Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 2, 2024, pp. 282 – 292

Comparison of Shallot Price Prediction In Pati City With LSTM, GRU and Linear Regression

Fajar Husain Asy'ari*[1], Ellen Proborini[2]
[1,2]*Sekolah Tinggi Teknik Pati, Jalan Pati-Trangkil Km.4 Pati, (0295) 382470, Pati*
E-mail : fajarhusain@sttp.ac.id[1], ellena@sttp.ac.id[2]*

Melina Dwi Safitri[3], Eko Hari Rachmawanto[4]
[3]*Sekolah Tinggi Teknik Pati,* [4]*Universitas Dian Nuswantoro*
E-mail : safitrimelinadwi@gmail.com[3], eko.hari@dsn.dinus.ac.id[4]

Abstract - Shallots are superior vegetable plant and contribute quite significantly to the development of the national economy. The price of shallots fluctuates almost every year. At certain times the price of shallots soars due to high demand while the supply in the market is insufficient. Therefore, an analysis is needed to see what phenomena significantly affect the increase in the price of shallots. The methods used in the study were LSTM, GRU and LR. The results of the analysis show that the LSTM algorithm gets a MAE value of 0.011072172783, MAPE 3.93678% and RMSE 0.03139695060, this error is the lowest compared to GRU getting MAE value is 0.01185741, MAPE 4.2282357% and RMSE 0.03122299395 and LR with MAE 0.0134737280395416, MAPE 5.45081% and RMSE is 0.0313332635305961, so LSTM is a suitable algorithm for predicting shallot data in Pati district.

Keywords – Shallot, LSTM, LR, GRU

1. INTRODUCTION

Agriculture plays a crucial role in improving the economic welfare of the community [1]. Within the realm of agriculture, there is a horticulture sub sector that has a significant role in the sector [2]. Horticulture is a part of agriculture that involves planting and maintaining fruits, vegetables, and ornamental plants [3]. According to Wahyudi, one of the horticultural crops commonly grown by farmers is shallots [4]. In 2019-2023, the Central Bureau of Statistics stated that the amount of shallot consumption has always increased. Based on Data BPS RI Shallot consumption in 2022 an increase of 8.33% or 60.81 thousand tons compared to 2020, the productivity also increases 10,42% from 2020 [5]. According to data from the National Strategic Food Price Information Center (PIHPS), the price of shallots experienced a significant increase from the beginning of 2022 to March 2022, reaching IDR 36,650/kg.

The price increase reached a peak in July 2022, reaching Rp61,950 per kilogram, and fluctuated until the end of the year to Rp38,300 per kilogram. The price increase in July 2022 was related to the La Nina phenomenon that occurred in May, which caused an increase in rainfall. This resulted in high rainfall intensity in May 2022, which could adversely affect the big harvest that fell in July 2022. This significant price increase triggered a rise in inflation in July 2022, reaching 0.64% (month-on-month), with shallots being the main contributor to inflation at 0.09% on an annual basis, reaching 4.94% (year-on-year), the highest rate since October 2015 [6]. One of the factors that can cause the increase in shallot prices is the involvement of farmers who influence the market, production costs that include various aspects such as seeds, seeds, fertilizers, water, and the distance of the garden to the collection point [7]. The Food Security Agency's Food Distribution and Reserve Center (PDCP) notes that certain phenomena, especially during religious holidays, can affect price patterns due to increased demand, in an effort to maintain supply and price stability, it is important to understand the basic characteristics of price movements and take appropriate action.

Creating price stability is very important as it relates to people's ability to meet food needs in the household. In an effort to avoid the adverse effects of such price fluctuations that can cause an increase in inflation and affect people's purchasing power in meeting their needs, shallot price forecasting in Indonesia is needed. The results of this forecasting can be the basis for designing appropriate strategies and policies related to the shallot price problem. In addition, this forecasting can also provide benefits to consumers by providing timely information so that they can make better decisions, whether to buy or refrain. The ability of consumers to make shallot purchase decisions is also one aspect of food security [8] [9]. Forecasting is an attempt to predict a future event, while the process results from using information from the past and present to estimate events that will occur in the future [10]. Time series data refers to information that is organized based on a time sequence of observations [11]. Time series data collection in the agricultural sector is used to record annual harvests, plant quality based on weather and record plant prices during a certain period in one year [12]. The price of shallots is an example of such time series data.

