Prediction of Oil Palm Production Using Recurrent Neural Network Long Short-Term Memory (RNN-LSTM)

Muhdan Syarovy[1,2], Andri Prima Nugroho[1(✉)], Lilik Sutiarso[1], Suwardi[1,3], Mukhes Sri Muna[1], Ardan Wiratmoko[1], Sukarman[3], and Septa Primananda[3]

[1] Smart Agriculture Research, Department of Agricultural and Biosystems Engineering, Faculty of Agricultural Technology, Universitas Gadjah Mada, Jln. Flora No.1 Bulaksumur, Yogyakarta 55281, Indonesia
andrew@ugm.ac.id

[2] Indonesia Oil Palm Research Institute, Jln. Brigjen Katamso, No. 51, Medan 20158, Indonesia

[3] Wilmar International Plantation, Region Kalimantan Tengah, Indonesia

Abstract. Prediction of oil palm production is essential so all activities can be planned effectively and efficiently, especially in financing. One can do many ways, one of which is utilizing production history data using the Recurrent Neural Network – Long Short-Term Memory (RNN-LSTM) model. RNN-LSTM is a Deep Learning model that can be used to predict based on sequential data. This study aims to see the performance of the RNN-LSTM model in predicting oil palm production. The annual production history data for 11 years from the division and estate levels were used. There were four inputs tested from the data used, namely 3, 5, 7, and 9 inputs. The results showed that nine inputs could predict well with MSE, MAE, and MAPE, respectively 1.186, 0.732, and 0.030 at the time of model validation and 39.711, 4.210, and 0.154 at the time of model evaluation.

Keywords: Artificial Intelligence · RNN · LSTM · Oil Palm · Prediction

1 Introduction

The increase in oil palm production is only carried out by area expansion or extensification, which can be seen from 2011–2019, an increase in oil palm area by more than 5 million hectares [1]. However, there is still a large gap between actual and potential production. Therefore, increasing production can also be done with the intensification [2]. The application of precision agriculture is one way of agricultural intensification to increase productivity by applying sustainability principles, protecting land resources, and minimizing production costs [3]. The ability to make predictions is one of the focuses of precision agriculture. Moreover, in the era of the industrial revolution 4.0, by utilizing the internet of things (IoT), big data, and artificial intelligence (AI), researchers have developed various models to improve prediction accuracy [4].

Prediction of production in oil palm plantations is essential to make the right decisions that all activities can do effectively and efficiently, especially to increase production and

J. Sumantyo et al. (Eds.): ICoSIA 2022, ABSR 29, pp. 55–66, 2023.
https://doi.org/10.2991/978-94-6463-122-7_6

its relation to cost. Making predictions can be made mechanistically or empirically. The mechanistic model has a higher complexity, so it is challenging to apply compared to the empirical model. Meanwhile, empirical models are built based on experiments [5]. Moreover, plantations always record all their activities, especially those related to production, so this data is not difficult to obtain on plantations. However, these data have yet to be appropriately utilized and have even been lost [6]. But compared to mechanical models, empirical models cannot explain the various mechanistic processes that occur and the accuracy of the empirical model is very low [7]. Currently, researchers are using intelligent algorithms to improve the accuracy of empirical models. Some of these models include Support Vector Machines (SVM), Back-Propagation Neural Network (BPNN), Artificial Neural Network (ANN), and Deep Neural Network (DNN) [5].

A Recurrent Neural Network (RNN) is an artificial neural network architecture whose processing is called repeatedly to process input, usually sequential data. [8]. RNN is included in the category of deep learning because the data is processed through many layers. RNN has progressed rapidly and has revolutionized fields such as natural language processing (NLP), speech recognition, music synthesis, time series financial data processing, DNA series analysis, video analysis, and others. RNN – Long Short-Term Memory (LSTM) is another type of processing module for RNN. LSTM was created by Hochreiter & Schmidhuber in 1997 and later developed and popularized by many researchers. LSTM is used because of the vanishing gradient problem as the length of the sequential data increases and the number of layers to be trained increases. A vanishing gradient is a situation where the gradient value used to update the weights on the neuron is 0 or close to 0 [9–11].

