Jurnal Sistem Teknik Industri
Journal homepage: https://talenta.usu.ac.id/jsti

Implementation of Long Short-Term Memory Network for Predicting The Cocoa Crop Yield

Anastasia Lidya Maukar*, Laesa Qotrun Nada Arrosyadi
Department of Industrial Engineering, Faculty of Engineering, President University, Jababeka, 17550, Indonesia
*Corresponding Author: almaukar@president.ac.id

ARTICLE INFO
Article history:
Received 10 January 2024
Revised 10 June 2024
Accepted 20 July 2024
Available online 29 July 2024

E-ISSN: 2527-9408
P-ISSN: 1411-5247

How to cite:
Maukar, A. L & Arrosyadi, L.Q. N. (2024). Implementation of Long Short-Term Memory Network for Predicting The Cocoa Crop Yield. *Jurnal Sistem Teknik Industri*, 26(2), 159-179.

https://doi.org/10.32734/jsti.v26i2.15359

ABSTRACT
Forecasting models with high accuracy become more important during uncertain conditions, such as climate change, that could have a high effect. The forecast model's accuracy in predicting cocoa crop yield must be high to determine decision-making in management. Seven different potential predictor variables have been analyzed in this research to see the influence of cocoa crop yield. Using a scatter plot diagram, six of seven variables, relative humidity, maximum temperature, minimum temperature, evapotranspiration, rainfall, and soil moisture, are proven to influence cocoa crop yield. Then, those datasets are divided into training and validation sets using multiple linear regression analysis and a Long Short-Term Memory (LSTM) network. The output model of those methods is assessed using two metrics: coefficient of determination and Root Means Square Error (RMSE). From those model performance metrics, LSTM outperformed multiple linear regression analysis. LSTM has an R-square of 98% and an RMSE of 0.3 while multiple linear regression just reached 82% of the R-square and 2.57 of the RMSE. The LSTM model has been proven to be valid.
Keyword: Coefficient of Determination, Crop Yield, Forecasting, Long Short-Term Memory Network, Regression

ABSTRAK
Model peramalan dengan tingkat akurasi tinggi menjadi lebih penting pada kondisi tidak menentu, seperti perubahan iklim. Hal ini diperlukan karena model peramalan digunakan untuk menentukan manajemen pertanian pada kakao. Terdapat tujuh variabel prediktor berbeda yang dianalisis dalam penelitian ini untuk melihat pengaruhnya terhadap hasil tanaman kakao. Dengan menggunakan diagram scatter plot, enam dari tujuh variabel terbukti mempunyai pengaruh terhadap hasil tanaman kakao yaitu kelembaban relatif, suhu maksimum dan minimum, evapotranspirasi, curah hujan, dan kelembaban tanah. Dataset yang ada kemudian dibagi menjadi dataset training dan dataset validasi dengan menggunakan analisis regresi linier berganda dan LSTM. Model yang dihasilkan dari dua metode tersebut kemudian diuji performanya menggunakan koefisien determinasi dan RMSE. Dari kedua metrik tersebut, LSTM mengungguli analisis regresi linier berganda. LSTM memiliki R-square sebesar 98% dan RMSE sebesar 0.3 sedangkan linear berganda hanya mencapai 82% RMSE dan memiliki RMSE sebesar 2.57. Model LSTM terbukti sebagai model yang valid.
Kata Kunci: Hasil Panen, Koefisien Determinasi, Long Short-Term Memory, Peramalan, Regresi

1. Introduction

Cacao is a species with significant economic significance on a global scale as the primary source of raw materials for chocolate production [1]. The export of goods containing cocoa rose by 2.12% in 2018–2019 worldwide. For smallholder farmers in Indonesia, cacao is one of the most important strategic commodities (besides coffee) and one of the major sources of revenue. It is one of the key products imported by Regional Comprehensive Economic Partnership (RCEP), which strengthens trade ties between ASEAN nations and their five trading partners—China, Japan, South Korea, Australia, Malaysia, and New Zealand [2].

Even though many studies show high prospects for Cocoa, including in Indonesia, unfortunately, from one hectare of the farm, there is just 655,515 kg of production, which is only 27.6% of 2,375,054 kg. It has a small

productivity value compared with Malaysia and Ivory Coast, which gain 1800 kg and 800 kg, respectively, with the same area of farm [3].

During 2003, 2005, and 2011, Indonesia was the second-largest country with a high production of cocoa worldwide, together with Côte d'Ivoire, which held the first rank, and Ghana, which held the third rank. Unfortunately, that position was taken by Ghana in 2005, 2012, and 2013, while Côte d'Ivoire was stable in the first rank. The declining trend is proven in another study. During 2013 until 2022, cocoa production in Indonesia has a problem of declining trend because of plantation area [4]. Many things can affect the productivity of cocoa in Indonesia. Climate change is the variable that has a significant impact to the growth and productivity of cocoa plants. It happens in Indonesia and almost all countries that cultivate cocoa [5]. Climate changes such as minimum temperature, maximum temperature, and rainfall has been proven as the variables that could influence cocoa crop yield production [6]. Because of that, forecasting is crucial to adjust the kind of step that could result in high cocoa production.

Even though numerous studies have been conducted to determine the best or most appropriate forecast model for predicting agricultural crop yield, there is still limited research related to crop yield prediction that specifically addresses cocoa and uses a machine-learning model. Long Short-Term Memory Network is the best method used to predict rice yield. The data comes from 81 counties in China for three years, compared to Support Vector Regression with model performance metrics using Mean Square Error (MSE) [7]. The same conclusion has been reached in the implementation of LSTM in yield data of soybean and corn with RMSE and MSE as the model performance metrics [8], [9]. While multiple linear regression has been proven to have a promising result for predicting the wheat yield through performance metrics such as relative approximation error (RAE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage (MAPE) [10]. The multiple linear regression also produces a good result in predicting rice yield [11].

As proven in several previous studies, cocoa clones can influence the cocoa plant's adaptability to climate change [12]. Cocoa's resistance to climate change and disease is influenced by genetics, this is what will produce cocoa beans that are stable against changes, including climate change and disease [13]. While soil moisture is important because it is related to the soil's ability to store water [14]. A study explained that damaged soil drought brought on by soil erosion will lower crop productivity and soil quality. Land production will decline due to the loss of organic matter and nutrients in damaged soil [15]. Specifically, soil chemical for cocoa plants requires cation exchange capacity, exchangeable bases, pH H2OH2O pH, salinity, base saturation, and organic carbon [16]. This study will assess some data between climate changes, soil conditions, and clones to determine what factors could influence the cacao crop yield and choose the best forecast model to predict cocoa crop yield more accurately. Thus, through the forecast model discussed in this research, any organization, company, or farmer that has cocoa production in Indonesia is expected to determine the farm management of cocoa cultivation preparation to maximize cocoa production. Furthermore, cocoa is a potential agricultural sector in Indonesia that has an important role in the country's economic development [17]. The climate changes tested in this research include rainfall, evapotranspiration, temperature maximum, temperature minimum, and humidity. To fill the gap in the previous research, genotype and soil data have been added to this research as the variables that influence crop yield. All these datasets would be trained and validated using multiple linear regression analysis and the LSTM network. Then, those models would be assessed using RMSE and coefficient of determination. This study is expected to create a forecast model with high accuracy for predicting cocoa crop yield, considering that cocoa is a promising commodity in Indonesia.

2. Methods

2.1. Research Materials

This research will discuss yield of cocoa crop yield and the factors that could influence cocoa crop yield, which will be the subject of this research. The research materials discussed include cocoa clone, yield, relative humidity, maximum temperature, minimum temperature, evapotranspiration, rainfall, and soil moisture. To verify the validity and reliability of this research, the reason for variables selected will be further discussed in the next section which selected a broad range of sources of some academic papers.

1. Cocoa Clone

The original habitat of the cocoa plant comes from tropical forest, which is a shade-loving plant. This plant is cultivated from seed and commonly produced by cloning to create new cocoa plants with specific traits that farmers or researchers seek. This cultivation method is frequently used to overcome the problem of low production potential [18]. The cocoa clones discussed in this research are ICCRI03, ICCRI09, KW516, KW562, MCC02, SUL01, and SUL02. Those clones were found and bred by the Indonesian Coffee and Cocoa Research Institute (ICCRI), which is a research institute that focuses on coffee and cocoa and is owned by the Indonesian government. This material is chosen because there is a significant result in cocoa crop yield production when using different clone used [19]. Genetics in the cocoa plant is proven to be the variable that effects the cocoa plant's adaptability to climate change [19]. The physical and chemical properties of the cocoa beans were considerably influenced by the clone types [12]. There is research that shows that along with climate conditions, cocoa clones can influence how resistant the cocoa plant is to weather and disease, so it can produce stable cocoa beans [13].