Based on the explanation above, the price of shallots has an unstable fluctuation value, so it is necessary to forecast the price of shallots with the best algorithm model. The shallot price forecasting process uses several algorithms, namely Linear Regression, LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit), of the three algorithms, the algorithm that has the lowest error value is sought.

According to research conducted by Trisya, et.al (2024), a comparison of ARIMA, LSTM and Support Vector Machines (SVM) models was carried out for Electricity Energy Consumption Analysis, from this comparison, it was found that the LSTM method had the best prediction accuracy compared to other models [13]. Then in further research conducted by Shahi et al. (2020), a comparison of the Long Short-Term Memory (LSTM) model with the Gated Recurrent Unit (GRU) model was carried out in forecasting stock prices, of the two methods, GRU provides better accuracy results compared to the LSTM model [14]. In the study entitled "Perbandingan Performa Algoritma Linear Regresi dan Random Forest untuk Prediksi Harga Bawang Merah di Kota Samarinda" from Pratama et.al (2024) Linear Regression has a lower error rate than Random Forest, the RMSE value obtained by LR is 53.74842694081432, LR is superior because the random forest has a high level of complexity so it only works on certain datasets [15].

In this research, researchers want to know how the performance of the LSTM, GRU and LR algorithms, in predicting the price of shallots. This study is expected to provide benefits that include (1) providing information on future increases in shallot prices, (2) To find out the level of data complexity used in the research process, (3) Obtaining the algorithm with the best performance in predicting future increases in shallot prices.

2. RESEARCH METHOD

2.1. Shallot dataset

The data used in this study amounted to 1189 data, with the category of food price data based on retail prices, where the data was taken from March 2021 to May 2024 from the website panelharga.badanpangan.go.id. There are several food commodities on the website, but in the implementation process the data used is only onion red data, an example of the data to be used can be seen in the table.

Table 1. Food prices are based on retail prices

Komoditas (Rp)	10/03/2021	11/03/2021	12/03/2021	13/03/2021	.....	.....	29/05/2024	30/05/2024	31/05/2024
Beras Premium	11.500	11.500	9.000	11.500	.....	.....	14.460	14.460	14.470
Beras Medium	11.000	11.000	9.000	11.000	.....	.....	13.300	13.300	13.300
Kedelai Biji Kering (Impor)	12.000	13.000	9.000	13.000	.....	.....	12.240	12.250	12.210
Bawang Merah	35.000	35.000	9.000	40.000	.....	.....	55.780	55.160	53.940
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
Tepung Terigu Kemasan (non-curah)	-	-	-	-	.....	.....	15.030	14.910	14.970

2.2. Metode can implement

2.2.1. LSTM (Long Short-term Memory)

LSTM is an RNN architecture equipped with memory cells [16]. With the memory cell, the LSTM architecture can function more effectively than ordinary recurrent neural networks, because it is able to remember information over a longer period of time, making it a superior algorithm for predicting time series data [17]. Another opinion states that LSTM is a special type of RNN that is more effective in practice because of the updates in the equations and the backpropagation dynamics applied. The Long Short-Term Memory (LSTM) architecture is also equipped with gates that function to delete or add information, namely forget gate, output gate, and input gate [18]. The formula for LTSM can be seen in the equation below [19]:

$$f_t = \sigma_g(W_f x_1 + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_1 + U_f h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_1 + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t * c_f + i_t * \sigma_g(W_C x_t + U_C h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t * \sigma_h(c_t) \quad (5)$$

2.2.2. GRU (Gated Recurrent Unit)

Gated Recurrent Unit (GRU) is an algorithm in Deep Learning that has similar performance with LSTM. However, GRU only has two gates, namely reset gate and update gate [20]. Another opinion states that the Gated Recurrent Unit (GRU) Model is a modified model of the Recurrent Neural Network (RNN) model. The GRU model has a simpler architecture than the Long Short-Term Memory (LSTM) architecture.

