**JURNAL RISET INFORMATIKA** **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 4, No. 3. June 2022 DOI: https://doi.org/10.34288/jri.v4i3. 409
**Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020**

# COMPARISON OF LINEAR REGRESSIONS AND NEURAL NETWORKS FOR FORECASTING COVID-19 RECOVERED CASES

**Tyas Setiyorini[1*)], Frieyadie[2]**

Teknik Informatika, Sistem Informatika
Universitas Nusa Mandiri
www.nusamandiri.ac.id
tyas.setiyorini@gmail.com[1*)], frieyadie@nusamandiri.ac.id[2]

(*) Corresponding Author

**Abstract**

The emergence of the Covid-19 outbreak for the first time in China killed thousands to millions of people. Since the beginning of its emergence, the number of cases of Covid-19 has continued to increase until now. The increase in Covid-19 cases has a very bad impact on health and social and economic life. The need for future forecasting to predict the number of deaths and recoveries from cases that occur so that the government and the public can understand the spread, prevent and plan actions as early as possible. Several previous studies have forecast the future impact of Covid-19 using the Machine Learning method. Time series forecasting uses traditional methods with Linear Regression or Artificial Intelligence methods with neural networks. The research proves a linear relationship in the time series data of Covid-19 recovered cases in China, so it is proven that Linear Regression performance is better than the Neural Network.

Keywords: Covid-19, Forecasting, Linear Regression, Neural Network

***Abstrak***

*Munculnya wabah Covid-19 untuk pertama kalinya di China membunuh ribuan hingga jutaan orang. Dari awal kemunculan jumlah kasus Covid-19 terus meningkat hingga saat ini. Peningkatan kasus Covid-19 sangat berdampak buruk bagi kehidupan kesehatan, sosial maupun ekonomi. Perlunya peramalan masa depan untuk memprediksi jumlah kematian dan kesembuhan dari kasus yang terjadi, agar pemerintah maupun masyarakat dapat memahami penyebaran, mencegah serta merencanakan tindakan sedini mungkin. Beberapa penelitian sebelumnya telah melakukan peramalan masa depan dampak Covid-19 dengan menggunakan metode Machine Learning. Peramalan time series dapat dilakukan dengan metode tradisional dengan Linear Regression atau metode Artificial Intelligent dengan neural network. Pada penelitian ini telah dibuktikan bahwa terdapat hubungan linear pada data time series kasus sembuh Covid-19 di China, sehingga terbukti bahwa kinerja Linear Regression lebih baik dibanding Neural Network.*

*Kata Kunci: Covid-19, Forecasting, Linear Regression, Neural Network*

## INTRODUCTION

In December 2019, the beginning that took the world by storm emerged a new corona outbreak, Covid-19, for the first time in China. In China, as of March 7th, 2020, a total of 80,813 cases have been confirmed, with 3,073 deaths (Livingston et al., 2020). On June 3rd, 2022, 3,184,961 cases were confirmed, with 17,127 deaths. Indonesia, on March 17th, 2021, a total of 1,437,283 patients were confirmed with 38,915 deaths (WHO, 2021), until now, on June 3rd, 2021, a total of 6,056,017 cases have been established with 156,604 deaths (Guan et al., 2020). Covid-19 has attacked and killed thousands to millions of people worldwide, and the number of Covid-19 cases has continued to increase from its first emergence to the present.

The increase in Covid-19 cases and deaths has significantly impacted changes in world life in terms of health, social and economic. It raises huge concerns for the government and society, such as when the Covid-19 outbreak will peak, how long the

spell will last, how many people will eventually be infected (Zhang et al., 2020), and how many people can survive and heal. Based on these things, there is a need for future forecasting to predict the number of Covid-19 cases so that the government and the public can understand the spread of Covid-19 (Fanelli & Piazza, 2020)(Fong et al., 2020)(Roosa et al., 2020), preparing for prevention as early as possible, as well as preparation of action planning (Rath et al., 2020).

