5th International Conference on Computer Science and Computational Intelligence 2020

A Comparison: Prediction of Death and Infected COVID-19 Cases in Indonesia Using Time Series Smoothing and LSTM Neural Network

Zulfany Erlisa Rasjid[a,*], Reina Setiawan[a*], Andy Effendi[b*]

[a]Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K.H. Syahdan No. 9, Jakarta 11480, Indonesia
[b]Information Systems Department, School of Information Systems, Bina Nusantara University, Jl. K.H. Syahdan No. 9, Jakarta 11480, Indonesia

Abstract

COVID-19 is a virus causing pneumonia, also known as Corona Virus Disease. The first outbreak was found in Wuhan, China, in the province of Hubei on December 2019. The objective of this paper is to predict the death and infected COVID-19 in Indonesia using Savitzky Golay Smoothing and Long Short Term Memory Neural Network model (LSTM-NN). The dataset is obtained from Humanitarian Data Exchange (HDX), containing daily information on death and infected due to COVID-19. In Indonesia, the total data collected ranges from 2 March 2020 and by 26 July 2020, with a total of 147 records. The results of these two models are compared to determine the best fitted model. The curve of LSTM-NN shows an increase in death and infected cases and the Time Series also increases, however the smoothing shows a tendency to decrease. In conclusion, LSTM-NN prediction produce better result than the Savitzky Golay Smoothing. The LSTM-NN prediction shows a distinct rise and align with the actual Time Series data.

* Corresponding author. Tel.: +62 818764566;
  E-mail address: zulfany@binus.ac.id

1. Introduction

COVID-19 is a virus causing pneumonia, also known as Corona Virus Disease. The first outbreak was found in Wuhan, China, in the province of Hubei on December 2019[1,2]. In Indonesia, the first outbreak was first reported on 2 March 2020 and by 26 July 2020, the spread of the infected individuals have reached more than 90,000 people based on dataset from Humanitarian Data Exchange (HDX)[3] , and currently have infected over 200 countries all over the world[4].

 Researches following this outbreak have conducted on many aspects, such as supply chain[5], medical, predictions in China, Italy, and France[6], also predictions using Time Series[7] and the effects of COVID-19 in the field of economy[8], education[9], the risk of being infected when individuals have chronic diseases[10], and others. In Indonesia, most research is the study of the effect of COVID-19 in several sectors such as education[11,12,13,14], economy[15,14,16], tourism[17,18], not many researches has been conducted on prediction and forecasting methods. Some research related to prediction was conducted by Nuraini et al. using a model with Richard's Curve[19], Rustan and Handayani researched on prediction using a modified SEIR (Susceptible Exposed Infectious Recovered) model[20], Parhusip uses SVM, Bayesian and Gaussian method[21], Arianto and Noviyanti uses Back Propagation and Fuzzy[22] .

This paper is to compare two methods for the prediction and forecast of the outbreak of COVID-19 in Indonesia using Time Series with Savitzky Golay smoothing and Long Short Term Memory Neural Network (LSTM-NN) in order to get the most accurate prediction of death and infected cases of outbreak and predict which method will provide a more accurate result. The results show that using both models the trend for death and infected cases are still increasing.

2. Related Works

The Susceptible-Infected-Recovered Model (SIR) has been used to analyze epidemic[23,24] and used to analyze and perform prediction in diseases such as the Ebola Virus[25], SARS[26] and covid-19[27]. Another model, the Susceptible-Exposed-Infectious-Recovered (SEIR) Model is also used in predicting infectious diseases[28,29].

Analysis of the spread of Covid-19 has been conducted by Fanelli and Piazza (March 2020) using a Simple Susceptible-Infected-Recovered-Death Model (SIRD) covering the spread of covid-19 in China, Italy and France. The results showed that the rate of recovery in the three countries mentioned above is relatively stable, however, with regards to infection and death rates, it is quite variable[6].

Yang, Zifeng et al., uses SEIR method modified with the addition of In(t) and Out(t) parameter, in order to create the graph of the epidemic. The in(t) and Out(t) is used to measure the inflow and outflow of and using Artificial Intelligence (AI) to perform the prediction on the epidemic. The graph is then used to predict the epidemic using AI, the Long-Short-Term Memory (LSTM) model, a type of the Recurrent Neural Network (RNN), trained using the SARS-CoV-2 data. The prediction result is within expectations[1].