In the LSTM model there are 3 gates (input gate, forget gate, and output gate), while in the GRU model there are only 2 gates (update gate and reset gate) [21].

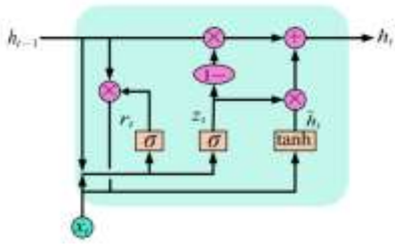

Figure 1. Structure GRU [22]

2.2.3. LR (Linear Regression)

Linear regression is a technique used to evaluate the relationship between two variables by analysing the correlation between the dependent variable and the independent variable through a straight line [23]. Another opinion defines that linear regression is a method for estimating numerical values based on historical data in a certain period of time. There are two types of linear regression, namely one-variable linear regression and multivariable linear regression. In one-variable linear regression, the main focus is to find the correlation between one variable x and independent variable y, while in multivariable linear regression, analysis is carried out to find the relationship between several variables at once [24]. The formula for linear regression can be seen below:

$$y_t = \beta_0 + \beta_1 x_1 + \varepsilon_t \qquad (6)$$

2.3. Design system

In this study, the method used can be seen in Figure 2. In the first step, a literature review is carried out where researchers collect information from various sources such as books or journals as references in preparing their research. then collect data related to shallot prices in the city of pati, then perform preprocessing and cleaning on the data that has been collected. Then the division of training, testing and validation data is carried out. After that, modeling comparisons based on the LSTM LR and GRU algorithms are carried out. data is divided into predictions using the LSTM, GRU and LR algorithms. After that, an evaluation is carried out to get the results.

Figure 2. Flowchart method used

In the first step, a literature review is carried out where researchers collect information from various sources such as books or journals as references in preparing their research. then collect data related to shallot prices in the city of Pati, then preprocessing and cleaning the data that has been collected. Then the division of training, testing and validation data is carried out. After that, modelling comparisons based on the LSTM LR and GRU algorithms are carried out. data is divided into predictions using the LSTM, GRU and LR algorithms. After that, an evaluation is carried out to get the results.

3. RESULTS AND DISCUSSION

In the process of modelling there are several stages carried out, namely, EDA (exploratory Data Analysis) this part to know, information data base on datasets can be used, Preprocessing includes (cleansing data, Minmax Scaler, shift data, split data), modelling (with comparison system) and evaluation using RMSE, MAPE and MAE.

3.1. Preprocessing data

Before the data is used in building a model, a preprocessing process is needed so that the data can work optimally, as for some of the preprocessing processes carried out, including:

3.1.1 EDA & Cleaning data

Data from *panelharga.badanpangan.co.id* is still raw data, so need to select date data and Onion dataset only. Based on table 1, it is known that the date data records are still in the form of columns, so it is necessary to transpose the data from columns to rows.

The process of transposing data and creating new data utilizes the pandas library. The new data contains date data as index and Onion data as value. In the data used does not have null data but there is some data that contains **-**, data that has the value **-** must be deleted because the data is the same as null, after the process of removing the data comes the number that can be used is 1082. In addition to checking the missing value, it is necessary to check the duplicate data, the checking results show that 1080 data are duplicated, so there is no need to remove duplicate data.

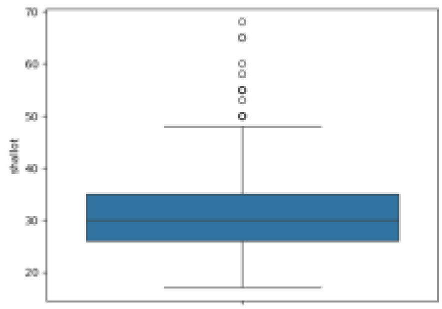

Figure 3. Box plot

Time series data is identical to the value of outliers, this value is obtained when prices experience unstable fluctuations, so the process of removing outliers is needed. The outlier removal technique in the research was conducted using IQR.