2. Crop Yield

In this research, the term 'unit' refers to a standardized measure of land area or cultivation area of cocoa crop, expressed in terms of hectares. The cocoa crop yield is measured in kilograms per unit to measure the amount of cocoa dry beans produced per block under cultivation. This metric allows researchers and farmers to evaluate the effectiveness and productivity of the agricultural practices used, including in cocoa crop yield practices. By expressing cocoa crop yield in kilograms per unit of cultivation area, it helps to provide a more thorough understanding of cocoa productivity and supports knowledgeable decision-making within the farming community.

3. Relative Humidity

Relative humidity, usually called RH, is the proportion of the entire capacity for holding water that the quantity of hydration in the air occupies in relation to its maximum moisture-holding capacity. RH would impact the air-drying capacity to evaporate water from the product being dried. The unit of RH is in percentage (%). A study shows that too high relative humidity could decrease cocoa crop yield production because it can damage the metabolic mechanisms that enable pod growth [15]. Therefore, this data can potentially be a variable in making a forecast model to predict cocoa crop yield.

4. Maximum and Minimum Temperature

Temperature is the measurement of the hotness and coldness of a place that is expressed on different scales, such as Celsius and Fahrenheit, that are commonly used. In this research, the scale of temperature used is Celsius (C). The maximum and minimum temperatures need to be measured because the growth of the cacao plant did not stop at any particular time. The maximum temperature means the top level reached during the period, in this case, during that month of observation. The minimum temperature refers to the minimum level that the temperature reached during that month of observations. The minimum temperature influences average cocoa production in a study case conducted in Sulawesi, Indonesia [20]. This variable needs to be considered as Indonesia is a tropical country.

5. Evapotranspiration

Evapotranspiration is a combination term between evaporation and transpiration that defines a process of water transfer to the atmosphere, which is water evaporation from the soil surface and transpiration through a plant that lives on the earth's surface. Evapotranspiration has high correlations with rainfall, relative humidity, and soil moisture. Cacao grown in the shade used less water because it was less exposed to high temperatures and evapotranspiration [21]. Because cocoa trees are deciduous plants, they deposit many leaves on the soil, further limiting water loss through evaporation [22]. Thus, combining those expected variables can make farm management decision-making easier.

6. Rainfall

Rainfall is the amount of rain that falls from cloud to earth within a specific time. The rainfall measurement in this research is in millimeters units in a month. This variable is proven to be one factor that could influence cocoa production. High rainfall has contributed to high cocoa production [23].

7. Soil Moisture

Soil moisture is the average soil water content as a volumetric ratio. It is measured in m^3/m^3, defines the volume of water per unit volume of soil. This volumetric ratio highlights the relationship between the volume of water and the total volume of the soil. To cultivate cocoa plant, the soil condition should be considered very well; it should have good drainage with a pH level in the range of 6–7, a height of 0–600 meters above sea level, and enough water in the depth of the soil [24]. The relationship between the amount of water in a section and the amount of solids in the soil sample, represented as a proportion, for example (%), is known as the moisture content of soils. When it comes to controlling how water and heat energy are exchanged between the earth's surface and the atmosphere through plant transpiration and soil evaporation, soil moisture is an essential factor [14].

The weather data and yield are recorded monthly from March 2020 until January 2023. These data are divided into two functions: data for modelling and testing. The modelling will use the data from March 2020 until March 2022, while the testing will use the rest. The weather data is taken from open data, which is Open-Meteo ERA5. Open-Meteo is an open-source platform that provides accurate and reliable weather data. It is the latest generation of Reanalysis 5 that the European Centre conducts for Medium-Range Weather Forecast (ECMWF). ERA5 is one of the most precise and complete reanalysis weather datasets. It generates data with high spatial and temporal precision [25].

2.2. Research Methods

2.2.1. Forecast Methods

The detailed flow of this study can be seen in Figure 1. After the observation of cocoa agriculture in Indonesia has been carried out, the research objectives and scope are determined. The area of this research is in East Lampung, which is located at latitude -5.249237 and longitude 105.55228. The data is provided by a company that runs cocoa research and production. They produce cocoa beans to fulfil their own demand and also export and import the cocoa to another company. But they also send their production to other companies to fulfill the demand for cocoa. Some activities were conducted during the initial observation, including finding out the potential of cocoa in Indonesia, the demand, and the productivity to determine the urgency of choosing the best forecast model for predicting cocoa crop yield.

Before choosing the best forecast model for predicting cocoa crop yield, the most important thing is to decide what kinds of factors could influence cocoa crop yield. To select the independent variables, the activity is conducted with a literature study-based approach, finds a gap in the research, and fills the gap by adding the variables that have not been carried out. All those variables are visualized in a scatter plot, where the influence of each variable is shown. Then, to make sure that the chosen variable affects cocoa crop yield, the dataset is analyzed using multiple regression analysis. The regression analysis consists of multiple regression analysis to see the influence of all variables on the response variable and linear regression to see the influence of each independent variable on the response variable. Then, Long Short-Term Memory (LSTM) will create the forecasting model process the relationship of the independent variable to the response variable, no matter how high and low the influential level of independent variables are while Multiple Linear Regression just creates the forecasting model with support of high levels of independent variables.

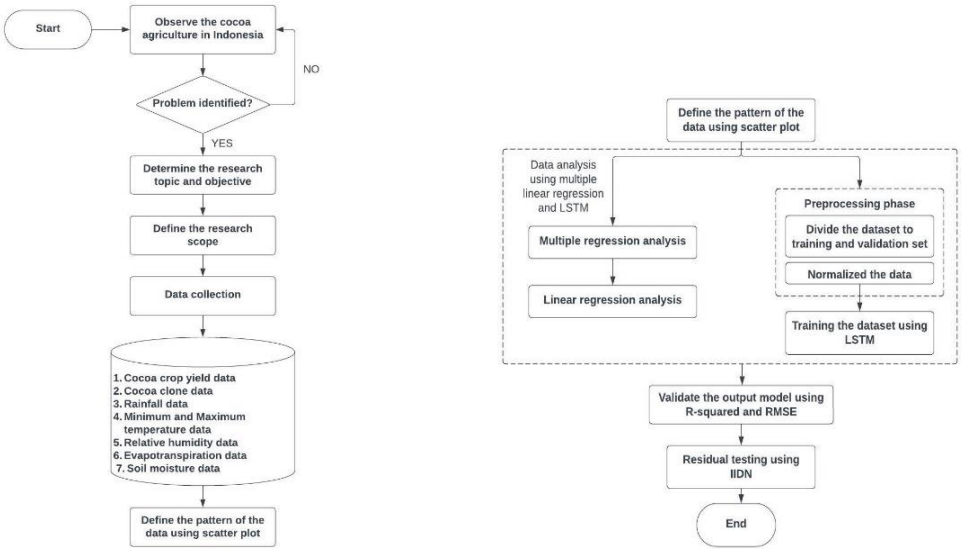

Figure 1. Flowchart of Long Short-Term Memory Network for Cocoa Crop Yield Forecast

1. Linear Regression Analysis

First, multiple linear regression needs to be analyzed. Multiple linear regression can show how rainfall, minimum temperature, maximum temperature, relative humidity, evapotranspiration, soil moisture, and cocoa clone affect cocoa crop yield. Equation (1) is the formula of multiple linear regression.

$$y_i = \beta_0 + \beta_{1\,X_{i1}} + \beta_{2\,X_{i2}} + ... + \beta_{p\,X_{ip}} + e_i \quad (1)$$

Where y_i refers to the predicting value, β (0) is the constant term, and e is the error or residual. Each observed residual in multiple regression analysis is given by Equation (2):

$$e_i = y_i - \hat{y}_1 = y_i - (\beta_0 + \beta_{1\,X_{i1}} + \beta_{2\,X_{i2}} + ... + \beta_{p\,X_{ip}}) \quad (2)$$

Figure 2 shows the steps of conducting multiple regression analysis regarding the forecasting model. Start, where the dataset is collected both dependent and independent variables. There should be more than two independent variables.