Research using LSTM has been carried out by several researchers, such as predicting oil palm production in Indonesia [12], predicting crude palm oil (CPO) prices [13], predicting various agricultural commodity prices [14], and many more. Therefore, the LSTM model can be used to predict oil palm production based on historical data on oil palm production. This study aims to see the performance of the LSTM model in predicting oil palm production based on time-series data on oil palm production.

2 Material and Methods

This study used production data that has been collected for 11 years. The plantation area in this study was 4630.60 ha which was divided into five divisions. Divisions 1 to 5 each had an area of 817.40 ha, 1044.87 ha, 757.14 ha, 989.09 ha, and 1022.10 ha. Production data from these five divisions were collected to become the dataset to be modeled. The oil palm annual data production needed a more evident pattern, so this study simulated several inputs. Some of these inputs were 3, 5, 7, and 9. After that, the inputs were modeled using LSTM (Fig. 1).

The sequential arrangement of oil palm production data in this study can be seen in Fig. 2 and Table 1. Data from 2011–2019 were used for training or model development. This data had been divided into 3, 5, 7, and 9 inputs, whose composition can be seen in Table 1. Then the resulting model was validated using 2020 data and tested using 2021 data.

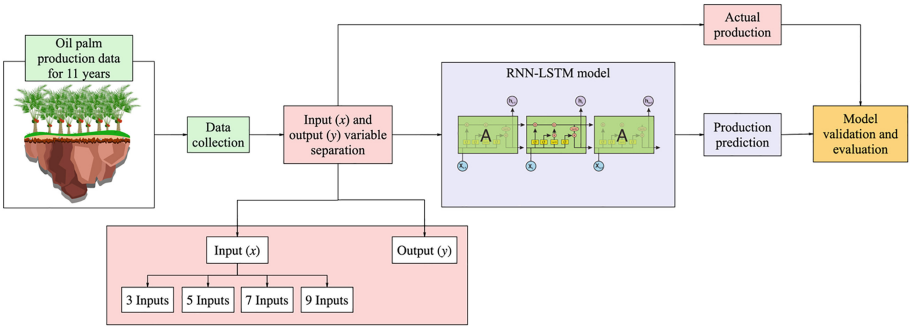

Fig. 1. Illustration of the research framework

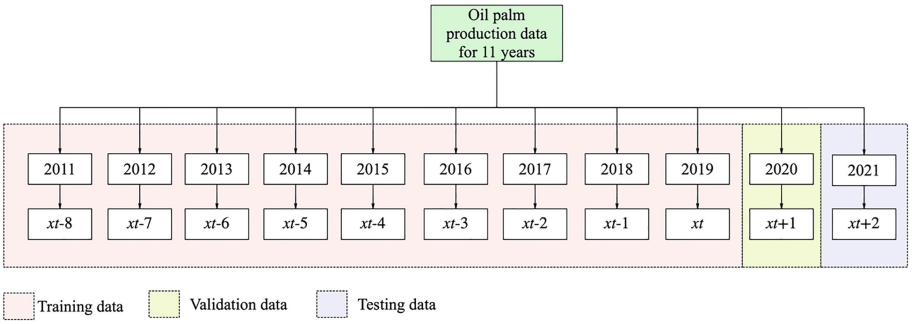

Fig. 2. The sequential arrangement of oil palm production data in the study

This research was divided into four stages: data collection and preprocessing, model development, model validation, and model evaluation (Fig. 3). Data collection and preprocessing began with an array of production data collected for 11 years. After that, the linear interpolation method imputed discarded outlier data and missing values. After that, this data was converted into a range of 0 – 1 using Min-max Normalization. Linear interpolation and Min-Max Normalization can be seen in Eqs. (1) and (2).

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{1}$$

Where:
$f_1(x)$ = Missing value
$f_1(x_0)$ = The value of the dependent variable from the previous data
$f_1(x_1)$ = The value of the dependent variable from the data afterwards
x_0 = The value of the independent variable from the previous data
x_1 = The value of the independent variable from the data afterwards.

$$X_{new} = \frac{(X_{old} - X_{min})x(X_{newmax} - X_{newmin})}{X_{max} - X_{min}} + X_{newmin} \tag{2}$$

Table 1. The sequential arrangement of oil palm production data for 3, 5, 7 and 9 inputs.