A research by Vaisha, R. et al., analyzes seven important AI applications for COVID-19 pandemic, and they are summarized in Table 1[30].

Table 1. AI Seven Application for Covid-19 Pandemic[30]

No.	AI Application
1.	Early Detection and Diagnosis of Infection
2.	Monitoring Treatment
3.	Contact Tracing of Individuals
4.	Projection of cases and mortality
5.	Development of Drugs
6.	Reducing Workload of Healthcare workers
7.	Prevention of Disease

Rustam, F. et al., uses forecasting based on Machine Learning to identify the number of patients affected by COVID-19. Four models are used, that is The Linear Regression, the Least Absolute Shrinkage and Selection Operator, the Support Vector Machine and Exponential Smoothing, showing the best result using Exponential Smoothing and the worst performance using Support Vector Machine[31].

A hybrid Machine Learning model with the adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) proposed by Pinter, G. et al., in order to obtain a prediction of COVID-19 outbreak and claimed the result are promising[32]. The result can be seen in in Figure 1.
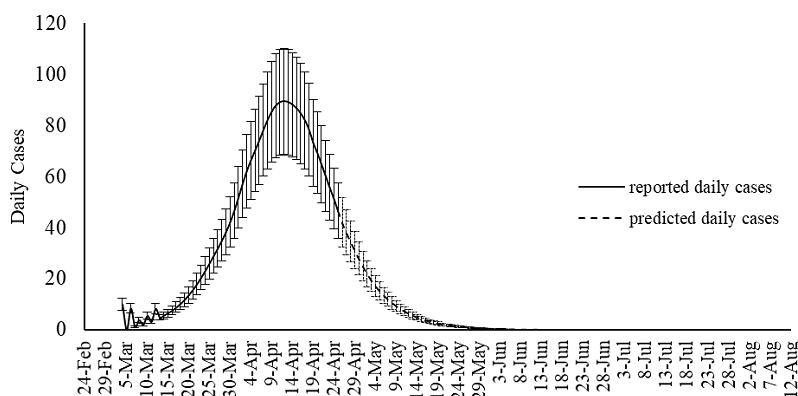

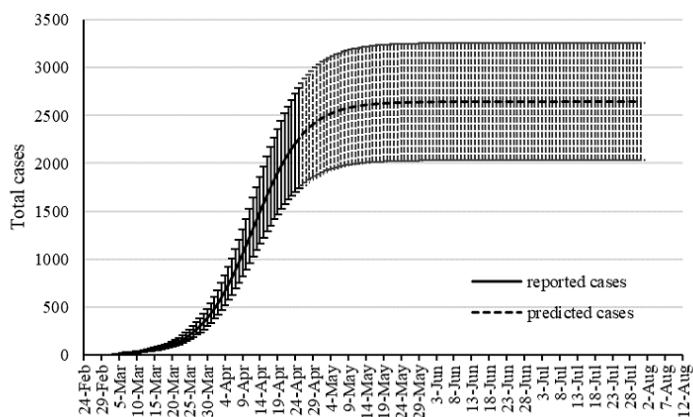Figure 1. Daily outbreak prediction rate (MPL-ICA)

Figure 2. Total outbreak prediction rate (MPL-ICA)

A hybrid Machine Learning model with the adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) proposed by Pinter, G. et al., in order to obtain a prediction of COVID-19 outbreak and claimed the result are promising[32]. The results can be seen in in Figure 1 and Figure 2. Another Hybrid AI model proposed by Zheng et al., to predict COVID-19 in China. The hybrid is prepared with the use of Susceptible-Infected(SI) model and the Natural Language Processing (NLP) and LSTM Network The result of this experiment shows smaller error compared to the basic traditional model[33]. Linear Regression is used by researchers Ghosal et al., to predict deaths in India due to COVID-19 from the first day up to 6 weeks, with a result of high adjusted R^2 values found in all of the predictive models[34].

Rustan and Handayani use SEIR model with modification by adding the factor of isolation in the equation to predict the COVID-19 outbreak in Indonesia. The SEIR model is computed and analyzed. Based on this analysis they predicted that the peak of the outbreak will be reached in May 2020[20]. A research conducted by Parhusip, proposed to predict the outbreak of COVID-19 in Indonesia and the world using Support Vector Machine with Bayesian Ridge regression. In comparison with the Gaussian model, the result using Bayesian Ridge regression shows more realistic results[21]. Based on Time Series data and the use of back propagation and Tsukamoto Fuzzy Method, researchers Arianto and Novianti showed that the simulation of the prediction of the outbreak in Indonesia has MSE of 1.632 and the correlation coefficient R = 0.843[22].

