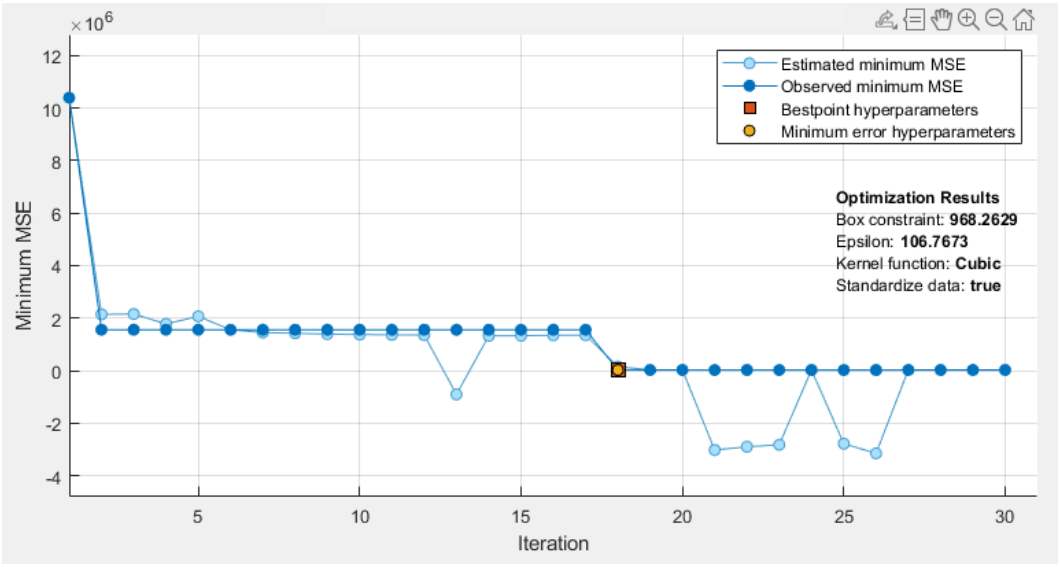
## 3. Results Analysis

There are two predictive models i.e, optimizable GPR and optimizable SVM used in this study which include two main processes i.e., training process and external validation process. Firstly, the training process by using training dataset, this process is used to find the best kernel function to obtain the optimal predictive model to represent the confirmed coronavirus cases in Indonesia.

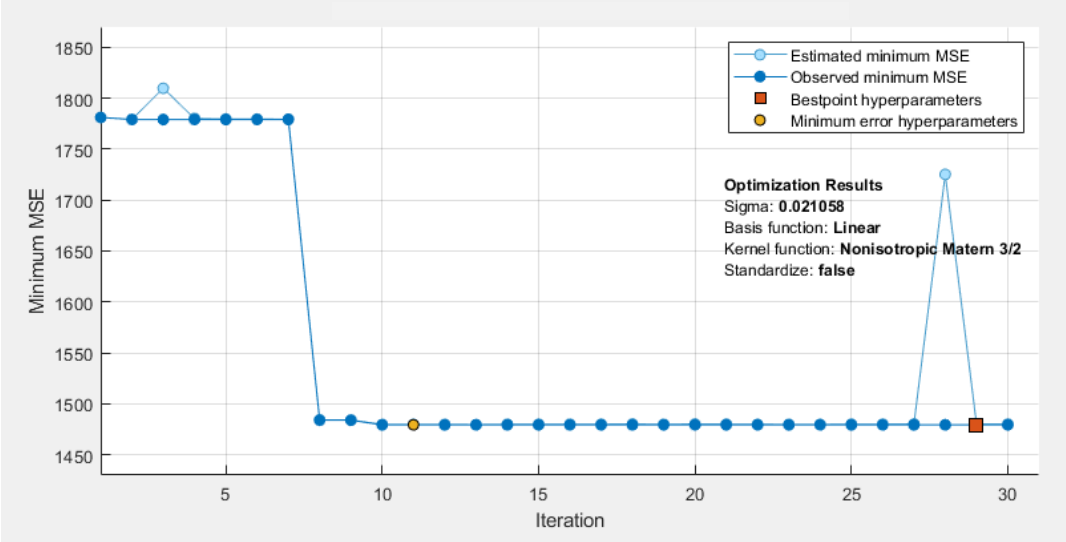
Furthermore, the external validation process includes the external validation dataset which used to verify the performance of the optimal predictive model obtained in the training process.