	Input (*x*)	**Output (*y*)**
3 inputs	*xt-3, xt-2, xt-1*	*xt*
	xt-4, xt-4, xt-2	*xt-1*
	xt-5, xt-4, xt-3	*xt-2*
	xt-6, xt-5, xt-4	*xt-3*
	xt-7, xt-6, xt-5	*xt-4*
	xt-8, xt-7, xt-6	*xt-5*
5 inputs	*xt-5, xt-4, xt-3, xt-2, xt-1*	*xt*
	xt-6, xt-5, xt-4, xt-3, xt-2	*xt-1*
	xt-7, xt-6, xt-5, xt-4, xt-3	*xt-2*
	xt-8, xt-7, xt-6, xt-5, xt-4	*xt-3*
7 inputs	*xt-7, xt-6, xt-5, xt-4, xt-3, xt-2, xt-1*	*xt*
	xt-8, xt-7, xt-6, xt-5, xt-4, xt-3, xt-2	*xt-1*
9 inputs	*xt-8, xt-7, xt-6, xt-5, xt-4, xt-3, xt-2, xt-1*	*xt*

where:
X_{new} = Normalized data
X_{old} = Data before normalization
X_{min} = The smallest data from a single column of data rows
X_{max} = The largest data from a single column of data rows
X_{newmin} = Minimum value limit of normalization
X_{newmax} = Maximum value limit of normalization

The dataset used to predict production was time series data based on divisions and estate for 11 years. This data was divided into input and output variables. The data used for model development or training was from 2010 to 2019. The model was validated using 2020 data and evaluated using 2021 data. Evaluation of the model used mean absolute percentage error (MAPE), mean squared error (MSE), and mean absolute error (MAE) in Eqs. (3), (4), and (5).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{Y - \hat{Y}}{\hat{Y}} \tag{3}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y - \hat{Y}\right)^2 \tag{4}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y - \hat{Y}\right| \tag{5}$$

