JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 6, No. 1. December 2023 DOI: https://doi.org/10.34288/jri.v6i1.263
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

COMPARISON OF THE APPLICATION OF NEURAL NETWORKS WITH K-FOLD CROSS VALIDATION AND SLIDING WINDOW VALIDATION FOR FORECASTING COVID-19 RECOVERED CASES

Tyas Setiyorini

Informatika
Universitas Nusa Mandiri
Jakarta, Indonesia
tyas.setiyorini@gmail.com

Abstract

The Covid-19 virus first appeared in China resulting in millions of confirmed cases, deaths and recovered cases to date. The spread and increase in the death rate due to Covid-19 is very worrying. Health workers and researchers continue to struggle to improve recovery from Covid-19 cases. There is a need for future forecasting to predict recovery from cases that occur, so that the public or government can understand the spread, take precautions and prepare for action as early as possible. Several previous studies have carried out forecasting the future impact of Covid-19 using Machine Learning methods. K-Fold Cross Validation is usually applied in classification and regression cases, but is not suitable for forecasting time series data. In this study, an RMSE result of 0.990 was obtained in the application of Neural Network with K-Fold Cross Validation, and an RMSE of 0.330 in the application Neural Network with Sliding Window Validation. This proves that the application of Neural Network with Sliding Window Validation is able to improve performance much better than K-Fold Cross Validation in forecasting Covid-19 recovery cases in China. The Sliding Window Validation method has been proven to be suitable for forecasting time series data so that it can be applied in various forecasting cases in the future. One of them, as in other fields, is the case of forecasting electricity consumption.

Keywords: Covid-19, Forecasting, Neural Network; Sliding Window Validation

Abstrak

Virus Covid-19 untuk pertama kalinya muncul di China mengakibatkan jutaan kasus terkonfirmasi, kasus kematian dan kasus sembuh hingga kini. Penyebaran dan kenaikan tingkat kematian akibat Covid-19 sangat memprihatinkan. Para tenaga kesehatan dan peneliti terus berjuang meningkatkan kesembuhan dari kasus Covid-19. Perlunya peramalan masa depan untuk memprediksi kesembuhan dari kasus yang terjadi, agar masyarakat atau pemerintah dapat memahami penyebaran, melakukan pencegahan serta melakukan persiapan tindakan sedini mungkin. Beberapa penelitian sebelumnya telah dilakukan peramalan masa depan dampak Covid-19 dengan menggunakan metode Machine Learning. K-Fold Cross Validation biasa diterapkan pada kasus klasifikasi dan regresi, namun tidak cocok diterapkan pada kasus peramalan data time series.. Pada penelitian ini diperoleh hasil RMSE sebesar 0,990 pada penerapan Neural Network dengan K-Fold Cross Validation, dan RMSE sebesar 0,330 pada penerapan Neural Network dengan Sliding Window Validation. Hal tersebut membuktikan bahwa penerapan Neural Network dengan Sliding Window Validation mampu meningkatkan kinerja yang jauh lebih baik dibanding dengan K-Fold Cross Validation pada peramalan kasus sembuh Covid-19 di China. Metode Sliding Window Validation telah terbukti tepat digunakan untuk peramalan data time series sehingga dapat diterapkan di berbagai kasus peramalan di masa depan. Salah satunya seperti pada bidang lain yaitu kasus peramalan konsumsi listrik.

Kata kunci: Covid-19, Forecasting, Neural Network, Sliding Window Validation

INTRODUCTION

The Covid-19 virus first appeared in Wuhan, China in December 2019. In China, as of March 15 2020, there were 81.058 confirmed cases and 3.204 deaths due to the virus. (Shi et al., 2020). China has implemented a lockdown, inspections in public places and isolation of people who may be suffering from the disease. However, the number of Covid-19 cases continues to increase day by day

(Roosa et al., 2020). Cases of the Covid-19 virus continue to spread and increase throughout the world. On March 17, 2021 the Indonesian government reported 1.437.283 confirmed cases of COVID-19, 38.915 deaths and 1.266.673 recovered cases from 510 districts in 34 provinces (WHO, 2021). The Covid-19 virus claimed millions of lives in a short time. Hospitals and health centers in various countries are flooded with patients, to the point where cemeteries are full (Das, 2020).