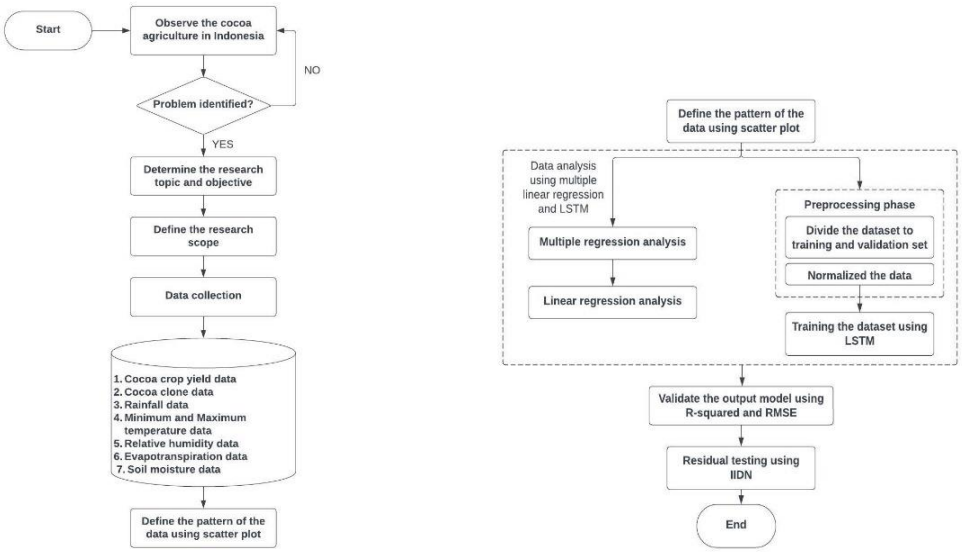

Figure 2. Step to Conduct a Multilinear Regression Analysis [26]

Second, decide whether the independent variable which is not in the equation can be led in. There are some ways to do this step as follows:

Scatter plot: this chart is beneficial to visualize the individual independent variable to dependent variable and see how the relationship is [27]. Some relationships can be seen through scatter plot, including positive, negative, and no relationship.

In this section, ANOVA, Analysis of Variance will be calculations that serve as the foundation for tests of significance for every independent or predictor variable and provide information about the degrees of variability within a regression model that can be calculated through Equation (3).

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\widehat{Y_i} - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 \quad (3)$$

Where it is equal can be computed using Equation (4)

$$SS\ (Total) = SS\ (Regression) + SS\ (Error) \quad (4)$$

SS (Total) is the total variation of the predictor variable being observed (y), SS (Regression) is the variation provided by the Y and x linear relationships or the variation brought on by the fitted linear relationship, and SS (Error) is the variation of error or residual, it is an unexplained variation.

Then, there would be an F distribution in this analysis. The minimum value of F is 0, and there is no maximum. These F-values would be beneficial to calculate the p-value. A larger F-value means it has a high influence on the response variable. Equation (5) is the formula to calculate the F-value.

$$F = \frac{MS\ (Regression)}{MS\ (Error)} \quad (5)$$

Then, the probability value (p-value) really correlated with the alpha of significant value. In hypothesis testing, the probability value (p-value) is used to help decide if the null hypothesis should be rejected. A result is deemed to be "not significant" or of "no importance" if the p-value is greater than 0.05 [28].

Where to identify the multicollinearity in the variable, the variance inflation factor (VIF) is used in this research. Multicollinearity is a situation where the predictor variables in this regression correlate with one another. The value of VIF can make standard errors of the coefficient in the variable high; because of that, it should be prevented [29]. Equation (6) shows how to calculate VIF value.

$$VIF = \frac{1}{1 - R_i^2} \quad (6)$$

Table 1 shows the interpretation of each VIF value.

Table 1. VIF Value Interpretation

No	VIF-value	Example and Use
1	VIF = 1	Not Correlated
2	$1 < VIF \leq 5$	Moderately correlated
3	VIF > 5	Highly correlated

Third, remove the variables. After conducting multiple linear regressions, there would be a conclusion about which predictor variables can influence the response variable and which could not. The variables that have no influence on cocoa crop yield can be removed.

Then, linear regression analysis is conducted. This analysis can see how each predictor variable affects the response variable independently—the prediction of the dependent variable that is represented by y_i and X_1 represent the independent variable. α is the intercept (constant term), β_1 represents the coefficient for X_1, and e refers to the error or residual. The linear regression model can be seen in Equation (7).

$$y_i = \alpha + \beta_1 X_1 + e \quad (7)$$

The error of linear regression can be calculated by Equation (8), where $\hat{y}_1$ represents the forecast result of the linear regression while y_1 is the observed value for the first observation.

$$e_1 = y_1 - \hat{y}_1 \quad (8)$$

Figure 3 shows the detailed steps for conducting a multiple linear regression analysis. In this analysis, variables will be removed if, according to the analysis, those variables are useless and have no influence on the response variable, cocoa crop yield.

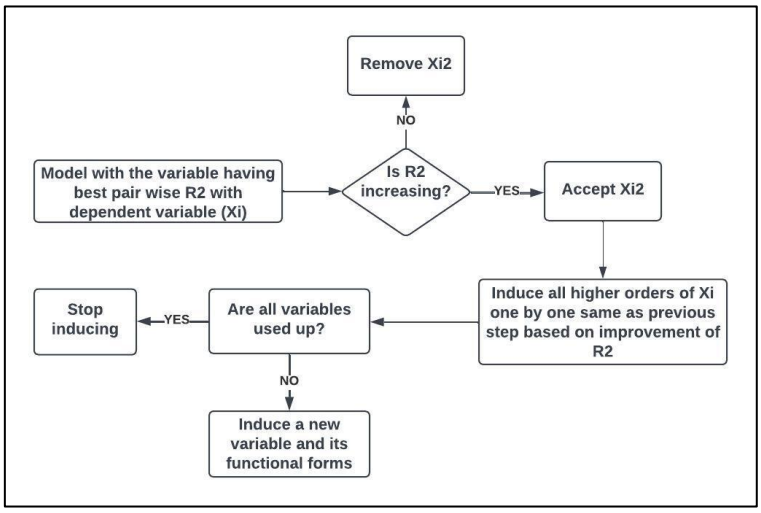

Figure 3. Steps to Conduct a Linear Regression Analysis [30]

2. Long Short-Term Memory (LSTM) Network

LSTM can overcome long-term dependencies on its input. Depending on the intricacy of the created network, LSTM networks are somewhat physiologically plausible and can learn more than 1,000 timesteps [31]. Figure 4 shows the steps to conduct LSTM network training. Sequential modelling using LSTM network requires transforming the dataset into a training and validation set. The training set is used to train the model, and the validation set is used to evaluate how accurate the model is. The ratio to divide the dataset varies. A common ratio is 70% for the training set and 30% for the validation set.

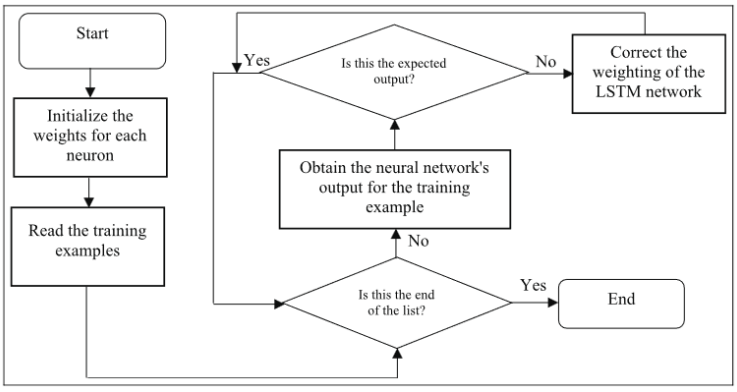

Figure 4. Steps to Conduct a LSTM Network Training [32]

Figure 5 shows the structure of LSTM network which consists of three different gates: forget, input and output gate. Forget the gate; there is a sigmoid function in this gate that is represented by σ . This gate uses the Sigmoid function to determine what data from the preceding state cell will be memorized [33]. To process the long-term sequence more effectively, the network structure may effectively forget the previous useless information, save the valid input information, and decide the necessary output information [34]. The procedure in this gate follows the equation below where W_f and b_f refers to the parameters must be decided upon following training, h_{t-1} refers to hidden layer at time epoch t−1, x_t refers to the input arrow at time epoch t, and f_t refers to the output of sigmoid function. Epoch is one pass through all samples in training dataset and updating the network weights [35]. The network will process the dataset in forget uses Equation (12).