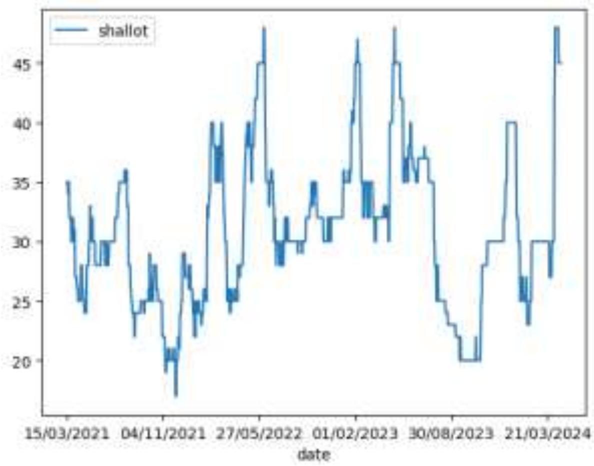

Figure 4. After the remove outlier process

After the data cleaning process, 1029 data were obtained that were relatively clean from outliers. The visualization of the outlier-cleaned data can be observed in Figure 3 which shows a more stable and representative price pattern.

3.1.2. MinMaxScaler

To reduce overfitting and make data confidence, it is necessary to normalize the data, the normalization technique used is minmaxscaler. Minmaxscaler is a way to transform data [22]. To make balancing between data where the content process uses a range of 0-1. The formula of minmaxscaler can be seen below:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (7)$$

The results of the minmaxscaller process can be seen in the table 2 below:

Table 2. Result minmaxscaller

date	actual	scaller
15/03/2021	35.0	0.580645
16/03/2021	35.0	0.580645
17/03/2021	35.0	0.580645
…..	….	….
24/05/2024	45.0	0.903226
25/05/2024	45.0	0.903226
26/05/2024	45.0	0.903226

3.1.3. Shape dataset with shift data

After scaling, the data is still unsupervised learning so the data needs to be converted into supervised learning by shifting the data, the number of data shifts used is 1, because in this study applying a comparison system with Linear Regression, where linear regression only needs 1 input. The results of the shifted can be seen in the table 3.

Table 3. Results of the shifted

x	y
0.58064516	0.58064516
0.58064516	0.58064516
0.58064516	0.5483871
....	......
0.90322581	[0.90322581
0.90322581	0.90322581
0.90322581	0.90322581

3.2. Split data

There are several percentages used in splitting data, such as training 70% testing 30%, training 80% testing 20%,[23], or training 70% validation 10% and testing 20% [24]. In this study using the provisions of 70% training data, 20% testing and 10% validation.

Table 4. Results of the shifted

Training	Testing	Validation
713 data	275 data	31 data

3.3. comparison for LSTM, GRU, LR

The comparison process aims to find the algorithm with the lowest error value using MAPE, MAE and RMSE metrics.

3.3.1. LSTM (Long Short-Term Memory)

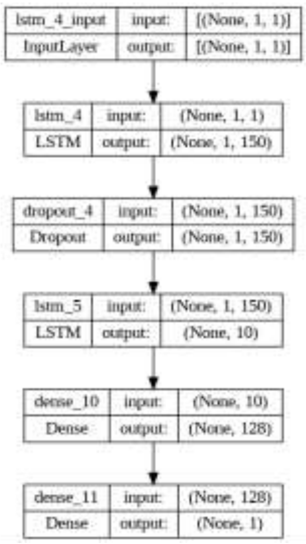

Figure 5. LSTM Flowchart Analysis

The LSTM training process uses a batch size of 32 with 100 epochs (but in the implementation process utilizes early stopping, this process stops the training process when there is no decrease in error). The oprimizer used in the training process is Adam with loss (MAE). The training results can be seen in Figure 6.

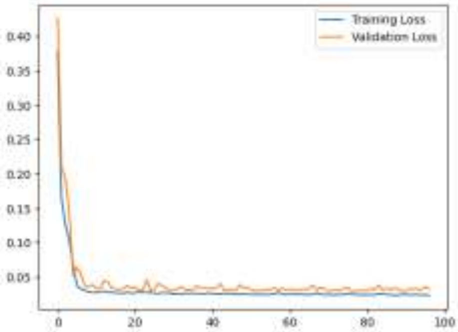

Figure 6. The loss graph based on the best results from training on the daily dataset

The prediction training can be seen in Figure 6, where the prediction value is almost accurate, but based on the training process the MAE value does not decrease so that it can be potentially overfitting, so a further modelling process is needed.