Forecasting the future of Covid-19 also aims to develop effective public health system planning. The accuracy of disease forecasting impacts the public health system (Ribeiro et al., 2020). it is affecting all areas of life. In recent years, many studies have emerged to predict the transmission of Covid-19 by applying several mathematical models (Shim et al., 2020)(Zhao et al., 2020). Several studies analyze the impact of Covid-19 by predicting future cases using machine learning methods (Castillo & Melin, 2020; Fong et al., 2020; Kavadi et al., 2020; Peng & Nagata, 2020)

A time series is a series of sequential data measured over time, such as hourly, daily, or weekly peak loads (Dodamani et al., 2015). Covid-19 is one type of time series data recommended for applying a sequential network to extract patterns. However, these data are dynamic, so they are often unclear when using epidemiological and statistical models (Krätschmer, 2006). Regression models are included in traditional time series methods (Yan et al., 2019), such as linear regressions unsuitable for predicting nonlinear and complex models (Satre-Meloy, 2019). Linear regression models present little focus, just as most anticipated qualities are lower, especially for chillers, indicating low linearity between due utilization and the original (Pombeiro et al., 2017). This complexity makes understanding the connection between data input and reactions challenging. In network modeling, some of the weaknesses of the existing model are non-temporal, linear, and several other possible constraints (Chimmula & Zhang, 2020). In general, for short-term load forecasting, the use of traditional methods such as statistical models is a linear regression method that is a linear model, which suffers from nonlinearity and provides only reasonable accuracy (Lee & Ko, 2009).

Compared to traditional methods, Artificial Intelligence (AI) is proliferating, providing short-term forecasting solutions essential for time series (Yan et al., 2019). Modern artificial intelligence methods are developed very fast, introducing artificial neural networks and population evolution algorithms for forecasting electricity (Taylor, 2010). Neural networks have become popular in terms of nonlinear in all areas of engineering, including load forecasting, and overcome functional dependence on forecasting models (Ferreira & Alves da Silva, 2007). Various neural variants of network artificialization are implemented to model complex and nonlinear relationships between features used for forecasting and achieve high accuracy (Agrawal et al., 2019). Nonlinear relationships of models in the complex structure of electricity demand allow overcome by ANN (Nadtoka & Al-Zihery Balasim, 2015).

Traditional methods can forecast time series with Linear Regression or Artificial Intelligent methods with neural networks. In this study, we want to prove whether there is a linear or nonlinear relationship in the Covid-19 time series data. After that, we will compare which performance is better using Linear Regression or Neural Network.

## MATERIALS AND METHODS

### Data

This study uses time series data on the incidence or number of COVID case recoveries in China from January 22nd, 2020, to May 10th, 2020. This dataset is obtained from kaggle.com.

The Time Series Recoveries Cases Covid-19 Dataset in China shown in Table 1 consists of 1 attribute predictor, date, and one attribute class, recoveries.

Table 1. Time Series Recoveries Cases Covid-19 Dataset in China

| No | Attributes | Description |
|---|---|---|
| 1 | Date | Date |
| 2 | Recoveries | Number of recovered per date |

Table 2. Significant Value of Linearity Test on the Time Series Recoveries Cases Covid-19 Dataset in China

| Model | Sig. |
|---|---|
| Date | 1.000 |
| Recoveries | 1.000 |

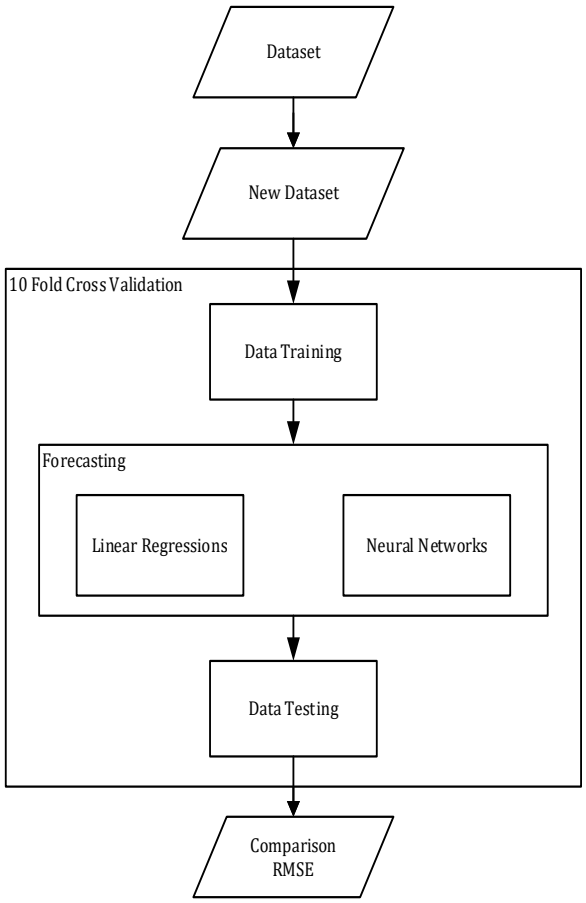
Source: (Setiyorini & Frieyadie, 2022)

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3. 409
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

After the linearity test shows that in Table 2, the Time Series Recoveries Cases Covid-19 Dataset in China A shows a significant value (sig.) on the Recoveries is 0.000. It shows that 1,000 >0.05, so the data has a linear relationship. It can be concluded that the Time Series Recoveries Cases Covid-19 Dataset in China dataset show data that have linear relationships.