$$f_t = \sigma (W_f[h_{t-1} x_t] + b_f) \quad (12)$$

Input gate, the gate that may select from the input which values to update the memory state. It manages the input and chooses which information will be added and saved in the current cell state. Two equations below are the procedures of this gate. Equation (13) is similar to the previous equation related to sigmoid functions. It determines what new information that must be added to the cell state h_{t-1} .

$$i_t = \sigma (W_i[h_{t-1} x_t] + b_i) \quad (13)$$

The output of Equation (13) will be saved in a cell state, the layer of tanh that produces a new layer. $\bar{C}_t$ or cell state layer that would be multiplied by I later as shown in Equation (14).

$$\bar{C}_t = \tanh (W_c[h_{i-1} x_t] + b_c) \quad (14)$$

The output gate decides what to output based on the input and the cell's memory. The output gate serves as the cell's real output limiter. Equation (15) is the procedure of processing in the output gate where it considers the weight metrics, the sigmoid activation, the current input (which is output from the input gate), and the bias of the output gate).

$$o_t = \sigma (W_o[y_{t-1} x_t] + b_o) \quad (15)$$

Then, the cell state in input gate $\bar{C}_t$ will be selected through the output gate to produce the hidden state (y_t) as in Equation (16), tanh refers to the hyperbolic tangent activation function.

$$y_t = o_t \tanh(C_t) \quad (16)$$

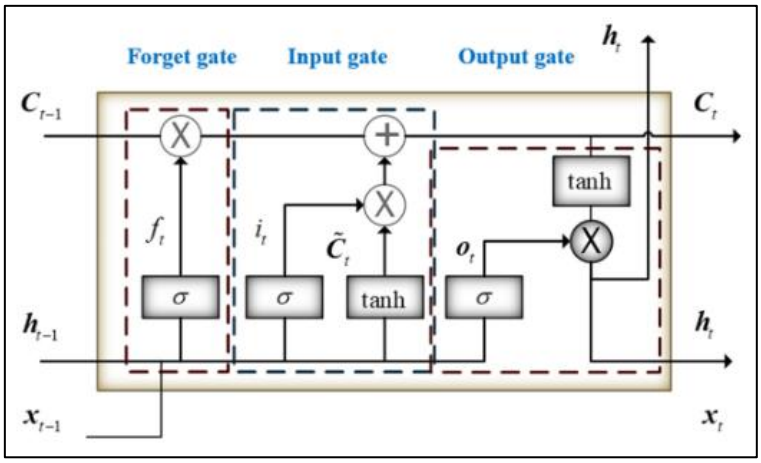

Figure 5. Structure of Long Short-Term Memory (LSTM) Network

2.2.2. Forecasting Model Validation Test

After all of the forecasting models produce the output, there should be a validation test to determine which is the best forecast to predict cocoa crop yield. The two most common metrics used to measure the accuracy of forecasting models are RMSE and R-squared. The best statistic for measuring normal errors is RMSE. RMSE typically does a better job of highlighting differences in model performance. RMSE means that the errors have a normal distribution and are unprejudiced. The fact that RMSE avoids the use of absolute value, which is highly undesirable in many mathematical calculations, is one key advantage of RMSE over MAE. It could be challenging to determine the gradient or sensitivity of the MAE regarding particular model parameters, for example [36]. It is easily differentiable and has simple computational requirements [37]. The formula of RMSE could be seen in the Equation (17) where it comes from calculation of the root square difference between predicted and actual value.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}i)^2} \quad (17)$$

While R^2 quantifies the proportion of variance that the independent variables contribute to the variance of the dependent variable. Depending on how the prediction model and the ground truth relate to one another, the

coefficient of determination can have values in the [(∞, 1] range [38]. For forecasting model, R^2 has been regarded as one of the most frequently employed and trustworthy statistical tools for evaluating the goodness of fit of a model or contrasting the effectiveness of multiple models [39]. Equation (18) is used to calculate R-squared where 1 is reduced by sum squared differences of predicted and observed value where SSE or Sum Squared Error is the difference sum square of predicted and actual value, and SST or Sum Squared Total is the difference sum square of predicted and total value of actual.

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} \tag{18}$$

Sum of Squared Errors can be calculated using Equation (19) where it is a value of error between predicted value compared to the actual value at a certain size of sample:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}i)^2 \tag{19}$$

While SST or Total Sum of Squares can be calculated through Equation (20) that refers to the error of actual value with the average of all actual value.

$$SST = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{20}$$

2.2.3. Residual Testing (IIDN)

The error or usually called residual after the model is fitted is beneficial to check whether the dataset has been sufficiently captured by a model. To ensure that the residuals are uncorrelated, have a zero mean, a constant variance, and are distributed normally, these properties are tested. Correlations between residuals indicate that there is information still present in the residuals that can be exploited to create forecasts. The forecasts are inaccurate if the residuals have a mean different from zero [40]. So that, IIDN tests that stands for Identical, Independence, and Normal Distribution need to be measured.

1. Normality Test

First, multiple linear regression needs to be analyzed. Multiple linear regression can show how rainfall, minimum A normal distribution could be seen through the Kolmogorov-Smirnov test. In contrast to regression analysis, the p-value in residual testing must reach more than the alpha of significance level. It indicates that the residuals of the dataset are independent. If the p-value is less than 0.5, it indicates that the residuals of the dataset are dependent and didn't pass the residual testing.

Other than p-value to see the normal distribution of the data, Anderson-Darling (AD) could be considered. Equation below shows the way to calculate the value of Anderson-Darling. Where F_0 refers sample parameters for estimation, $Z_{(i)}$ refers to normalized, sorted, and sample value, n refers to the sample size, In refers to the base e, and i runs from value 1 until n. Smaller value of AD means that the data or model follows a normal distribution [41].

$$AD = \sum_{i=2}^{n} \frac{1 - 2i}{n} \{In (F_0 [Z_{(i)}]) + In(1 - F_0 [Z_{(n+1-i)}]) \} -n, ... \tag{21}$$

2. Autocorrelation Test

Choosing To examine the autocorrelation of residuals, the common test is Durbin-Watson test. When the residuals from one observation are associated with the residuals from earlier observations, this is known as autocorrelation. An autocorrelation analysis can find that the output of a model is invalidated. The equation below is the formula of Durbin-Watson. For the accuracy of the model, DW value should be in range 1 to 3 to define it as the model that has no autocorrelation. When DW value is less than 1 or bigger than 3 means that the model has a problem and has autocorrelation [42]. While during the graph of autocorrelation, the graph of ACF should be stationary in the line of FAK and FAKP diagram [27].

$$DW = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2} = \frac{\sum_{i=1}^{n-1} (\Delta\varepsilon_i)^2}{\sum_{i=1}^{n} \varepsilon_i^2} = 2(1 - \rho) \tag{22}$$

Where:

$$\rho = \frac{\sum_{i=2}^{n} \varepsilon_i \varepsilon_{i-1}}{\sum_{i=1}^{n} \varepsilon_i^2} \quad (23)$$

3. Result and Discussion

3.1. Scatter Plot

A scatter plot showing the influence of each variable on cocoa crop yield is provided in this research. It is provided to see the relationship between two variables, each predictor variable and one response variable. Figure 6 is the scatter plot of yield versus clone. Seven different cocoa clones are tested in this research and coded in that plot. 0 is MCC02, 1 is SUL01, 2 is SUL02, 3 is ICCRI03, 4 is ICCRI09, 5 is KW516, and 6 is KW562. The data points represent the yield variance is included for each cocoa clone. The scatter plot shows that there isn't a discernible pattern or correlation between cocoa yield and clone. There is substantial diversity within each clone, as evidenced by the yield values, which are distributed across a broad range for each cocoa clone. This variable can be removed. Figure 7 shows the scatter plot between relative humidity and cocoa crop yield. The trend line of those data looks quite sharp, meaning that relative humidity can influence the cocoa crop yield, and higher relative humidity can produce higher crop yields. There are many points that are far from the trend line, and its spread is large, which means that the data variability is high. There are many potential causes of these problems, such as human error in measuring the data, outliers, and unpredictable reasons. This chart shows a difference from the previous research, stating that extreme relative humidity could reduce the yield of cocoa crops.