Based on the evaluation process, the MAE value is 0.011072172783, MAPE 3.93678% and RMSE 0.03139695060. The MAPE value is included in the low error limit because it is less than 10%.

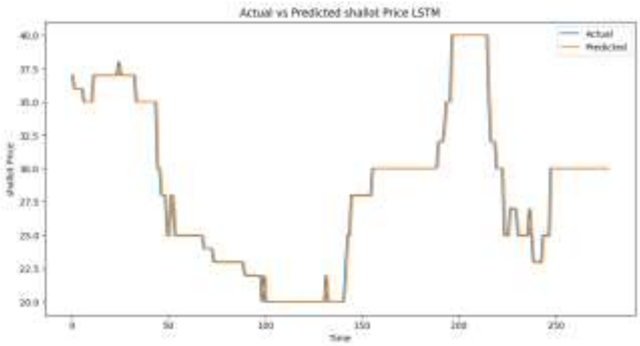

Figure 7. prediction using LSTM

3.3.2. GRU (Gated Recurrent Unit)

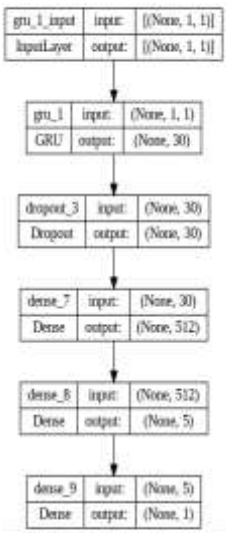

Figure 8. GRU Flowchart Analysis

The training process uses the same scheme as LSTM, but the model experiences overfitting so the process stops at epoch 28. There are several possibilities that occur because GRU experiences overfitting, the first is that the model used is not suitable, there is not much data, the CPU/GPU used is not fast enough so that the data experiences overfitting.

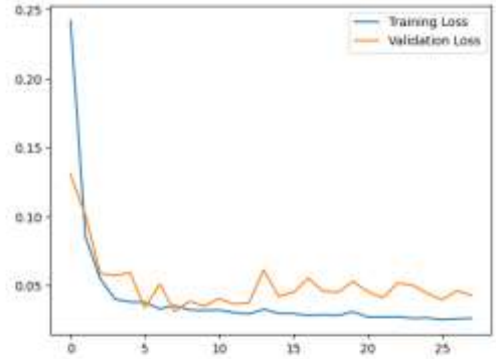

Figure 9. The loss graph training in GRU

Prediction results can be seen in Figure 10, where the MAE value is 0.01185741, MAPE 4.2282357% and RMSE 0.03122299395. These results show that the MAPE value of LSTM higher than LSTM, so if used the prediction results are less accurate.

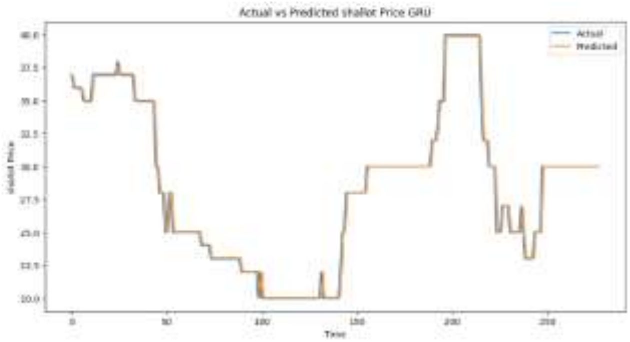

Figure 10. Prediction using GRU

3.3.3. Linear Regression

Unlike LSTM and GRU, Linear Regression is not included in the deep learning family, so the structuring model used is different, for the case of the dataset used using 2 variables, 1 as training data and 1 as a label. The MAE value is 0.0134737280395416, MAPE is 5.45081% and RMSE is 0.0313332635305961. Where the modelling process does not set certain units, but uses the default model, if using more complex training data then use multiple regression.