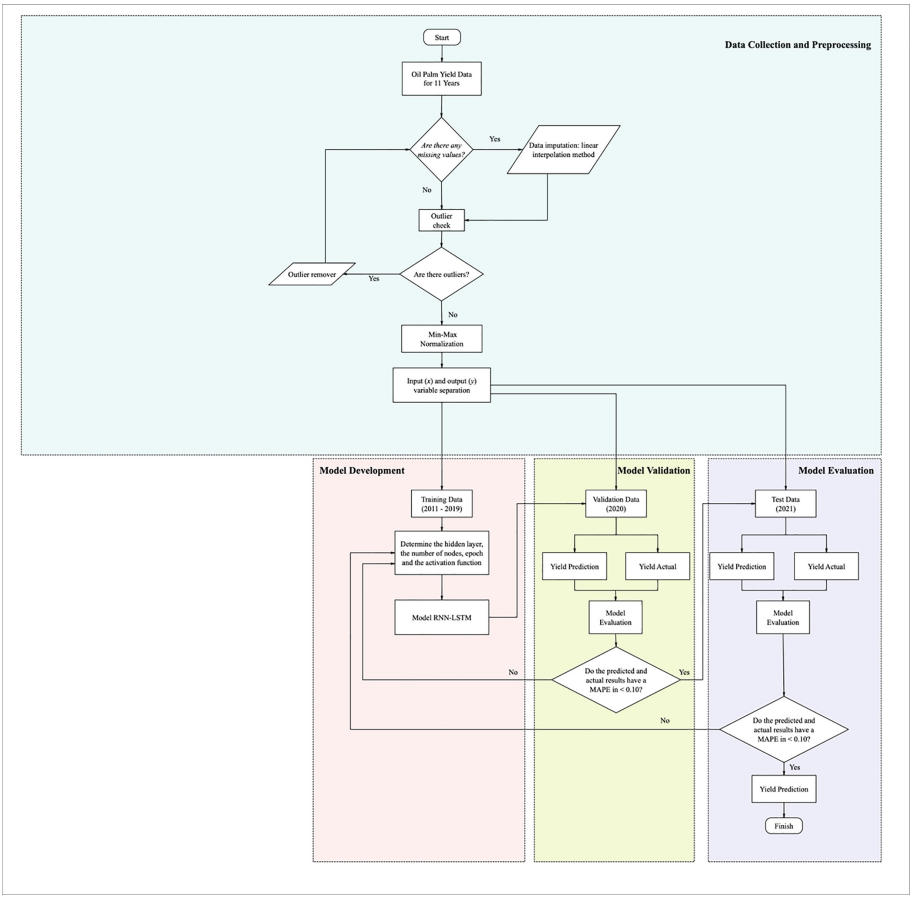

Fig. 3. Stages in the research

where:
n = Number of data
Y = Actual
$\hat{Y}$ = Prediction

In general, LSTM calculations used python 3.7 with Keras and Tensorflow libraries. LSTM architecture has more complex cell contents than RNN. LSTM has four structures that will process time series data (sequential).

Figure 4 shows that the LSTM has two outputs, where one output will be used as the following cell input, and the other output is the cell state. Inside each cell, there are four structures that each have a function to process data.

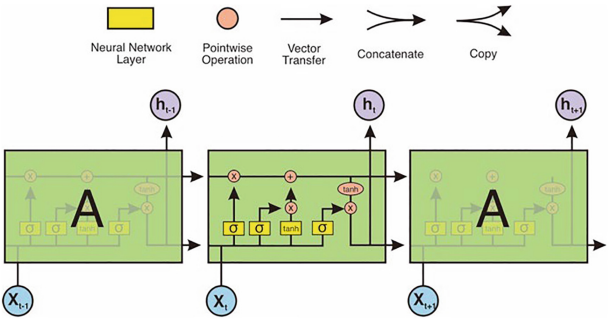

Fig. 4. Structures in the LSTM architecture

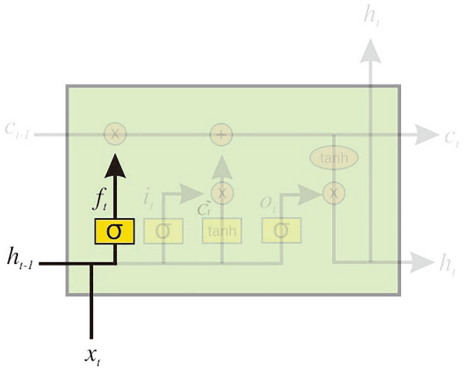

Fig. 5. Information flow forget gate

2.1 Forget Gate

The LSTM's first step is to decide what information to keep or discard. This decision is based on a neural network architecture with a sigmoid activation function with an output of 0 to 1. If the result is close to 0, the greater the information from the previous output (h_{t-1}) and input (x_t) will be forgotten or will not be a correction factor in the cell state. The forget gate equation (f_t) can be seen in Eq. 6 (Fig. 5).

$$f_t = \sigma(W_f \cdot [h_t - 1, x_t] + b_f) \tag{6}$$

2.2 Input Gate

The input gate (i_t) (in Eq. 7) with the sigmoid neural network layer decides the value to update. Then the tanh neural network layer creates a new candidate value vector ($\tilde{C}_t$) (Eq. 8) (Fig. 6).

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \tag{7}$$

$$\tilde{C}t = tanh(W_c \cdot [h_t - 1, x_t] + b_c) \tag{8}$$

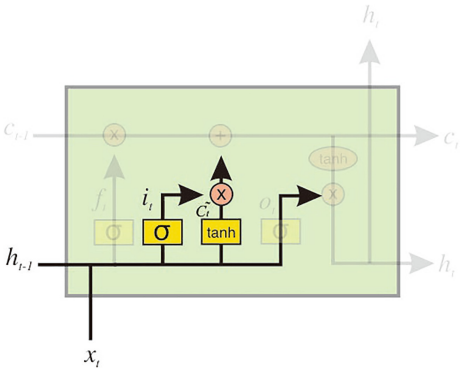

Fig. 6. Information flow input gate

2.3 Update Cell

The next step is to update the old cell state (C_{t-1}) into the new state (C_t) by multiplying the old condition by f_t, then adding $i_t * \tilde{C}_t$ to get the new Candidate value (Ct) as in Eq. 9.

$$C_t = f_t * C_t - 1 + i_t * C_t \tag{9}$$

2.4 Output Gate

At the output gate (O_t), the first thing that is executed is the sigmoid neural network layer which determines what part of the cell will be output (Eq. 10). Next, the cell state is processed on tanh and multiplied by the output of the sigmoid neural network layer (h_t) as in Eq. 11.