In previous researches, mostly Time Series method is used to predict the effect of COVID-19 on various aspects such as education, economy etc. In this research the Time Series is used to predict deaths and infected cases, especially for Indonesia where the population is high.

3. Method

Based on data from Humanitarian data exchange (HDX), there are 147 data ranging from 2 March 2020 until 26 July 2020. The data contains Date, Cumulative cases, Recovered cases, Total death, Patient under treatment, New case per Day, Recovered cases per Day, Death cases per Day and Treatment cases per Day. New Cases per day and Death cases per day is used in this research. 121 data (the first 4 month) are used for training, and the rest (1 month) used for testing. LSTM-NN model is applied to the training and testing dataset. Actual data is processed using Time Series forecasting with Savitzky Golay smoothing. Savitzky Golay smoothing technique is extensively used due to its capability to eliminate noise[35]. This model is applied to Death and Infected cases in Indonesia. For each case, a comparison is performed based on the actual data, the Savitzky Golay[36] smoothing data and the LSTM -NN. LSTM-NN is chosen because it is known to have a good performance for the prediction of long-sequenced data (e.g. time series data)[33]. Therefore this research compares those two methods namely the Time Series with Savitzky Golay smoothing and the LSTM-NN method.

The programming language used is python and several libraries. The algorithm for the prediction is as follows:

Algorithm: Time Series with Savitzki Golay smoothing and LSTM-NN
Input: dataset from HDX
Output: Graph of New cases and Death cases
Pre-condition: None
Post-condition: None

[1] import dataset from HDX
[2] transform dataset using minmaxscalar from sklearn.preprocessing
[3] training dataset
[4] create layers using sequential from keras.model
[5] process LSTM using LSTM from keras.layers
[6] fit model from keras.layers
[7] testing dataset
[8] process prediction
[9] process smoothing using savgol_filter from scipy.signal

The parameters used for LSTM-NN are:
     hidden layer = 50 neurons
     input = 1
     epoch = 100
     batch size = 32
     dropout=0.2

4. Results and Discussions

The data processed using Time Series with Savitzky Golay smoothing, LSTM-NN is fitted together with the Actual Time Series data. In figure 3 below, the red curve shows the actual Time Series data, the green curve shows the Time Series with Savitzki Golay smoothing and the blue curve represents the prediction using LSTM-NN. The results show that for Infected case, the trend is still rising no matter which method is used. In the actual Time Series, there is a significant fluctuation whereas from the smoothing model, it is still difficult to determine whether the prediction will increase or decrease. But the result using LSTM-NN the curve is ascending and some crossing values with the actual Time Series. Based on observation the smoothing was poor, because at the end of the curve, it is undecided whether the prediction will increase or decrease. The same increasing result can also be seen in the Death case as can be seen in Figures 4.
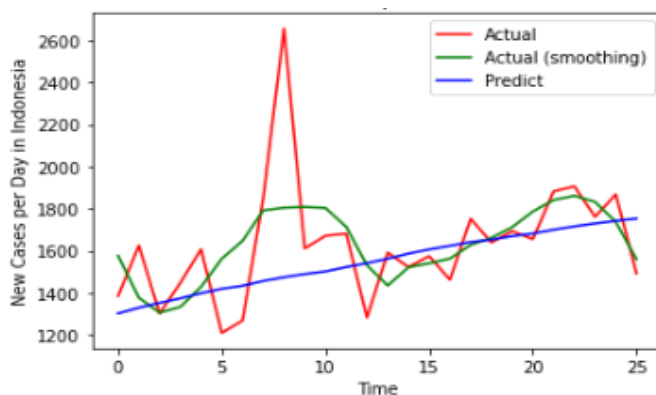

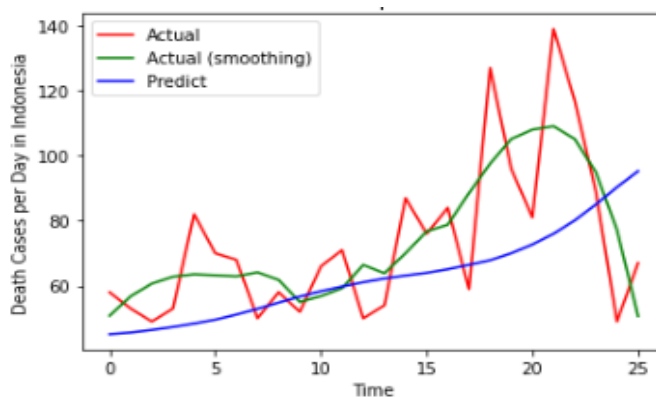Figure 3. New Cases Prediction

Figure 4. Death Cases Prediction

Similarly in Figure 4, the same condition applies, however the fluctuation is more constant than in the new casses graph, but still increasing for all three curves. Therefore it can be predicted that the death and infected will still rise in the next month.