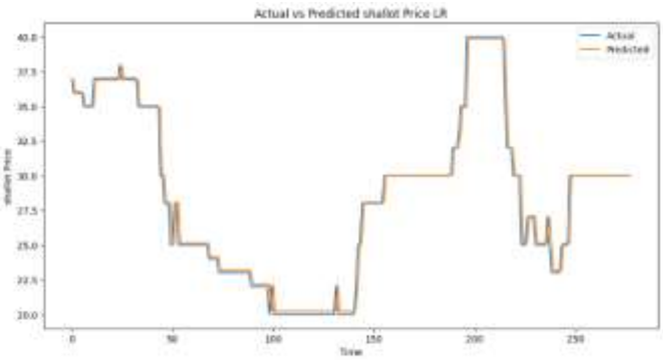

Figure 11. Prediction using Linear Regression

4. CONCLUSION

Based on the evaluation of the LSTM, GRU, and Linear Regression algorithms using data from March 2021 to May 2024, it is found that LSTM has the lowest error value with MAE value is 0.011072172783, MAPE 3.93678% and RMSE 0.03139695060 Although the MAPE value of 3.93678% is included in the excellent category, to improve prediction accuracy, it is necessary to strive for the MAPE value to be below 1% in subsequent models. To achieve this, several steps can be taken, such as increasing the amount of data with higher complexity and using a different optimizer for LSTM to avoid overfitting.

REFERENCES

[1] C. D. Y. Lensun, J. R. Mandei, and J. F. J. Timban, "Adopsi Petani Terhadap Inovasi Alat Pertanian Modern Padi Sawah di Kelurahan Woloan Dua Kecamatan Tomohon Barat Kota Tomohon," *Agri-SosioEkonomi Unsrat,* vol. 15, no. 2, pp. 355–362, 2019.

[2] S. Kasuba, V. Panelewen, and E. Wantasen, "POTENSI KOMODITI UNGGULAN AGRIBISNIS HORTIKULTURA DAN STRATEGI PENGEMBANGANNYA DI KABUPATEN HALMAHERA SELATAN," 2015.

[3] D. Pitaloka, "Hortikultura: Potensi, Pengembangan Dan Tantangan," *G-Tech (Jurnal Teknologi terapan)*, vol. 1, no. 1, pp. 1–4, 2017.

[4] T. Wahyudi, *Pengelolaan komoditas hortikultura unggulan berbasis lingkungan*. Lombok Tengah: Forum Pemuda Aswaja, 2020.

[5] R. P. Wibowo and N. J. R. Surbakti, "Faktor-Faktor yang Mempengaruhi Permintaan dan Penawaran Bawang Merah di Indonesia," *Agro Bali : Agricultural Journal*, vol. 6, no. 2, pp. 326–336, Oct. 2023, doi: 10.37637/ab.v6i2.1312.

[6] M. K. Alfin, A. Alim Murtopo, and N. Fadilah, "Penerapan Metode Clustering untuk Prediksi Produksi Bawang Merah (Ensemble K-Nearest Neighbors)," *IJIR*, vol. 3, no. 2, pp. 30–37, 2022.

[7] O. A. D. Setyowati, "Peramalan Harga Cabai Rawit Di Provinsi Jawa Timur Menggunakan Metode Arimax," Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya, 2020.

[8] P. Harga Telur Ayam Ras Pada Hari Besar, R. Prastika Destiarni Universitas Pembangunan Nasional, J. Timur Jalan Raya Rungkut Madya, G. anyar, and G. Anyar, "The Forecasting of Broiler Egg Price on Religious Holiday In East Java Market," *Berkala Ilmiah Agribisnis AGRIDEVINA*, vol. 7, no. 1, pp. 62–76, 2018.

[9] K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," *JUNIF (Jurnal Nasional Informatika)*, vol. 1, no. 2, pp. 82–94, 2020, [Online]. Available: www.data.jakarta.go.id.