In the training process, the selection process to find the optimal predictive model for positive cases, recovered cases, and death cases are shown in Figure 3, Figure 4, and Figure 5 respectively. Using optimizable SVM, the best kernel function was achieved by cubic function for all confirmed coronavirus cases in Indonesia i.e, positive cases, recovered cases, and death cases. Furthermore, for optimizable GPR, the best kernel function was obtained by nonisotropic matern function for positive cases and isotropic matern function for recovered cases and death cases.

As explained in section of materials and methods, in this study, we used the training dataset from confirmed coronavirus cases in Indonesia started from 2nd March 2020 to 30th April 2020. Figure 6 shows the comparison of predictive modeling in the training process using optimizable SVM and optimizable GPR. According to Figure 6, it is clearly seen that the optimizable GPR is more able to represent the training dataset for positive cases, recovered cases, and death cases than optimizable SVM.

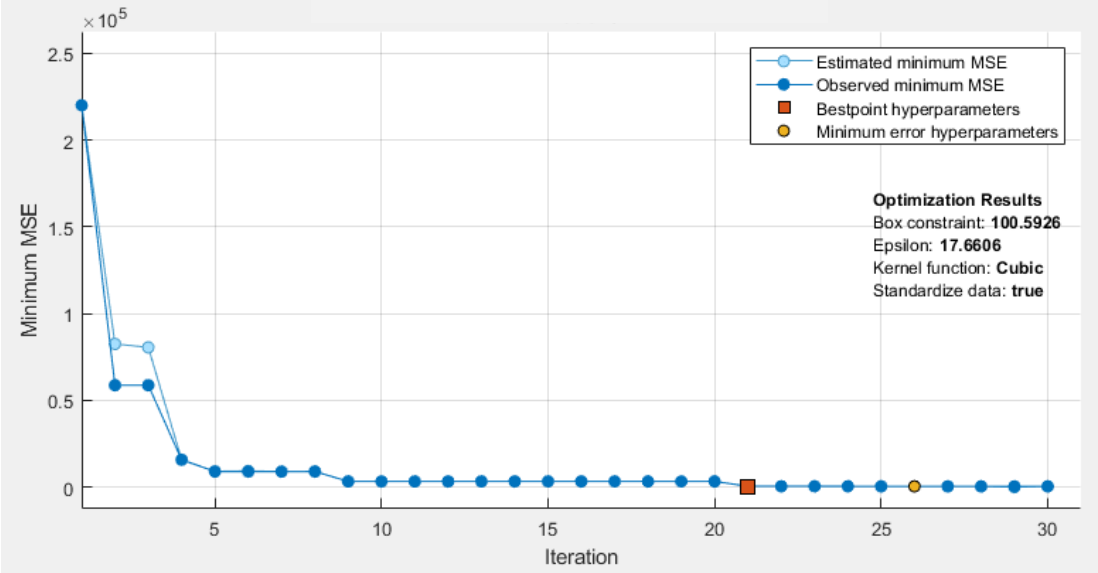



**(a)**

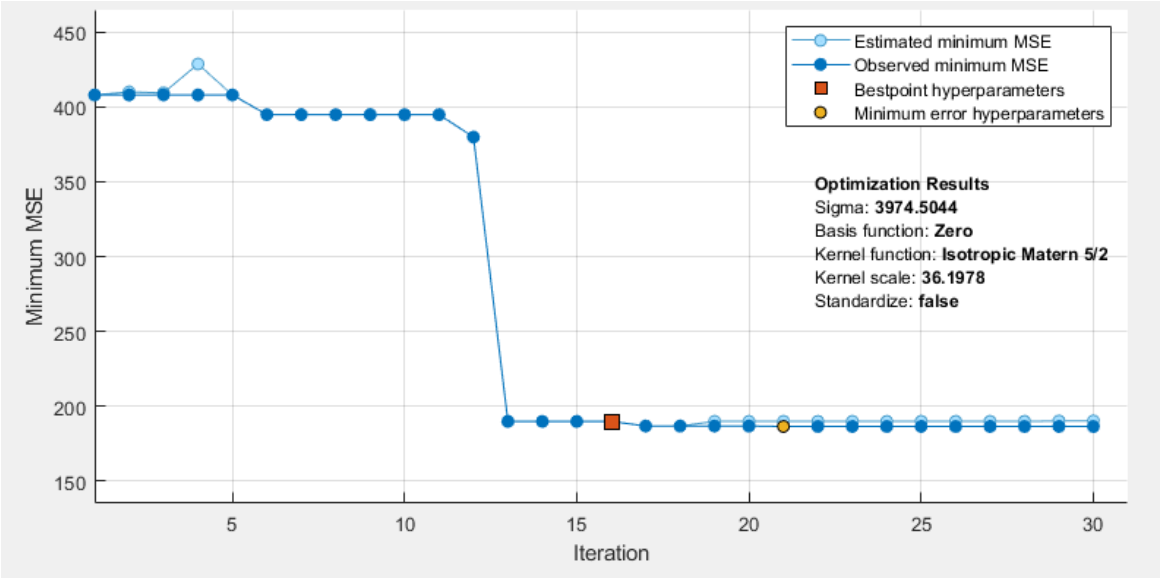



**(b)**

**Figure 1** Searching for the optimal predictive model parameter for positive cases using: **(a)** Optimizable SVM; **(b)** Optimizable GPR.