5. Conclusion

The research on COVID-19 has been conducted in many countries including Indonesia. Several methods have been used, including hybrid methods such as Bayesian Ridge Regression, Support Vector Machine and Time Series

with back propagation. In this research Time Series with Savitzki Golay smoothing is compared with LSTM-NN because the smoothing using Savitzki Golay technique is widely used for prediction and LSTM-NN is known for its performance. It can be concluded that LSTM-NN prediction produce better result than the Savitzky Golay smoothing. The reason is depicted in the results, where the curve of LSTM-NN shows a distinct rise and align with the actual Time Series data. This prediction reveals that the COVID-19 pandemic will continue to ascend in Indonesia.

6. Future Works

In future, the model can also be used with data representing other countries. There are several smoothing techniques available such as moving average, exponential smoothing, Multiplicative Winter's Smoothing Method, Brown's Double Exponential Smoothing and many more. Therefore different smoothing model can be applied to obtain a better fitted curve thus a better prediction.

References

1. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis [Internet]. 2020;2019. Available at: http://dx.doi.org/10.21037/jtd.2020.02.64

2. Harapan H, Itoh N, Yufika A, Winardi W, Keam S, Te H, et al. Coronavirus disease 2019 (COVID-19): A literature review. J Infect Public Health [Internet]. 2020;13(5):667–73. Available at: https://doi.org/10.1016/j.jiph.2020.03.019

3. HDX. No Title [Internet]. 2020 [cited July 30, 2020]. Available at: https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province

4. Setiati S, Azwar MK. COVID-19 and Indonesia. Acta Med Inonesia- Indones J Intern Med. 2020;52(1):84–9.

5. Ivanov D. Predicting the impacts of epidemic outbreaks on global supply chains : A simulation-based analysis on the coronavirus outbreak. Transp Res Part E [Internet]. 2020;136(March):101922. Available at: https://doi.org/10.1016/j.tre.2020.101922

6. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China , Italy and France. Chaos, Solitons Fractals Interdiscip J Nonlinear Sci Nonequilibrium Complex Phenom [Internet]. 2020;134:109761. Available at: https://doi.org/10.1016/j.chaos.2020.109761

7. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One [Internet]. 2020;15(5):1–8. Available at: http://dx.doi.org/10.1371/journal.pone.0231236

8. Nicolaa M, Alsafib Z, Sohrabic C, Kerwand A, Al-Jabird A, Iosifidisc C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. Int J Surg. 2020;78:185–93.

9. Cao W, Fang Z, Hou G, Han M, Xu X, Dong J. The psychological impact of the COVID-19 epidemic on college students in China. Psychiatry Res [Internet]. 2020;287(March):112934. Available at: https://doi.org/10.1016/j.psychres.2020.112934

10. Fang L, Karakiulakis G, Roth M. Correspondence hypertension and increased risk for. Lancet Respir [Internet]. 2020;2600(20):30116. Available at: http://dx.doi.org/10.1016/S2213-2600(20)30116-8

11. Variations of Models and Learning Platforms for Prospective Teachers During the COVID-19 Pandemic Period. Indones J Teach Educ. 2020;1(2):61–70.

12. Mailizar, Almanthari A, Maulina S, Bruce S. Secondary School Mathematics Teachers ' Views on E-learning Implementation Barriers during the COVID-19 Pandemic : The Case of Indonesia. 2020;16(7).

13. Barid M, Wajdi N, Faruq U Al. Education Policy Overcome Coronavirus , A Study of Indonesians. J Educ Technol. 2020;3(2):96–106.

14. Milzam M, Mahardika A, Amalia R. Corona Virus Pandemic Impact on Sales Revenue of Micro Small and Medium Enterprises (MSMEs) in Pekalongan City. J Vocat Stud Appl Res. 2020;2(April):7–10.