$$O_t = \sigma(W_O \cdot [h_t - 1, x_t] + b_O) \tag{10}$$

$$h_t = O_t + \tanh(C_t) \tag{11}$$

3 Results and Discussion

This study used inputs 3, 5, 7, and 9 to predict oil palm production. The nine inputs were the inputs with the fastest errors reaching the lowest error pointed during training (Fig. 7). Seven inputs also provided more closed error reduction than others, but the smallest error performance was when more than 350 epochs. Meanwhile, 3 and 5 inputs require more than 100 epochs to reach the lowest error point, which indicates that a large variety of inputs would be made if the model continued to learn patterns for optimal accuracy (Table 2).

Figure 8 shows a graph of the actual and predicted values for various inputs at the model validation stage. At this stage, the data used by the validation model was 2020

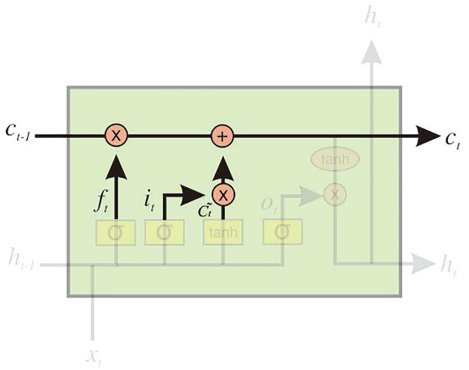

Fig. 7. Information flow update cell

data. Nine inputs showed the slightest error with MSE, MAE, and MAPE values of 1.186, 0.732, and 0.030, respectively. Meanwhile, seven inputs gave the most significant error with MSE, MAE, and MAPE values of 2,512, 1. 325, and 0. 059, respectively (Fig. 11).

Table 2. Model validation dan evaluation of the LSTM Model

		3 inputs	5 inputs	7 inputs	9 inputs
Model Validation	MSE	2.266	2.069	2.512	1.186
	MAE	1.181	1.247	1.325	0.732
	MAPE	0.048	0.050	0.059	0.030
Model Evaluation	MSE	10.042	9.656	39.711	4.538
	MAE	2.797	2.544	4.210	1.681
	MAPE	0.123	0.109	0.154	0.069