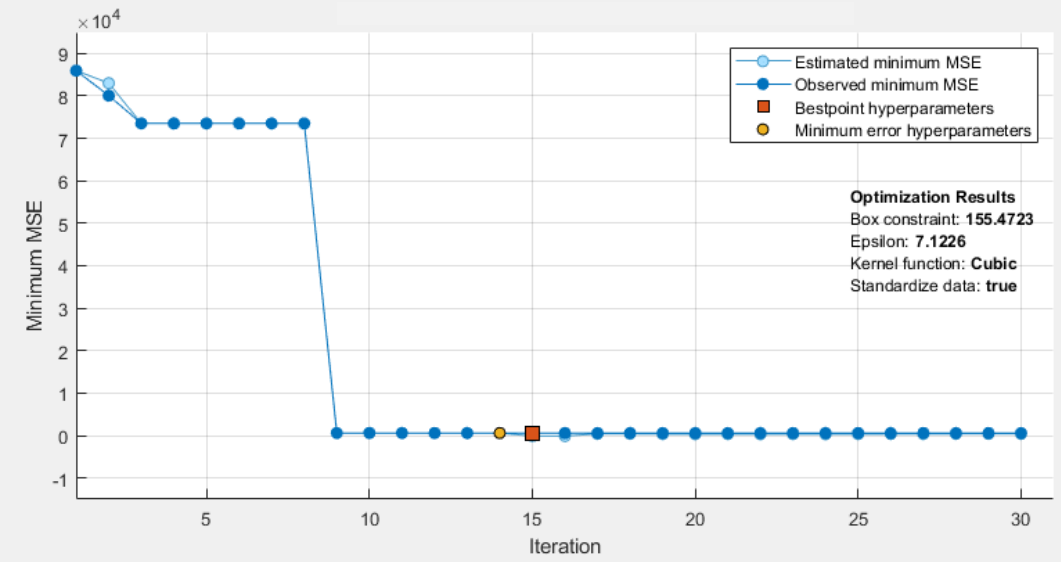

**(a)**



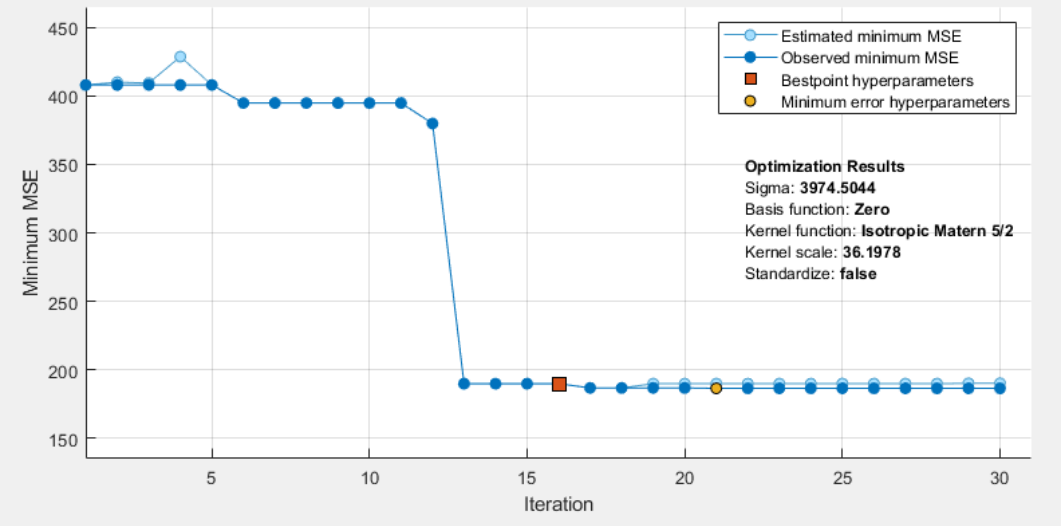

**(b)**

**Figure 2** Searching for the optimal predictive model parameter for recovered cases using: **(a)** Optimizable SVM; **(b)** Optimizable GPR.