The spread and increasing death rate due to the Covid-19 virus raises deep concern. Huge concerns continue to arise among the public and government about how long the outbreak will last or peak, how many people will be infected (Zhang et al., 2020), and how many people will recover. Researchers and health workers continue to work hard to improve recovery from Covid-19 cases.

It is necessary to understand the spread of Covid-19 to predict the number of recovered cases of Covid-19 (Fanelli & Piazza, 2020) to prepare countermeasures as soon as possible (Rath et al., 2020). The new outbreak of Covid-19 presents challenges for modeling researchers, as limited data is available on the initial growth trajectory, and the epidemiological characteristics of the new coronavirus are not yet fully elucidated (Roosa et al., 2020). A number of studies have been carried out regarding forecasting and modeling the COVID-19 problem using machine learning methods (Ribeiro et al., 2020)(Saba & Elsheikh, 2020)(Peng & Nagata, 2020)(Kavadi et al., 2020)(Castillo & Melin, 2020)(Fong et al., 2020). Avoiding and predicting Covid-19 disease as accurately as possible is something that is urgently needed for the public health system (Ribeiro et al., 2020). To ensure such accuracy, Artificial Intelligence models have been widely used over the years to forecast epidemiological time series (Ribeiro et al., 2020).

Time series data is a very important need for future forecasting, especially forecasting the Covid-19 pandemic. Public data sets available from 10 countries in the world in the form of time series data of confirmed cases, recovered cases, and death cases from Covid-19 have been used to build models accurately and efficiently (Castillo & Melin, 2020). A collection of data in a regular time series and each data point has the same distance throughout time is called time series data (Chimmula & Zhang, 2020). A time series is a sequential range of time such as hourly, daily, weekly, monthly or yearly (Dodamani et al., 2015). Artificial Intelligence is developing very quickly, providing important short-term forecasting solutions for time series (Yan et al., 2019).

Developing a forecasting model to predict the spread of Covid-19 is an important issue. In Saba and Elsheikh's research, a statistical and artificial intelligence-based approach has been proposed to model and predict the spread of Covid-19 in Egypt using the Arima method and artificial neural networks (Saba & Elsheikh, 2020). Neural Networks are able to overcome limitations in traditional time series forecasting techniques by adapting the nonlinearity of specific Covid-19 datasets and can produce sophisticated results on temporal data (Chimmula & Zhang, 2020). Neural networks have become popular in all areas of engineering, including load forecasting, and overcoming nonlinearities and functional dependencies in forecasting models (Ferreira & Alves da Silva, 2007). Various types of artificial neural networks have been used to model complex and nonlinear relationships between features used for forecasting and are able to achieve high accuracy (Agrawal et al., 2019).

Apart from the Neural Network method which is suitable for forecasting, the Sliding Window method has also been used for forecasting time series data (Norwawi, 2021). This is very important because Sliding Window is able to restructure lost data in a short period of time, thereby disrupting transmission performance and Machine Learning-based predictive algorithms (Papadopoulos et al., 2023). The Sliding Window method is more suitable for application to time series data than the K-Fold Cross Validation method. K-Fold Cross Validation is one of the most important tools in evaluating regression and classification methods, but when applied to the case of time series forecasting it causes problems (Bergmeir & Benítez, 2012). The classic K-Fold Cross Validation method is applied to data that is independent and identically distributed, so it is not applicable to long time series data (Monnier, 2018). Therefore special care is required for computational machine learning models for time series data (Monnier, 2018). To achieve high accuracy in prediction, a hybrid prediction model combining a convolutional neural network with a sliding time window is carried out (Zhen et al., 2022). Sliding Window has been proven to improve forecasting performance (Papadopoulos et al., 2023).

The Neural Network and Sliding Window methods are appropriate methods to use for forecasting time series data and are proven to be able to achieve high accuracy. Sliding Window has been stated as a more suitable method for forecasting data series compared to K-Fold Cross Validation. Therefore, this research will compare

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i1.263
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

which performance is better in applying Neural Network with K-Fold Cross Validation and Neural Network with Sliding Window Validation for predicting recovery rates in Covid-19 sufferers. Apart from that, to prove whether Sliding Window Validation has better performance than K-Fold Cross Validation in forecasting time series data.

MATERIALS AND METHODS

Data

The dataset used in this research was obtained from public data on kaggle.com. This dataset is time series data on the number of recovered Covid-19 cases in China from 22 January 2020 to 10 May 2020. The Time Series Data on Recovered Covid-19 Cases in China shown in Table 1 consists of predictor attribute, namely date, as well as a class attribute, namely number of recoveries per date.