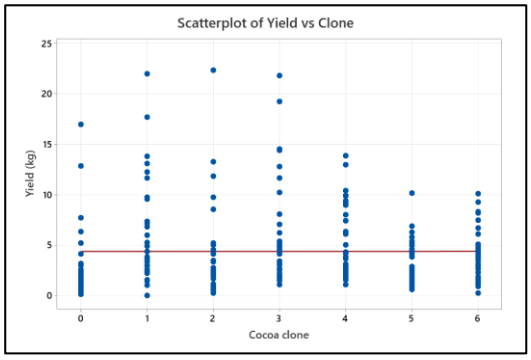

Figure 6. Scatter Plot of Yield vs Clone

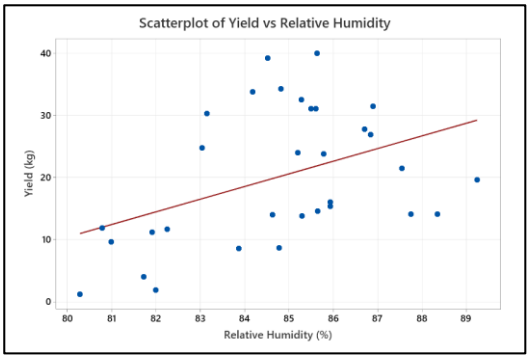

Figure 7. Scatter Plot of Yield vs Relative Humidity

Figure 8 visualizes the relationship between maximum temperature and cocoa crop yield. The plot has quite a constant line; the trend line's slope is not sharp but still shows the negative trend. Even though the relationship between those variables is weak, it means that cocoa crop yield cannot survive at a maximum temperature, as supported in the previous study that a high temperature can also indirectly result in stress of the cocoa plant [43].

Figure 9 shows the scatter plot of cocoa crop yield versus minimum temperature, showing that there is a relationship even though it is weak. The opposite of maximum temperature, minimum temperature, has a positive trend. It means that cocoa plants also cannot survive in temperatures too low. From Figure 8 and 9, it can be predicted that temperature has a weak relationship to cocoa crop yield. However, this variable still needs to be maintained in cocoa cultivation so that the farm has the stable temperature that the cocoa plant needs.

The relationship between evapotranspiration and cocoa crop yield is shown in Figure 10, which shows evapotranspiration has quite a high negative correlation with cocoa crop yield; lower evapotranspiration has a better influence on cocoa crop yield. The measurement of this variable is important to decide the water management in the farm. The only graph that shows a perfect relationship between the variables is the graph in Figure 11, which shows the correlation between rainfall and cocoa crop yield. Higher rainfall is following the increase in cocoa crop yield. This chart shows that rainfall has the most influence on cocoa crop yield. This prediction is supported by a study that rainfall is high correlated with cocoa crop yield prediction, especially four months before harvest activity [23], [44].

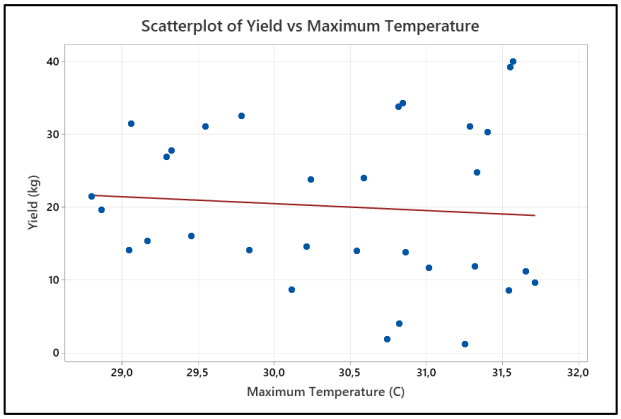

Figure 8. Scatter Plot of Yield vs Max Temp

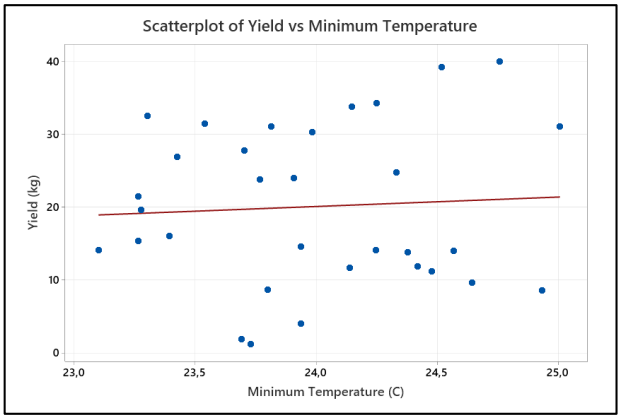

Figure 9. Scatter Plot of Yield vs Min Temp

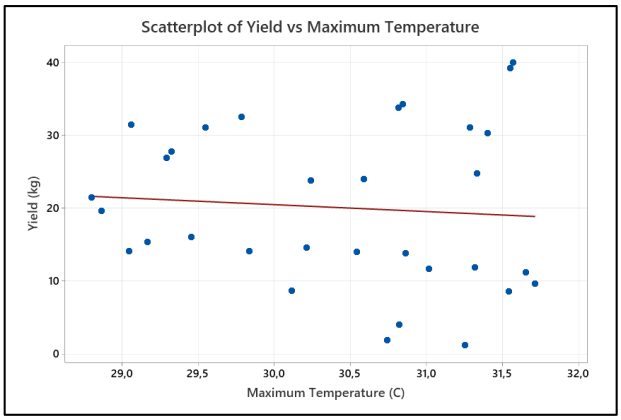

Figure 10. Scatter Plot of Yield vs et0

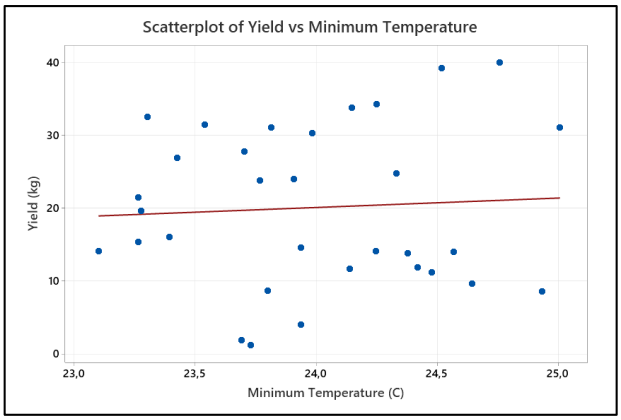

Figure 11. Scatter Plot of Yield vs Rainfall

Figure 12 is the scatter plot of cocoa crop yield versus soil moisture. It has a sharp trend line showing the positive relationship between the variables. This plot shows that higher soil moisture is one of the reasons the production of the cocoa crop yield is higher cause the soil moisture represents the drainage in soil that can prevent water logging that could affect plant cultivation [45].

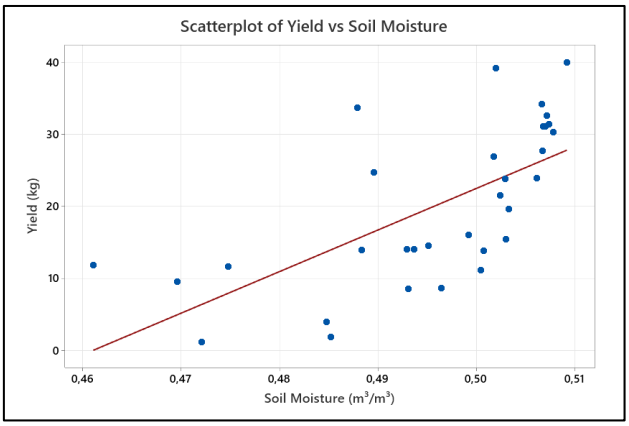

Figure 12. Scatter Plot of Yield vs Soil Moisture

Table 2 presents the summary of independent variables that could influence cacao crop yield based on the analysis that has been explained previously.

Table 2. Independent Variables

Independent Variables	Description
Relative Humidity	Higher relative humidity could produce higher cocoa crop yields. RH would impact the capacity of drying air to evaporate water from the product being dried.

Independent Variables	Description
Maximum and Minimum Temperature	Cocoa crop yield cannot survive at a maximum temperature because it can indirectly result in stress of the cocoa plant as well as the minimum temperature. Extreme temperatures have a negative impact on the subregion's cocoa production.
Evapotranspiration	Lower evapotranspiration can increase the value of cocoa crop yield. The low evapotranspiration can cause the moisture of the soil while cocoa plant has to cultivated in a good irrigation area and that system can be built better through the evapotranspiration measurement.
Rainfall	Higher rainfall is following the increase of cocoa crop yield. Rainfall has related to humidity. More high rainfall would cause high humidity that could causing fungal black pod disease, which has a high influence on cocoa production.
Soil Moisture	Higher soil moisture is one of the reasons the production of the cocoa crop yield is higher cause the soil moisture represents the drainage in a soil that can prevent water logging that could affect plant cultivation.