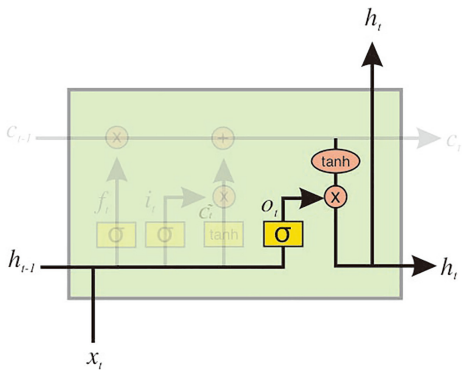

Fig. 8. Information flow output gate

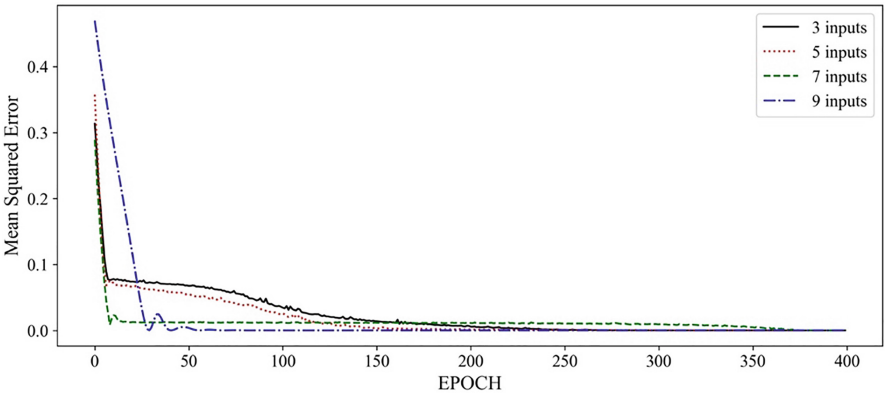

Fig. 9. Epoch at 3, 5, 7, dan 9 input during training

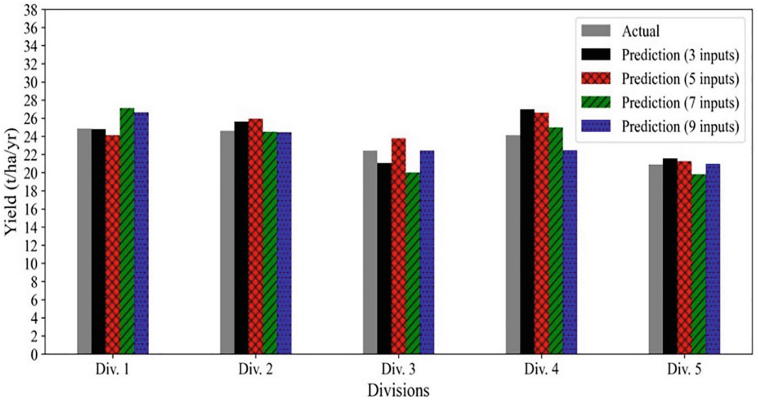

Fig. 10. The actual and predicted yield on the division level on model validation

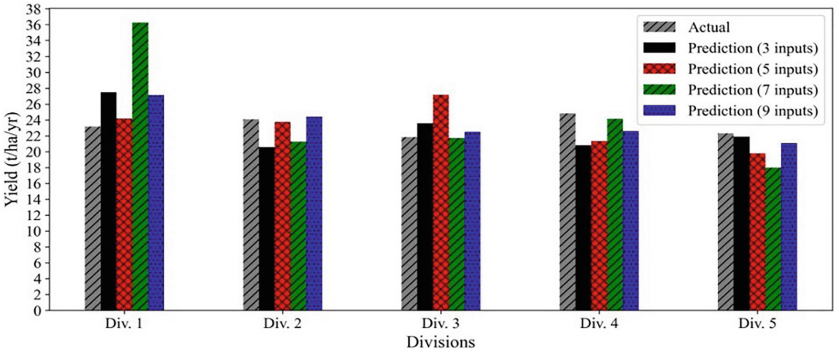

Fig. 11. The actual and predicted yield on the division level on model evaluation

At the model evaluation stage using 2021 data, nine inputs also gave the smallest error even though the error increased compared to the model validation stage. These errors had MSE, MAE, and MAPE values of 4.538, 1.681, and 0.069, respectively. Meanwhile, the seven inputs still gave the biggest error with MSE, MAE, and MAPE values of 39.711, 4.210, and 0.154, respectively. Figure 9 shows a graph of the actual and predicted values for various inputs at the model evaluation stage.

Figure 10 shows the performance of the predictions of various inputs with the actual production conditions of the plantation. Predictions with nine inputs were very close to actual production in 2020 and 2021. In contrast to predictions with inputs 3, 5, and 7, wherein in 2020, they were pretty good at predicting, but in 2021 there had been an enormous prediction gap.

The LSTM model used sequential data with a pattern as the input unit. For oil palm plantations, using the LSTM model was very appropriate if a plantation had long production history data, had the smallest unit of time (e.g., monthly), and only one production variable was recorded on the same timestamp. In contrast, other variable data could have been recorded better. In addition, this model was exact in predicting large plantation areas, for example, at the estate to division level. In contrast, the block level was complicated because blocks were usually related to the planting year, which sometimes did not have long sequential data and required repeated training processes depending on the number of blocks in the estate. Thus, the LSTM model could not be used to simulate various conditions based on historical data, such as the multi-layer perceptron model carried out in some studied [6, 7, 15, 16].

In general, research using LSTM uses datasets that have repeating patterns. Thus, determining the sequence as input can consider the pattern [17]. In this study, various inputs 3, 5, 7, and 9 were used because no clear pattern was found in the sequential annual production data. If monthly production was used, it usually had the same monthly production distribution patterns yearly. Thus, the input pattern used could be used every multiple of 12 months, 24 months, and so on, depending on the amount of data. Therefore, further research could be carried out with sequential monthly production data.