Table 1. The Time Series Data on Recovered Covid-19 Cases in China

No	Attribute	Description
1	Date	Date
2	Recoveries	Number of recovered per date

Methodology

Figure 1 shows the methods used in this research, two experiments were carried out, namely applying Neural Network with K-Fold Cross Validation and Neural Network with Sliding Window Validation for the Covid-19 Time Series Dataset in China to predict the number of recovered cases over time in a certain period. After carrying out the training and testing process with Neural Network with K-Fold Cross Validation and Neural Network with Sliding Window Validation , the final result is Root Mean Square Error (RMSE). From the RMSE results, they are then compared, which RMSE result is the smallest. The smallest RMSE results show the best results.

Figure 2 and Figure 3 illustrate the differences in how K-Fold Cross Validation and Sliding Window Validation work. Figure 2 shows K-Fold Cross Validation working by dividing the dataset into k folds of equal size. The model is trained using k-1 folds as training data and testing data on the remaining folds. The training and testing process is repeated k times, with each fold used as test data exactly once.

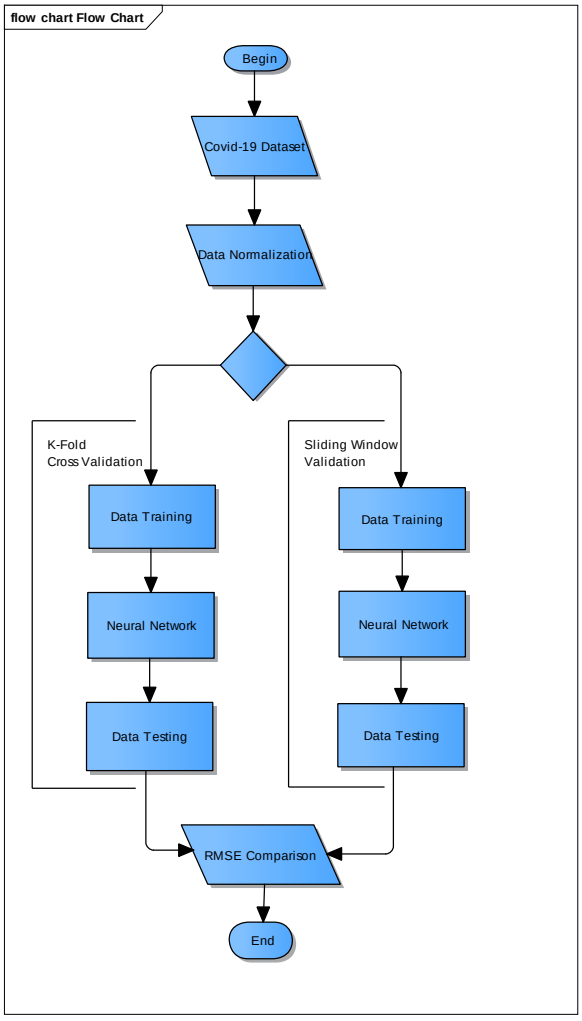

Figure 1. Comparison of the Application of Neural Networks with K-Fold Cross Validation and Neural Networks with Sliding Window Validation

Figure 2 shows Sliding Window Validation working by dividing the dataset into two parts, namely one for initial training and the other for testing. The training window shifts across the data, and the model is iteratively updated over time. In K-fold Cross-Validation, dataset splitting is done by dividing the dataset into folds, while Sliding Window Validation focuses on splitting data based on time windows for testing and training respectively. The revised version of the K-Fold Cross Validation standard including Sliding Window Validation is for handling time series data that is evaluated based on future observations, not random observations. This revised version is known as time series splitting on which training is performed the set is split at each iteration such that the validation set is always at the front opening of training (Mustafa Qamar-ud-Din, 2019).