15. Muzakki F. THE GLOBAL POLITICAL ECONOMY IMPACT OF COVID-19 AND THE IMPLICATION TO INDONESIA. J Soc Polit Sci. 2020;1(2):76–93.

16. Hanoatubun S. DAMPAK COVID - 19 TERHADAP PEREKONOMIAN INDONESIA. EduPsyCouns J. 2020;2(1):146–53.

17. Wachyuni SS, Kusumaningrum DA. The Effect of COVID-19 Pandemic : How are the Future Tourist Behavior ? J Educ Soc Behav Sci. 2020;33(4):67–76.

18. Yunity IGAD, Sasmita N, Komara LL, Purba JH, Pandawani NP. The Impact of Covid-19 on Community Life in the Province of Bali , Indonesia. Int J Psychosoc Rehabil. 2020;24(10):1918–29.

19. Nuraini N, Khairudin K, Apri M. Modeling Simulation of COVID-19 in Indonesia based on Early Endemic Data. 2020;3(1):1–8.

20. Rustan, Handayani L. THE OUTBREAK ' S MODELING OF CORONAVIRUS (COVID-19) USING THE MODIFIED SEIR MODEL IN INDONESIA. SPEKTRA J Fis dan Apl. 2020;5(1):61–8.

21. Parhusip HA. Menelusuri COVID-19 di Dunia dan di Indonesia dengan Model Regresi SVM , Bayesian dan Gaussian Study on COVID-19 in the World and Indonesia Using Regression Model of SVM , Bayesian Ridge and Gaussian. J Ilm Sains. 2020;20(2):49–

57.

22. Siskus F, Arianto D. PREDIKSI KASUS COVID-19 DI INDONESIA MENGGUNAKAN METODE BACKPROPAGATION DAN FUZZY TSUKAMOTO. J Teknol Inf. 2020;4(1):120–7.

23. Kermack WO, McKendrick A. A Contribution to the Mathematical Theory of Epidemics. In: The Royal Society A. 1927. p. 700–21.

24. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013;19(9):999–1005.

25. Ku Z, Cheng L. The Forecast of Ebola Virus Propagation. In: Dang P, Ku M, Qian T, Rodino L (eds) New Trends in Analysis and Interdisciplinary Applications. Birkhäuser, Cham; 2017. p. 97–103.

26. Jia N, Tsui L. Epidemic Modelling Using SARS As A Case Study. North Am Actuar J. 2005;9(4):28–42.

27. Ozdinc M, Senel K, Ozturcan S, Akgul A. Predicting the Progress of COVID-19 : The Case for Turkey. Turkiye Klin J Med Sci Press. 2020;19–22.

28. Biswas MHA, Paiva LT, Pinho M de. A SEIR MODEL FOR CONTROL OF INFECTIOUS DISEASES. Math Biosci Eng. 2014;11(4):761–84.

29. Wu G. The Ebola virus based on SIR and SEIR model. Int Journall Sci. 2018;5(10):163–8.

30. Vaishya R, Javaid M, Haleem I, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr Clin Res Rev. 2020;14:337–9.

31. Rustam F, Reshi AA, Mehmood A, Ullah S, On WON, Aslam W, et al. Future Forecasting Using Supervised Machine Learning Models. IEEE Access. 2020;8:101489–99.

32. Pinter G, Felde I, Mosavi A, Ghamisi P. COVID-19 Pandemic Prediction for Hungary ; A Hybrid Machine Learning Approach. Mathematics. 2020;8(890).

33. Zheng N, Du S, Wang J, Zhang H, Cui W, Kang Z, et al. Predicting COVID-19 in China Using Hybrid AI Model. IEEE Trans Cybern. 2020;1–14.

34. Ghosal S, Sengupta S, Majumder M, Sinha B. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th. Diabetes Metab Syndr Clin Res Rev [Internet]. 2020;14(4):311–5. Available at: https://doi.org/10.1016/j.dsx.2020.03.017

35. Chen H, Song Q, Tang G, Feng Q, Lin L. The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models. ISRN Spectrosc. 2013;2013.

36. Luchini S, Teschl M, Pintus PA, Degoulet M, Baunez C, Moatti J. Urgently Needed for Policy Guidance : An Operational Tool for Monitoring the COVID-19 Pandemic [Internet]. Social Science Research Network. 2020 [cited July 31, 2020]. p. 1–14. Available at: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3563688