4 Conclusion

The LSTM model could predict sequential data, including oil palm production. For the prediction of the annual output using data for 11 years, this model could predict production quite well with low error. In this study, nine inputs gave the best results indicating that more sequential data would make the best prediction. However, it was highly recommended for further research to use monthly production data to have long data sequences and input layers set every 12 months (January to December), which usually had a similar pattern every year. This model was also highly recommended for predicting with extensive unit data (e.g., a combination of various plantations in a particular area or several blocks to form a divisional unit). For predicting block level or simulating multiple conditions (e.g., in different varieties, watered deficit conditions, etc.), the multilayer perceptron model was very appropriate.

Acknowledgments. Acknowledgments to PT. Kerry Sawit Indonesia, Wilmar Plantation for the opportunity to conduct this research.

References

1. Direktorat Jenderal Perkebunan. Statistik Perkebunan Unggulan Nasional 2019 - 2021. *Sekretariat Dirjend Perkebunan Kementerian Pertanian* 2020; 1056 pp.
2. Monzon JP, Slingerland MA, Rahutomo S, et al. Fostering a climate-smart intensification for oil palm. *Nature Sustainability* 2021; 4: 595–601.
3. Hakkim V, Joseph E, Gokul A, et al. Precision Farming: The Future of Indian Agriculture. *J App Biol Biotech* 2016; 068–072.
4. Serraj R, Pingali P. *Agriculture & Food Systems to 2050: Global Trends, Challenges and Opportunities*. 2019. Epub ahead of print 1 January 2019. DOI: https://doi.org/10.1142/11212.
5. García-Rodríguez L del C, Prado-Olivarez J, Guzmán-Cruz R, et al. Mathematical Modeling to Estimate Photosynthesis: A State of the Art. *Applied Sciences* 2022; 12: 5537.
6. Syarovy M, Nugroho AP, Sutiarso L, et al. Utilization of Big Data in Oil Palm Plantation to Predict Production Using Artificial Neural Network Model. In: *Manuscript submitted for publication*. Bangka Belitung, 2022, p. 11.
7. Harahap IY, Lubis MES. Penggunaan Model Jaringan Saraf Tiruan (Artificial Neuron Network) untuk Memprediksi Hasil Tandan Buah Segar (TBS) Kelapa Sawit Berdasar Curah Hujan dan Hasil TBS Sebelumnya. *Jurnal Penelitian Kelapa Sawit* 2018; 26: 59–70.
8. Tarkus ED, Sompie SRUA, Jacobus A. Implementasi Metode Recurrent Neural Network pada Pengklasifikasian Kualitas Telur Puyuh. *Jurnal Teknik Informatika* 2020; 15: 137–144.
9. Manaswi NK. RNN and LSTM. In: Manaswi NK (ed) *Deep Learning with Applications Using Python : Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*. Berkeley, CA: Apress, pp. 115–126.
10. Syahram EF, Effendy MM, Setyawan N. Sun Position Forecasting Menggunakan Metode RNN – LSTM Sebagai Referensi Pengendalian Daya Solar Cell. *Journal Of Electrical Engineering And Technology* 2021; 8.
11. Noh S-H. Analysis of Gradient Vanishing of RNNs and Performance Comparison. *Information* 2021; 12: 442.
12. Sugiyarto AW, Abadi AM. Prediction of Indonesian Palm Oil Production Using Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). In: *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. 2019, pp. 53–57.
13. Amal I, - T, - S. Crude palm oil price prediction using multilayer perceptron and long short-term memory. *J Math Comput Sci* 2021; 11: 8034–8045.
14. Gu YH, Jin D, Yin H, et al. Forecasting Agricultural Commodity Prices Using Dual Input Attention LSTM. *Agriculture* 2022; 12: 256.
15. Hermantoro H, Rudyanto R. Modeling and Simulation of Oil Palm Plantation Productivity Based on Land Quality and Climate Using Artificial Neural Network. *International Journal of Oil Palm* 2018; 1: 65–70.
16. Kartika ND, Astika W, Santosa E. Oil Palm Yield Forecasting Based on Weather Variables Using Artificial Neural Network. *Indonesian Journal of Electrical Engineering and Computer Science* 2016; 3: 626–633.
17. Jang J, Han J, Leigh S-B. Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks. *Energy and Buildings* 2022; 255: 111647.