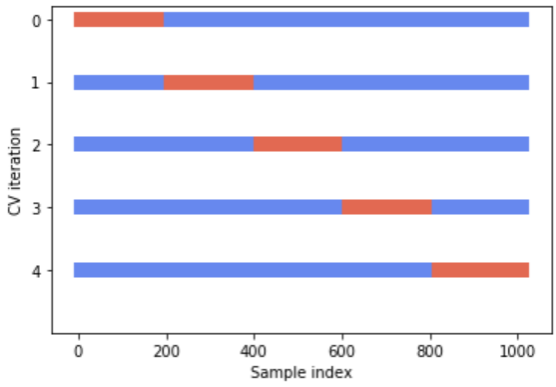

Figure 2. K-Fold Cross Validation
Source: (Mustafa Qamar-ud-Din, 2019)

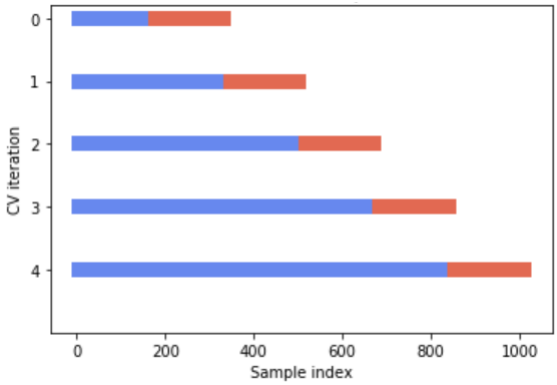

Figure 3. Sliding Window Validation
Source: (Mustafa Qamar-ud-Din, 2019)

Neural Network

Artificial neural networks or Neural Networks are developing very quickly and are able to predict in various fields (Yan et al., 2019). Neural Networks are able to extract complex and nonlinear relationships well through a training process using historical data for forecasting cases (Lee & Ko, 2009). Neural Networks apply partial computational models in processing information to determine the relationship between a set of patterns or variables in data. The Neural Network method in learning imitates biological neural networks, especially those in the human brain. The nonlinear and nonparametric nature of neural networks is more wired to model complex data problems in data mining (Brockmann et al., 2006). Neural Network is a model that imitates the function of the human brain. The human brain consists of millions of neurons which are small processing units that work in parallel. Neurons are connected to each other through neuronal connections. Each individual neuron takes input from a set of neurons. Then the input is processed through the output to a set of neurons. The output is collected by other neurons for further processing (Shukla, 2010). The Neural Network algorithm works through an iterative process using training data, comparing the predicted values from the network with each data contained in the training data (Han et al., 2012).

Sliding Window

Sliding window is one of the methods applied at the preprocessing stage to restructure data into a classification problem according to time period (Norwawi, 2021). Sliding Window is applied to extract data samples to achieve dynamic data acquisition. Convolutional Neural Networks train on samples to obtain the final combined prediction model (Zhen et al., 2022).

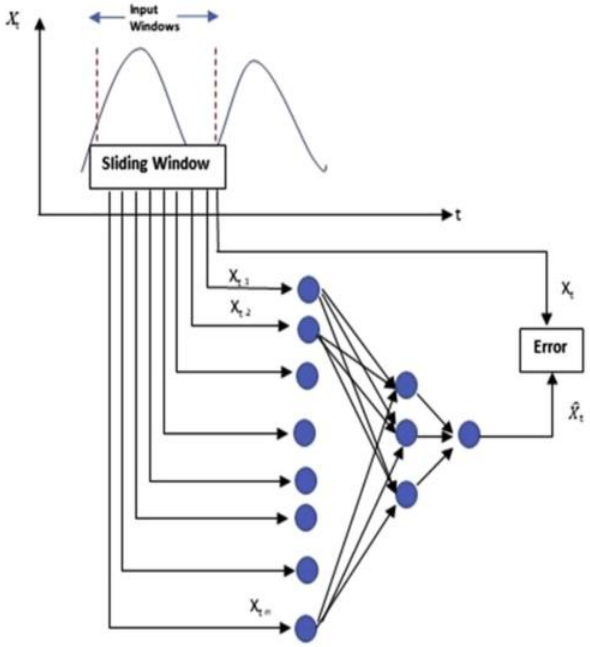

Figure 4. Application of Sliding Window in Neural Network

The key to success in building a prediction model is in selecting training data samples. When the selected training data samples are unable to reflect the effective relationship between the samples and the predicted values, this will hinder the learning process of network parameters, and reduce the efficiency and accuracy of predictions. Therefore, Sliding Window is implemented to solve this problem (Zhen et al., 2022).