3.2. Forecasting Model Using Linear Regression Analysis

Seven different independence variables are tested in this research. The consideration to choose those variables are based on previous research. Even though clones have been proven to be a variable that has no influence on cocoa crop yield through scatter plots, this variable still needs to be tested for significance, including the t, F, and VIF tests. The t and F tests are carried out to clarify whether the independent variable has an influence or has no influence on the response variable. VIF is carried out to see how deep or how weak the relationship is explained by the t and F tests.

In a case study to forecast cereal yield, the variables that influence the yield are the Normalized Difference Vegetation Index (NDVI), Normalized Difference Red Edge (NDRE), rainfall, solar radiation, evapotranspiration, potential evapotranspiration, maximum temperature, minimum temperature, vapor pressure, and relative humidity [46]. In other research with study cases of soybean and corn, precipitation, solar radiation, snow water equivalent, maximum temperature, minimum temperature, vapor pressure, and soil conditions (bulk density, drained upper limit, per cent clay, wilting point, hydraulic conductivity, soil pH, per cent organic matter, saturated volumetric water, and per cent sand) are factors that influence the crop yield [47], [48].

In a study case to maize yield, genotype, yield, day length, precipitation, solar radiation, vapor pressure, maximum temperature, and minimum temperature influence crop yield. Meanwhile, rainfall and temperature are the factors that could affect wheat yield [49]. Specifically speaking about factors influencing cocoa crop yield, there are minimum temperature, maximum temperature, and rainfall [6]. In this study, the variables that can be carried out are minimum and maximum temperature, relative humidity, and rainfall, which are factors that can influence cocoa crop yield. As shown in the previous research, those variables can affect crop yield. In addition, the researcher adds clone, soil moisture, and evapotranspiration as no study discusses these factors.

Figure 13 shows ANOVA, which is one of the steps of multiple linear regression analysis. The F-value in this analysis is used to see how significantly different each independent variable is from the response variable. Equation (5) is used to calculate the F-value. Of those seven independent variables, rainfall is the most influential because the higher the f-value, the more significant that variable is. Rainfall has 5.72 as its value, while the other mostly has less than 2, i.e., soil moisture is 21, evapotranspiration is 1.69, minimum temperature 1.05, maximum temperature is 0.26, relative humidity is 0.03, and clone is 0. This statement is strengthened by the smaller p-value of rainfall that just reached 0.018. It is significantly different from other variables and makes it the only variable that could influence cocoa crop yield. Another variable has a high p-

value of more than 0.05, as the alpha of significance level is 5%, which means it is not significantly different from another variable. The rest of the variables should influence the crop yield based on some previous research, but they do not work well in this research. Many possibilities may occur in this case, as agriculture is a field or sector that has high variabilities in its processes depending on its plantation area.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	118,68	16,9549	0,97	0,450
Clone	1	0,01	0,0059	0,00	0,985
Relative Humidity	1	0,45	0,4494	0,03	0,872
Maximum Temperature	1	4,54	4,5401	0,26	0,610
Minimum Temperature	1	18,25	18,2477	1,05	0,307
Evapotranspiration	1	29,40	29,4013	1,69	0,195
Rainfall	1	99,59	99,5872	5,72	0,018
Soil Moisture	1	38,41	38,4114	2,21	0,139
Error	237	4123,08	17,3970		
Total	244	4241,77			

Figure 13. ANOVA of Multiple Regression Analysis

Based on the coefficient value in Figure 14, almost all the variables tested negatively correlate to the response variable: clone, relative humidity, maximum temperature, minimum temperature, evapotranspiration, and soil moisture. Rainfall is the only variable that has a positive relationship, and it influences cocoa crop yield. As in the ANOVA, rainfall is the most influential variable that can predict the cocoa crop yield, as the p-value is 0.018, which is less than the alpha of significance value of 5%. Even if we consider the VIF value by calculating it using Equation (6) and to interpret the VIF value is guided in Table 1, all the variables except maximum temperature do not exceed 10, indicating no multicollinearity between variables. Rainfall is the second variable with the lowest value of the VIF, 1.79, after clone, which is 1.00. This was followed by evapotranspiration at 2.92, soil moisture at 3.07, relative humidity at 8.41, minimum temperature at 8.48, and maximum temperature that reached the highest VIF value and indicated there is a multicollinearity in this variable with 15.47.

A linear regression analysis would be used to strengthen the previous statement that rainfall is the variable that most influences the cocoa crop yield. Unlike the multi-linear regression which tests more than two variables to one response variable, linear regression is conducted to see the influence of one predictor variable on one response variable. Because the rainfall would be the same for different clones in a certain period or time, the linear regression in this research would be conducted per clone to prevent the overlapping and clearly see the relationship of rainfall to yield.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	57,3	39,3	1,46	0,146	
Clone	0,002	0,133	0,02	0,985	1,00
Relative Humidity	-0,055	0,344	-0,16	0,872	8,41
Maximum Temperature	0,59	1,15	0,51	0,610	15,47
Minimum Temperature	-1,52	1,49	-1,02	0,307	8,48
Evapotranspiration	-1,79	1,38	-1,30	0,195	2,92
Rainfall	0,01142	0,00477	2,39	0,018	1,79
Soil Moisture	-51,3	34,5	-1,49	0,139	3,07

Figure 14. Coefficients of Multiple Regression Analysis

As shown in Figure 15, rainfall got a high value of F-statistics by calculating it using Equation (5), which is 1178.197. It got along with the low significance of the sum of squares. It indicates that the linear regression analysis model of rainfall is statistically significant. This means that rainfall influences cocoa crop yield. This statement is strengthened by the p-value of the coefficient of the rainfall variable that just reached 0.087 and the low VIF that was calculated using Equation (6), which is just 1.00, which can be seen in Figure 16.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50.806	1	50.806	1178.197	.000[b]
	Residual	10.479	243	.043		
	Total	61.284	244			

a. Dependent Variable: Cocoa_yield

b. Predictors: (Constant), Rainfall

Figure 15. ANOVA of Linear Regression

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3,186	0,734	4,34	0,000	
Rainfall	0,00610	0,00355	1,72	0,087	1,00

Figure 16. Coefficients of Linear Regression

The influence of the rainfall variable on cocoa crop yield can also be seen in Figure 17. Model Summary of the Rainfall Variable on Cocoa Crop Yield. It shows that the R2 is 0.829, it is calculated using Equation (18). This value means around 82.9% of cocoa crop yield data is explained by rainfall, leaving just 17.1% of the variable undefined.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.911[a]	.829	.828	.20766

a. Predictors: (Constant), Rainfall

b. Dependent Variable: Cocoa_yield

Figure 17. Model Summary of Linear Regression Model

3.3. Forecasting Model Using Long Short-Term Memory (LSTM) Network

Then, there will be a forecasting model creating using LSTM. This study will create the forecast model using LSTM, which is a type of recurrent neural network, a machine learning model. This research uses Python with the TensorFlow library as an open-source machine learning framework.

1. Training and Validation Process

After importing all libraries needed, the next step is to read the dataset. Then, the data is split into two functions: training and validation. The training took place from March 2020 until March 2022. While validation takes data from April 2022 until January 2023. Then, the data for variables that would be taken as input is extracted by the neural network through Equation (13). After reading all the datasets, they need to be adjusted to be compatible with the LSTM layer in Keras, and because of that, the training and validation datasets need to be reshaped from 2D to 3D. Then, as mentioned before, the clones need to have separate features.

After capturing all the information, it would be processed as an input and an LSTM layer. It would define the input shape of the forecast model. Meanwhile, the LSTM layer has one layer that can understand different patterns in the data, either linear or non-linear. The difference between the predicted and actual yield is measured using the loss function. Then the model's weight is updated during training using the optimizer function, which is Adam. The next step is training the data. Training data in this research consists of two arrays, which have already been reshaped and cloned, as well as validation data with the same arrays. Those data were trained 50 times, and the neural network will learn the pattern using Equation (12). It should be considered so that the model's output cannot learn the patterns in the data well enough and not overfit the training data. To see the loss in the training and validation process, a graph shows the loss in Figure 18. Even though the training

data is harder to model than validation at the first epoch, the data has successfully reached a good fit at a suitable epoch with a range of 40–50.