## Methodology

Figure 1 illustrates the method used in this study which compares the implementation of the linear regressions and neural networks. The initial step is training and testing with 10-fold cross-validation on the electricity consumption dataset using linear regression and neural networks (Kaytez et al., 2015) to produce RMSE. The RMSE produced by linear regressions and neural networks is then compared to the smallest RMSE.




Source: (Setiyorini & Frieyadie, 2022)

Figure 1. Comparison of Linear Regressions and Neural Networks

## Linear Regression

The regression analysis model is the most well-known model for forecasting electricity consumption (Abdel-Aal & Al-Garni, 1997). Linear regression is included in the statistical analysis method, which is applied to characterize the impact of selected independent (predictors) variables on the dependent (response) variable (Fang & Lahdelma, 2016). Linear regression is used for numerical data analysis and modeling (Han et al., 2012). Linear regression still can't be related to nonlinear problems, so it must be studied to find out whether it can be applied to short-term predictions (Shao et al., 2020).

Multiple Linear Regression (MLR) is the generalization of the simple linear regression technique (Aiken et al., 2013)(Fumo & Rafe Biswas, 2015). MLR is an algorithm that describes the relationship between one dependent variable and several independent variables (Shao et al., 2020).

The model in multiple linear regression consists of more than one predictor variable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon \quad \text{.....(1)}$$

Where Y is the response variable, $X_1$; $X_2$; … $X_p$ is the predictor variables with p as the number of variables, $\beta_0$; $\beta_1$; …$\beta_p$ are the regression coefficients, and ε is an error to account for the discrepancy between predicted data and the observed data (Fumo & Rafe Biswas, 2015).

The linear regression model characterizes the behavior of the unknown quantity y in terms of known quantities x, parameters, and random noise $\varepsilon$ (Fang & Lahdelma, 2016). Linear regression forecasting models are expressed in the following format:

$$Y_t = \beta X_t + \varepsilon_t \quad \text{..... (2)}$$

where $Y_t$ $i$s the predicted value at time t; $X_t$= (1, $X_{1t}$, $X_{2t}$, $X_{3t}$, ..... $X_{nt}$) is a vector of k explanatory variables at time t, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)^T$ is the vector of coefficients, and $\varepsilon_t$ is a random error term at time t, t = 1, …, N (Fang & Lahdelma, 2016).

## Neural Network

Artificial intelligence's rapid development and neural networks forecast various fields. (Yan et al., 2019). In health and care technology, mobile computing and artificial intelligence are one of the

keys to success (George & Huerta, 2018). The neural network is the deep learning method developed at this time to forecast energy consumption with very high predictive accuracy (Choi et al., 2018).

Neural networks are partial computational models for information processing beneficial for figuring out essential relationships among a set of patterns or variables in data. They are intelligent where the learning methods mimic the biological neural networks, especially those in the human brain. The nonlinear and nonparametric nature of neural networks is more of a cable for modeling complex data problems in data mining (Brockmann et al., 2006).

## RESULTS AND DISCUSSION

The Time Series Recoveries Cases Covid-19 Dataset in China experimented with linear regressions and neural networks. Then the RMSE results are compared on the linear regression and neural network. Table 3 shows the experimental results in the Time Series Recoveries Cases Covid-19 Dataset in China obtained by RMSE of 0975 used the linear regression, and RMSE of 0.990 used neural network.

Table 3. Comparison of RMSE Results with Linear Regression and Neural Network on the Time Series Recoveries Cases Covid-19 Dataset in China

| No | Method | RMSE |
|---|---|---|
| 1 | Linear Regression | 0.975 |
| 2 | Neural Network | 0.990 |

Source: (Setiyorini & Frieyadie, 2022)

The comparison results in Table 3 show significant differences in the value of RMSE between the use of linear regressions and neural networks. It also indicates decreased RMSE values in linear regressions and neural networks. The use of neural networks indicates a smaller RMSE value compared to the use of linear regressions. It shows that neural networks have better performance than linear regressions.