Figure 4 shows the application of Sliding Window in Neural Networks. Figure 4 shows the application of Sliding Window in Neural Networks. Given a time series data set that can be described by y=f(x), using a Sliding Window, the data will be restructured where the previous output or historical data will be used to predict the next output. For example, for a window width of 1, the output of the current time t is based on the output of the previous time, t-1 as shown in Eq. :

$y_t=f(y_{t-1})$ …….. (1)

RESULTS AND DISCUSSION

Two experiments were carried out, namely applying a Neural Network with K-Fold Cross Validation and Neural Network with Sliding Window Validation on the Time Series Dataset of Covid-19 Recovered Cases in China. Then the final result, namely the RMSE of the two methods, will be compared.

Experiments were carried out using default parameters shown in Table 2. Parameters in Neural Network consisting of training cycles=200, learning rate=0.01, momentum=0.9 and hidden layer 1=2. Parameters in K-Fold Cross Validation consisting of Folds=10. Parameters in Sliding Window Validation consisting of training window size=20, step size=1, and window size=1.

Table 2. Parameters used in Neural Network, K-Fold Cross Validation and Sliding Window Validation

Parameter		
Neural Network	K-Fold Cross Validation	Sliding Window Validation
Training Cycles=200	Folds=10	Training window size=20
Learning Rate=0.01		Step Size=1
Momentum=0.9		Window Size=1
Hidden layer 1=2		

Table 3 shows the Comparison of RMSE results of applying a Neural Network with K-Fold Cross Validation is 0.990, and applying a Neural Network with Sliding Window Validation is 0.330. The comparison results in Table 3 show significant differences in the application of Neural Networks with K-Fold Cross Validation and Neural Network with Sliding Window Validation. It is clear that the RMSE value is much smaller when using the Sliding Window Validation method compared to using K-Fold Cross Validation. The smaller the RMSE value means the better the performance. This proves that the performance of Sliding Window Validation is better than K-Fold Cross Validation, and it is proven that the Sliding Window method is suitable for forecasting time series data. As mentioned by Papadopoulos, Sliding Window shows a significant increase in accuracy and can improve performance. Implementation of a training data approach using a Sliding Window has been proven to result in significantly higher accuracy compared to offline implementation fore forecasting time series data (Papadopoulos et al., 2023). The Sliding Window is a convenient method for conducting time series forecasting (Norwawi, 2021).

Table 3. Comparison of RMSE Results in the Application of Neural Network with K-Fold Cross Validation and Neural Network with Sliding Window Validation

No	Method	RMSE
1	Neural Network with K-Fold Cross Validation	0.990
2	Neural Network with Sliding Window Validation	0.330

CONCLUSIONS

An experiment has been carried out to apply a Neural Network with K-Fold Cross Validation and a Neural Network with Sliding Window Validation on the Time Series Dataset of Covid-19 Recovered Cases in China. Obtained RMSE of 0.990 in the application of Neural Network with K-Fold Cross Validation, and RMSE of 0.330 in application of Neural Network using Sliding Window Validation. The application of a Neural Network using Sliding Window Validation shows a much smaller RMSE value compared to K-Fold Cross Validation. This shows that Sliding Window Validation has much better performance than K-Fold Cross Validation. This proves that the Sliding Window method is more suitable to be applied in the case of time series data forecasting than K-Fold Cross Validation.

REFERENCES

Agrawal, R. K., Muchahary, F., & Tripathi, M. M. (2019). Ensemble of relevance vector

machines and boosted trees for electricity price forecasting. *Applied Energy*, *250*(May), 540–548. https://doi.org/10.1016/j.apenergy.2019.05.062

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). Data Mining and Knowledge Discovery Handbook. In *Springer*. https://doi.org/10.1038/nature04292

Castillo, O., & Melin, P. (2020). Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic. *Chaos, Solitons and Fractals*, *140*, 110242. https://doi.org/10.1016/j.chaos.2020.110242

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons and Fractals*, *135*. https://doi.org/10.1016/j.chaos.2020.109864

Das, R. C. (2020). Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries. *Chaos, Solitons and Fractals*, *140*, 110248. https://doi.org/10.1016/j.chaos.2020.110248

Dodamani, S. N., Shetty, V. J., & Magadum, R. B. (2015). Short term load forecast based on time series analysis: A case study. *Proceedings of IEEE International Conference on Technological Advancements in Power and Energy, TAP Energy 2015*, 299–303. https://doi.org/10.1109/TAPENERGY.2015.7229635