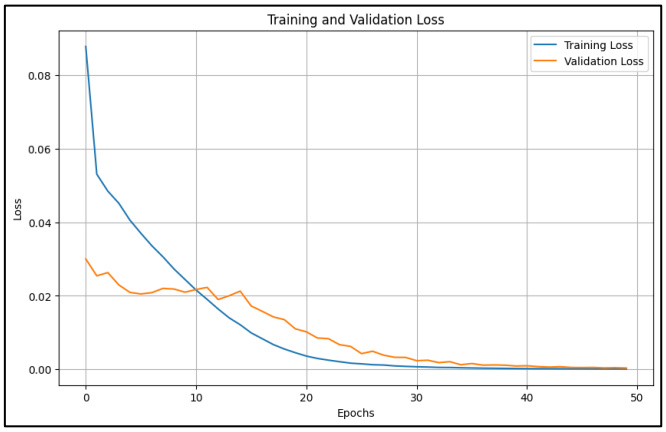

Figure 18. Training and Validation Loss

2. Analysis Forecast Result Model

The validation between the actual value of cocoa yield and the predicted value that results from the forecast model output could be seen in the following graph for April 2022 until January 2023. The graph shows that the line between the actual and predicted values is close. The distance between the lines is very close. Because clone does not influence yield, the total production in the farm could be the sum of all clones or plants.

Table 3. Actual vs Predicted Yield - Validation Test of LSTM

VIF-value	Actual Yield (kg)	Predicted Yield (kg)
April 2022	13.74	14.49
May 2022	18.20	18.91
June 2022	29.55	31.37
July 2022	53.92	53.55
August 2022	14.94	18.53
September 2022	15.23	17.41
October 2022	14.48	17.48
November 2022	24.49	26.40
December 2022	18.61	21.14
January 2023	9.11	13.12

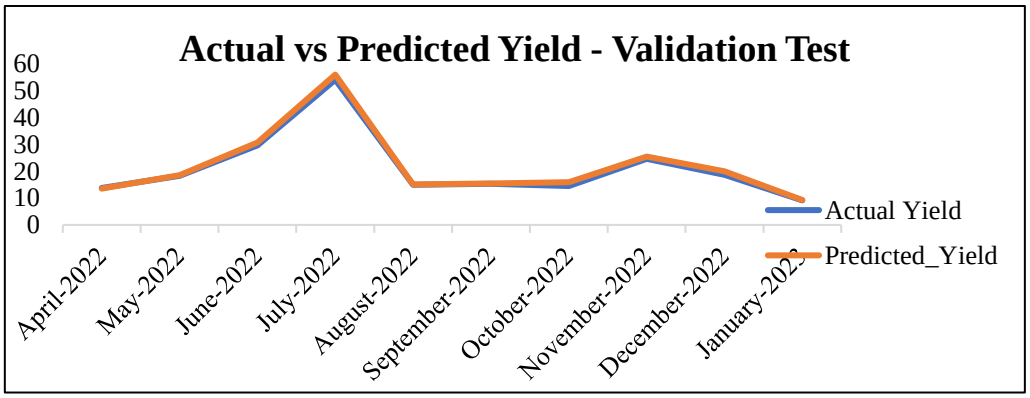

Figure 19. Actual vs Predicted Yield - Validation Test of LSTM

3.4. Performance Comparison Between the Models

After deciding which variable highly influences cocoa crop yield and creating the model analysis, the next step is comparing the models. It is done to see the best models based on some metrics: coefficient of determination and RMSE. These metrics can be calculated using Equations (18) and (17). RMSE is calculated through Equation (16), and R-squared is calculated through Equation (19). According to Table 3, the R-squared

of linear regression analysis is 82%, meaning that the forecasting result can be defined by regression analysis at 82%, and the rest, which is around 18%, is undefined. This result is lower than the forecasting result calculated by the LSTM network, which can explain 97% of the results; just 2% are undefined. At the same time, the same result is obtained with another metric, which is RMSE. LSTM architecture just produced 0.36 errors compared to regression analysis, which had 2.57 errors.

Table 4. Model Performance Metrics

Metrics	Linear Regression Analysis	LSTM Architecture
R-squared	82%	97%
RMSE	2.57	0.36

3.5. Forecasting Model Output Using Chosen Method

In the previous subsection, the reason why LSTM is the best model was already explained. The model created by the LSTM architecture is saved in a Keras-type file. TensorFlow, an open-source machine learning framework, is used to open the model. Figure 20 shows the model produced by the LSTM architecture. This final model has four different layers: Input, Long Short-Term Memory, Concatenate, and Dense layer.

There are two different layers, which are input_1 and input_2. First, input layer 1, which takes sequences of data with shape (None, 1, 7), means that ('None') refers to the batch size and single step. 7 refers to the predictor variables: rainfall, relative humidity, evapotranspiration, soil moisture, minimum temperature, maximum temperature, and clone. Because clone is a variable that does not influence the cocoa crop yield, there would be a separate column in feature 1. The output from input_1 will be transferred to the next layer, layer_2, as the final output to be transferred to the Long Short-Term Memory layer with a forget, input, and output gate.

Then, in the concatenate layer, this layer would concatenate the output from the LSTM layer and input_2 that produced output (None, 71). Then, this output is carried to the dense layer, a fully connected layer that produces a single unit or neuron. Here, this layer would predict a continuous value that would like to be predicted, in this case, cocoa crop yield. As the model is in Keras, to use this model, TensorFlow is needed.

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 1, 7)]            0         []

 lstm (LSTM)                 (None, 64)                18432     ['input_1[0][0]']

 input_2 (InputLayer)        [(None, 7)]               0         []

 concatenate (Concatenate)   (None, 71)                0         ['lstm[0][0]',
                                                                  'input_2[0][0]']

 dense (Dense)               (None, 1)                 72        ['concatenate[0][0]']

=================================================================================================
Total params: 18504 (72.28 KB)
Trainable params: 18504 (72.28 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

Figure 20. Forecasting Model of LSTM

3.6. Residual Testing

IIDN, which stands for identical, independent, and normal distribution, is a residual test for data that consists of a normality and auto-correlation testing for residual data from analysis output, not original data. It could be further explained in this section.

1. Normality Test

The hypothesis test for residual testing in terms of normality is as follows:

H_0: Data comes from normal distribution population

H_1 : Data doesn't come from normal distribution population

Figure 21 shows the normality test of LSTM residual. It shows that the data closely follow the pattern of the red line, which is the line of data values following a normal distribution. The p-value is also quite big for residual data, which is 0.104, which is bigger than the number of significance (0.05), which means do not reject the null hypothesis. It concludes that the data comes from a normal distribution population. This statement is also strengthened by the small value of AD that can be calculated using Equation (21), which is 0.617, it indicates that the data or method follows a normal distribution. In contrast with residual testing of LSTM's residual, Figure 22 shows that the p-value of the regression model is 0.005, which means it should reject the null hypothesis. The residual does not come from a normal distribution population.

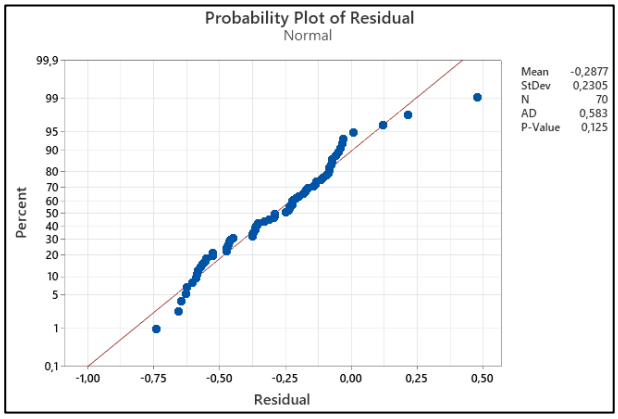

Figure 21. LSTM's Residual Probability Plot

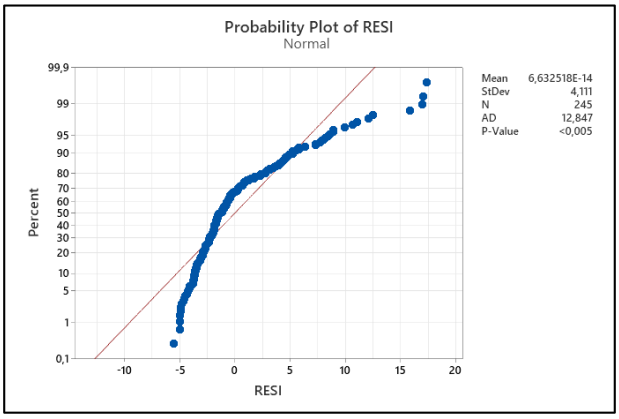

Figure 22. Linear Regression's Probability Plot

2. Autocorrelation Test

After the normality test is performed, the next step is the autocorrelation test to see the correlation between the variables and determine whether the data are independent or associated or not. It refers to the relationship between individual variables within a sequence number series. The data are considered dependent if the lag of this test exceeds the significance limits. As shown in Figure 23, which shows the autocorrelation graph of LSTM model's residual, there is no lag of data (represented by the blue line) that exceeds the red line, representing the significance limit. It means that the data are independent and not related to each other. This data passed the autocorrelation test.