[10] A. Dwi, A. Nasharudin, and U. Ependi, "Analisis Peramalan Penjualan Produk Pada PT.Enseval Putera Mega……," *Jurnal JUPITER*, vol. 15, no. 1, pp. 317–326, 2023.

[11] W. Wiratama, L. Aulia Alifah, A. Gurusinga, E. Indra, J. Sistem Informasi, and F. Sains Dan Teknologi, "Prediksi Turis Mancanegara ke Indonesia Menggunakan Metode EDA Time Series dan LSTM," *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)*, vol. 8, no. 2, pp. 524–537, 2023, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

[12] A. O. Sihombing *et al.*, "Analisis Korelasi Sektor Pertanian Terhadap Tingkat Kemiskinan di Provinsi Sumatera Utara," *Jurnal Agribisnis Sumatera Utara*, vol. 12, no. 1, 2019, doi: 10.31289/agrica.v12i1.2220.g1899.

[13] C. P. Trisya, N. W. Azani, L. M. Sari, H. Handayani, and M. R. M. Alhamid, "Performance Comparison of ARIMA, LSTM and SVM Models for Electric Energy Consumption Analysis,"

Public Research Journal of Engineering, Data Technology and Computer Science, vol. 1, no. 2, Feb. 2024, doi: 10.57152/predatecs.v1i2.869.

[14] T. B. Shahi, A. Shrestha, A. Neupane, and W. Guo, "Stock price forecasting with deep learning: A comparative study," *Mathematics*, vol. 8, no. 9, Sep. 2020, doi: 10.3390/math8091441.

[15] M. Aditya Pratama, M. Munawaroh, W. Joko Pranoto, P. Studi Teknik Informatika, F. Sains dan Teknologi, and U. Muhammadiyah Kalimantan Timur, "Perbandingan Performa Algoritma Linear Regresi dan Random Forest untuk Prediksi Harga Bawang Merah di Kota Samarinda," *Jurnal Ilmu Teknik*, vol. 1, no. 2, pp. 172–182, 2024, doi: 10.62017/tektonik.

[16] A. M. Bahador, "The accuracy of the LSTM model for predicting the S&P 500 index and the difference between prediction and backtesting," *DEGREE PROJECT TECHNOLOGY*, 2018.

[17] H. Prasetyanwar, "Peramalan Nilai Tukar IDR-USD Menggunakan Long Short Term Memory," in *e-Proceeding of Engineering*, 2018, pp. 3820–3826.

[18] S. Sen, D. Sugiarto, and A. Rochman, "Komparasi Metode Multilayer Perceptron (MLP) dan Long Short Term Memory (LSTM) dalam Peramalan Harga Beras," *ULTIMATICS*, vol. XII, no. 1, p. 35, 2020.

[19] U. P. Iskandar and M. Kurihara, "Long Short-term Memory (LSTM) Networks for Forecasting Reservoir Performances in Carbon Capture, Utilisation, and Storage (CCUS) Operations," *Scientific Contributions Oil and Gas*, vol. 45, no. 1, pp. 35–50, 2022, doi: 10.29017/SCOG.45.1.943.

[20] N. Giarsyani, A. F. Hidayatullah, and R. Rahmadi, "Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 3, pp. 48–57, 2020.

[21] A. Nilsen, "Perbandingan Model RNN, Model LSTM, dan Model GRU dalam Memprediksi Harga Saham-Saham LQ45," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 1, 2022.

[22] C. Li, Q. Guo, L. Shao, J. Li, and H. Wu, "Research on Short-Term Load Forecasting Based on Optimized GRU Neural Network," *Electronics (Switzerland)*, vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223834.

[23] A. A. Suryanto and A. Muqtadir, "Penerapan Metode Mean Absolute Error(MEA) dalam Algoritma Regresi Linear untuk Prediksi Produksi Padi," *SAINTEKBU: Jurnal Sains dan Teknologi*, vol. 11, no. 1, p. 11, 2019.

[24] C. Haryawan and M. M. Sebatubun, "Implementation of Multilayer Perceptron for Student Failure Prediction," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 18, no. 2, p. 125, Jul. 2020, doi: 10.12962/j24068535.v18i2.a990.