Table 1 shows the RMSE value of several cases using optimizable SVM and optimizable GPR in the training process. According to Table 1, optimizable GPR has lower RMSE value in positive cases, recovered cases, and death cases than optimizable SVM has. Therefore, the average RMSE produced by optimizable GPR is lower than produced by optimizable SVM i.e, 19.54. This result reveals that the optimizable GPR model has more favorable prediction capability than the optimizable SVM does.



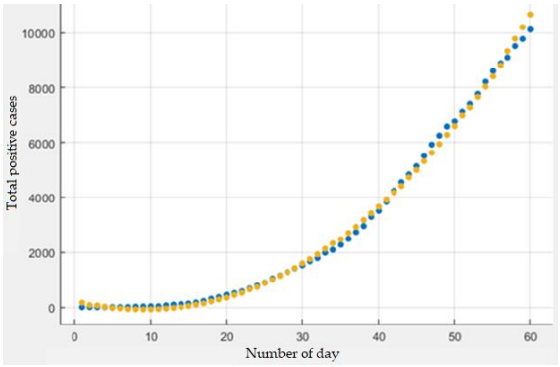

**(a)**



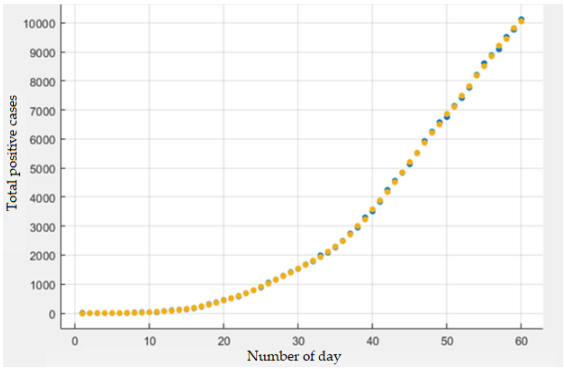

**(b)**

**Figure 3** Searching for the optimal predictive model parameter for death cases using: **(a)** Optimizable SVM; **(b)** Optimizable GPR.

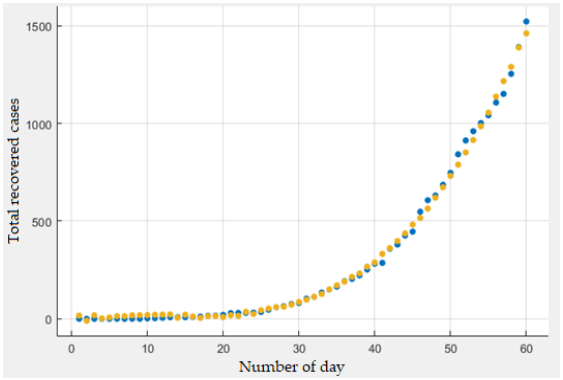
The external validation process was used for testing to assess the optimizable GPR and optimizable SVM performance. This study used the confirmed coronavirus cases in Indonesia started from 1st May 2020 to 12th May 2020 as external validation dataset. The RMSE value in the external validation process is shown by Table 2. According to Table 1 and Table 2, the average RMSE value achieved by optimizable GPR is more stable than obtained by optimizable SVM i.e., 19.54 and 15.85 for training and external validation process respectively. These results indicate that the optimizable GPR has more robust model than optimizable SVM.