As explained earlier, the electricity consumption dataset A and electricity consumption dataset B show nonlinear data relationships. Referring to previous research (Satre-Meloy, 2019), linear regression was unsuitable for nonlinear models, while neural networks were implemented to model nonlinear relationships to achieve high accuracy (Agrawal et al., 2019). This study proves that neural networks can overcome nonlinear problems in the electricity consumption dataset A and the electricity consumption dataset B so that the linear regressions can improve performance better than neural networks.

## CONCLUSION

Experiments conducted with linear regressions and neural networks on the Time Series Recoveries Cases Covid-19 Dataset in China obtained by RMSE of 0.975 used the linear regression and an RMSE of 0.990 used the neural network. The use of neural networks shows a smaller RMSE value compared to the use of linear regressions. It shows that the problem proven in the Time Series Recoveries Cases Covid-19 Dataset in China is that it has a linear relationship. It can be overcome by linear regression so that the linear regression improves performance better than the neural network.

## REFERENCE

Abdel-Aal, R. E., & Al-Garni, A. Z. (1997). Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis. *Energy*, *22*(11), 1059–1069. https://doi.org/10.1016/S0360-5442(97)00032-7

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

Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2013). Multiple Linear Regression. In *Handbook of Psychology*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0219

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.2020.11024

2

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

Choi, H., Ryu, S., & Kim, H. (2018). Short-Term Load Forecasting based on ResNet and LSTM. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2018*. https://doi.org/10.1109/SmartGridComm.2018.8587554

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

Fang, T., & Lahdelma, R. (2016). Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. *Applied Energy*, *179*, 544–552. https://doi.org/10.1016/j.apenergy.2016.06.133

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

Fumo, N., & Rafe Biswas, M. A. (2015). Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, *47*, 332–343. https://doi.org/10.1016/j.rser.2015.03.035

George, D., & Huerta, E. A. (2018). Deep Learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*, *778*, 64–70. https://doi.org/10.1016/j.physletb.2017.12.053

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/nejmoa2002032

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

Kaytez, F., Taplamacioglu, M. C., Cam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power and Energy Systems*, *67*, 431–438. https://doi.org/10.1016/j.ijepes.2014.12.036

Krätschmer, V. (2006). Strong consistency of least-squares estimation in linear regression models with vague concepts. *Journal of Multivariate Analysis*, *97*(3), 633–654. https://doi.org/10.1016/j.jmva.2005.04.009

Livingston, E., Bucher, K., & Rekito, A. (2020). Coronavirus Disease 2019 and Influenza 2019-2020. In *JAMA - Journal of the American Medical Association* (Vol. 323, Issue 12, p. 1122). https://doi.org/10.1001/jama.2020.2633

Nadtoka, I. I., & Al-Zihery Balasim, M. (2015). Mathematical modelling and short-term forecasting of electricity consumption of the power system, with due account of air temperature and natural illumination, based on support vector machine and particle swarm. *Procedia Engineering*, *129*, 657–663. https://doi.org/10.1016/j.proeng.2015.12.0

87

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

Pombeiro, H., Santos, R., Carreira, P., Silva, C., & Sousa, J. M. C. (2017). Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. fuzzy modeling vs. neural networks. *Energy and Buildings*, *146*, 141–151. https://doi.org/10.1016/j.enbuild.2017.04.032

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*, 1–10. 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

Satre-Meloy, A. (2019). Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models. *Energy*, *174*, 148–168. https://doi.org/10.1016/j.energy.2019.01.157

Setiyorini, T., & Frieyadie, F. (2022). *Laporan Akhir Penelitian Mandiri:*

Shao, M., Wang, X., Bu, Z., Chen, X., & Wang, Y. (2020). Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society*, *57*(March), 102128. https://doi.org/10.1016/j.scs.2020.102128

Shim, E., Tariq, A., Choi, W., Lee, Y., & Chowell, G. (2020). Transmission potential and severity of COVID-19 in South Korea. *International Journal of Infectious Diseases*, *93*, 339–344. https://doi.org/10.1016/j.ijid.2020.03.031

WHO. (2021). *Coronavirus Disease 2019 (COVID 19): Situation Report - 53*. Covid 19. https://cdn.who.int/media/docs/default-source/searo/indonesia/covid19/external-situation-report-53_28-april-2021.pdf?sfvrsn=c2563ad9_9

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

Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., Lou, Y., Gao, D., Yang, L., He, D., & Wang, M. H. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, *92*, 214–217. https://doi.org/10.1016/j.ijid.2020.01.050