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals*, *134*, 109761. https://doi.org/10.1016/j.chaos.2020.109761

Ferreira, V. H., & Alves da Silva, A. P. (2007). Toward estimating autonomous neural network-based electric load forecasters. *IEEE Transactions on Power Systems*, *22*(4), 1554–1562. https://doi.org/10.1109/TPWRS.2007.908438

Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied Soft Computing Journal*, *93*(December 2019), 106282. https://doi.org/10.1016/j.asoc.2020.106282

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques. In *Data Mining*. https://doi.org/10.1016/b978-0-12-381479-1.00001-0

Kavadi, D. P., Patan, R., Ramachandran, M., & Gandomi, A. H. (2020). Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19. *Chaos, Solitons and Fractals*, *139*. https://doi.org/10.1016/j.chaos.2020.110056

Lee, C. M., & Ko, C. N. (2009). Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm. *Neurocomputing*, *73*(1–3), 449–460. https://doi.org/10.1016/j.neucom.2009.07.005

Monnier, S. (2018). *Cross-validation tools for time series*. Medium.Com. https://medium.com/@samuel.monnier/cross-validation-tools-for-time-series-ffa1a5a09bf9

Mustafa Qamar-ud-Din. (2019). *Cross-Validation strategies for Time Series forecasting [Tutorial]*. Packt Editorial Staff. https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/

Norwawi, N. (2021). forecasting with multilayer. In *Data Science for COVID-19 Volume 1*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-824536-1.00025-3

Papadopoulos, D. N., Dadras, F., Najafi, B., Haghighat, A., & Rinaldi, F. (2023). Energy & Buildings Handling complete short-term data logging failure in smart buildings : Machine learning based forecasting pipelines with sliding-window training scheme. *Energy & Buildings*, *301*(October), 113694. https://doi.org/10.1016/j.enbuild.2023.113694

Peng, Y., & Nagata, M. H. (2020). An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos, Solitons and Fractals*, *139*. https://doi.org/10.1016/j.chaos.2020.110055

Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus

disease (COVID-19) pandemic using multiple linear regression model. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(5), 1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & Coelho, L. dos S. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons and Fractals*, *135*. https://doi.org/10.1016/j.chaos.2020.109853

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P., & Chowell, G. (2020). Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infectious Disease Modelling*, *5*, 256–263. https://doi.org/10.1016/j.idm.2020.02.002

Saba, A. I., & Elsheikh, A. H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Safety and Environmental Protection*, *141*, 1–8. https://doi.org/10.1016/j.psep.2020.05.029

Shi, Y., Wang, J., Yang, Y., Wang, Z., Wang, G., Hashimoto, K., Zhang, K., & Liu, H. (2020). Brain , Behavior , & Immunity - Health Knowledge and attitudes of medical staff in Chinese psychiatric hospitals regarding COVID-19. *Brain, Behavior, & Immunity - Health*, *4*(March), 100064. https://doi.org/10.1016/j.bbih.2020.100064

Shukla, A. (2010). Real Life Applications of Soft Computing. In *Real Life Applications of Soft Computing*. https://doi.org/10.1201/ebk1439822876

WHO. (2021). *Coronavirus Disease 2019 (COVID-19) Coronavirus Coronavirus Disease Disease Situation World Health World Health Organization Organization 28 April 2021*. Covid 19. https://cdn.who.int/media/docs/default-source/searo/indonesia/covid19/external-situation-report-46_10-march-2021-update.pdf?sfvrsn=1859ffc2_5

Yan, K., Li, W., Ji, Z., Qi, M., & Du, Y. (2019). A Hybrid LSTM Neural Network for Energy Consumption Forecasting of Individual Households. *IEEE Access*, *7*, 157633–157642. https://doi.org/10.1109/ACCESS.2019.2949065

Zhang, X., Ma, R., & Wang, L. (2020). Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. *Chaos, Solitons and Fractals*, *135*. https://doi.org/10.1016/j.chaos.2020.109829

Zhen, L., Zhang, L., Yang, T., Zhang, G., Li, Q., & Ouyang, H. (2022). Simultaneous prediction for multiple source–loads based sliding time window and convolutional neural network. *Energy Reports*, *8*, 6110–6125. https://doi.org/10.1016/j.egyr.2022.04.041

P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v6i1.263 Vol. 6, No. 1. December 2023
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023