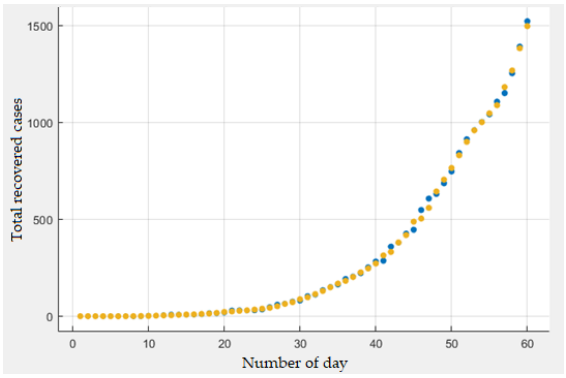

**(a1)**



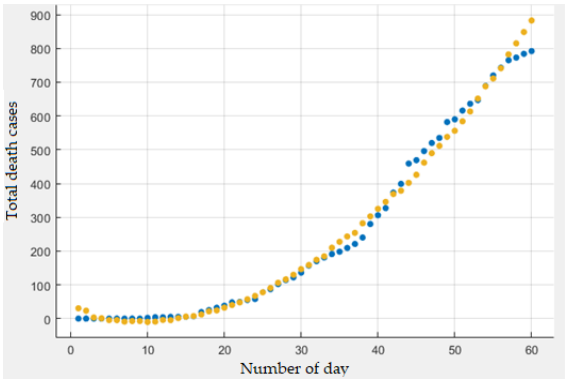

**(a2)**



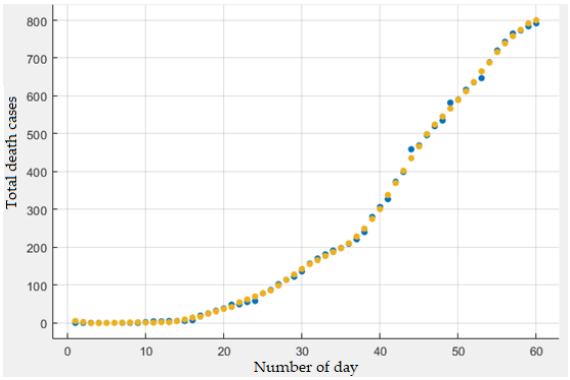

**(b1)**




**(b2)**




**(c1)**




**(c2)**

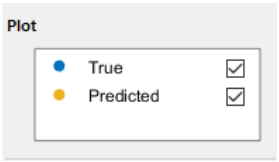



**Figure 4** The prediction curve of disease outbreak at training process using optimizable SVM model for: (**a1**) Positive cases; (**b1**) Recovered cases; (**c1**) Death cases, and using optimizable GPR model for: (**a2**) Positive cases; (**b2**) Recovered cases; (**c2**) Death cases.

**Table 1.** The RMSE value at the training process

| Method | RMSE | | | Average RMSE |
|---|---|---|---|---|
| | Positive cases | Recovered cases | Death cases | |
| Optimizable SVM | 164.90 | 22.65 | 24.68 | 70.74 |
| Optimizable GPR | 38.46 | 13.77 | 6.40 | 19.54 |

**Table 2.** The RMSE value at the external validation process

| Method | RMSE | | | Average RMSE |
|---|---|---|---|---|
| | Positive cases | Recovered cases | Death cases | |
| Optimizable SVM | 31.75 | 11.04 | 14.21 | 19 |
| Optimizable GPR | 25.04 | 9.87 | 12.65 | 15.85 |

## 4. Conclusions

This study successfully demonstrated the optimal predictive model to predict the coronavirus disease outbreak in Indonesia which includes three cases i.e., positive cases, recovered cases, and death cases. The dataset of each case is divided into training dataset and external validation dataset. There are two model applied in this study i.e., optimizable support vector machine (SVM) and optimizable gaussian process regression (GPR). According to the training process, the optimizable SVM has average RMSE of 70.74 and optimizable GPR has average RMSE of 19.54. Furthermore, in the external validation process, the optimizable SVM and optimizable GPR have average RMSE of 19 and 15.85 respectively. This finding indicates that the optimizable GPR model is more suitable to represent dataset. Therefore, the future work will concern about using optimizable GPR model to predict the coronavirus disease outbreak in Indonesia by including the distribution of location coordinates (longitude and latitude) from infected people.

**Author Contributions:** Anifatul Faricha and M. Achirul Nanda; conceptualization, methodology, and formal analysis, Siti Maghfirotul Ulyah and Enny Indansyah; validation, review, and editing, M. Achirul Nanda and Ni'matut Tamimah, and Robin Addwiyansyah Alvaro Samrat; visualization and writing.

## References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. February 2020. 395 (10223): 497–506.
2. Coronavirus cases in Indonesia. Available: https://kawalcovid19.blob.core.windows.net/viz/statistik_harian.html (accessed on 12 May 2020).
3. Worldometer: coronavirus cases in Indonesia. Available: https://www.worldometers.infocoronavirus/country/indonesia (accessed on 12 May 2020).
4. COVID-19 Pandemic in Indonesia. Available: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Indonesia (accessed on 12 May 2020).
5. Optimization in Regression Learner App. Available: https://www.mathworks.com/help/stats/hyperparameter-optimization-in-regression-learner-app.html (accessed on 12 May 2020).
6. Kernel function in Regression. Available: https://www.mathworks.com/help/stats/fitrkernel.html (accessed on 12 May 2020).
7. Vapnik, V. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.
8. Nocedal, J. and S. J. Wright. *Numerical Optimization*, Second Edition. Springer Series in Operations Research, Springer Verlag, 2006.
9. Rasmussen, C. E. and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press. Cambridge, Massachusetts, 2006.