As in contrast with the LSTM model's residual, Figure 24, which represents the autocorrelation residual of the regression model, shows that many blue lines exceed the red line as the significance limit. It means that the residual comes from the data that has autocorrelation.

Table 5 shows the summary of validation and verification tests among the methods: linear regression and LSTM network. It proves that LSTM passed the validation and verification as the forecast model that can predict cocoa crop yield.

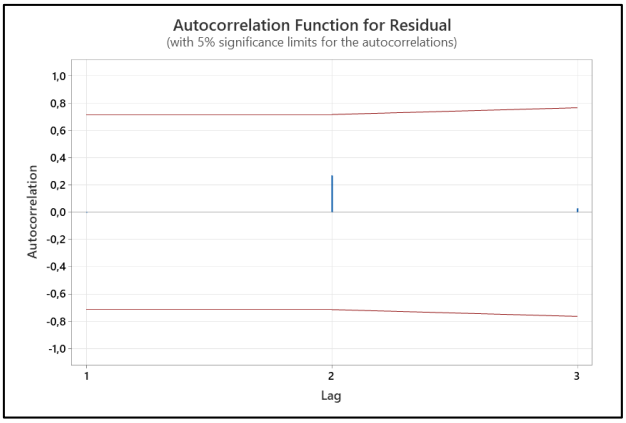

Figure 23. Autocorrelation Graph for LSTM's Residual

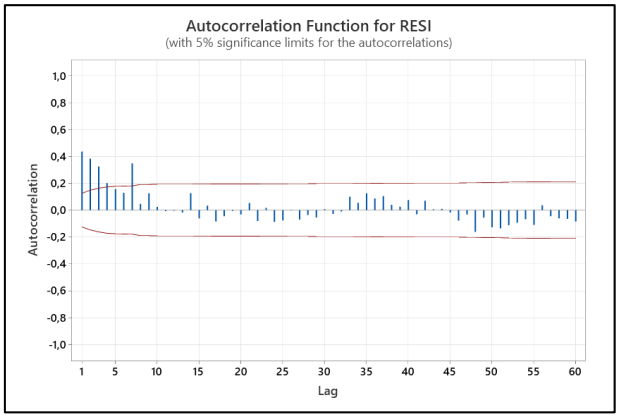

Figure 24. Autocorrelation Graph for Regression Residual

Table 5. Summary Among the Models

Method	Validation	Verification
Multiple Linear Regression	NO	NO
Long Short-Term Memory	PASSED	PASSED

From those seven different variables tested in this research—clone, relative humidity, maximum temperature, minimum temperature, evapotranspiration, rainfall, and soil moisture. Except for clones that have been proven to have no influence on cocoa crop yield, the trend and its correlation to cocoa crop yield can be found to create a forecast model using LSTM. Extremely high rainfall would indeed cause high humidity and cause fungal black pod disease, which can decrease cocoa harvesting. But, according to the data in this research, the range of the rainfall actually has a positive correlation with the yield produced. Relative humidity is proven to impact the capacity of frying air to evaporate from the plant being dried. High relative humidity can cause high evapotranspiration because when relative humidity and high evapotranspiration are high, it can attack the metabolism processes that enable cocoa pod growth, which refers to the cocoa beans before they are cultivated. While this research shows that relative humidity has a positive correlation to cocoa crop yield, evapotranspiration has a negative correlation to cocoa crop yield. Significant changes in weather conditions need to be considered here; therefore, it would be good if this were something that could be discussed in further research.

Both extreme minimum and maximum temperatures impact cocoa yield because minimum temperatures can decrease nutrient availability in the soil while maximum temperatures can cause block disease. All of these can negatively impact cocoa crop yield productivity. The soil condition, which is caused by rainfall, relative humidity, evapotranspiration, and temperature, has to have good moisture that is suitable for cocoa plants because good soil moisture can provide nutrition and water content for the plant. LSTM can capture those datasets better than regression. This model can also produce a more accurate and validated forecast result with a minimum error rather than regression.

4. Conclusion

This study has the purpose of making the best forecast model to predict cocoa crop yield using variables that have a high influence on the production of yield, considering that Indonesia is one of the largest countries that has a high production of cocoa worldwide, even though together with countries in West Africa, it is not enough to fulfil the demand for cocoa worldwide. But it needs to be emphasized again that the production of agricultural products has very high variability. Based on the data analysis and literature review of the previous chapter, it can be concluded that the conclusion is as follows:

From the seven variables tested in this research according to climate changes, genotype, and soil condition aspects, which are clone, relative humidity, maximum temperature, minimum temperature, rainfall, evapotranspiration, and soil moisture, it turns out that except clone, all of the variables tested influence different levels. Rainfall, relative humidity, evapotranspiration, and soil moisture greatly influence cocoa crop yield. Meanwhile, maximum or minimum temperature affects cocoa crop yield at a low level.

Regression analysis is insufficient to capture a dataset and make a forecasting model to predict cocoa crop yield. Based on R-squared and root mean square error (RMSE) calculations, the forecast model of the LSTM architecture gets a value of 98%, which is higher than the regression analysis. It means that the model could explain 98% of the data; just 2% is undefined. Same result as RMSE: forecast models using LSTM architecture just produce 0.3 errors compared to regression analyses that get 2.57 errors. Other than that, through the residual testing, the regression also did not pass the test, while LSTM could pass both the normality and autocorrelation tests. From those metrics, it could be concluded that a forecasting model using LSTM is fit to predict cocoa crop yield.

5. Acknowledgements

The authors would like to thanks the Department of Industrial Engineering of President University for their partial support of this research.

6. Conflict of Interest

The author declares there is no conflict of interest.

References

[1] A. Krähmer *et al.*, "Fast and neat - Determination of biochemical quality parameters in cocoa using near infrared spectroscopy," *Food Chem.*, vol. 181, no. February, pp. 152–159, 2015, doi: 10.1016/j.foodchem.2015.02.084.

[2] S. K. Dermoredjo, S. M. Pasaribu, D. H. Azahari, and E. S. Yusuf, "Indonesia's coffee and cocoa agribusiness opportunities in Regional Comprehensive Economic Partnership trade cooperation," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 892, no. 1, 2021, doi: 10.1088/1755-1315/892/1/012071.

[3] I. M. Fahmid, H. Harun, M. M. Fahmid, Saadah, and N. Busthanul, "Competitiveness, production, and productivity of cocoa in Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 157, no. 1, 2018, doi: 10.1088/1755-1315/157/1/012067.

[4] M. S. Dr. Ir. Anna A. Susanti and M. A. Rhendy Kencana Putra Widiyanto, S.Si, "Outlook Komoditas Perkebunan Kakao," *Outlook Komod. Perkeb. Kakao*, vol. 4, no. Ii, pp. 29–34, 2022.

[5] E. O. Ajayi *et al.*, "We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %," *Intech*, vol. 11, no. tourism, p. 13, 2016, [Online]. Available: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics

[6] S. S. Olofintuyi, E. A. Olajubu, and D. Olanike, "An ensemble deep learning approach for predicting cocoa yield," *Heliyon*, vol. 9, no. 4, p. e15245, 2023, doi: 10.1016/j.heliyon.2023.e15245.

[7] H. Mo, Y. Zhang, Y. Liu, and Y. Zheng, "Prediction of rice yield based on LSTM long short term memory network," *J. Phys. Conf. Ser.*, vol. 1952, no. 4, 2021, doi: 10.1088/1742-6596/1952/4/042033.

[8] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting," *Expert Syst. Appl.*, vol. 184, no. May, p. 115511, 2021, doi: 10.1016/j.eswa.2021.115511.

[9] M. Mukhlis, A. Kustiyo, and A. Suharso, "Peramalan Produksi Pertanian Menggunakan Model Long Short-Term Memory," *Bina Insa. Ict J.*, vol. 8, no. 1, p. 22, 2021, doi: 10.51211/biict.v8i1.1492.

[10] G. Niedbała, "Application of multiple linear regression for multi-criteria yield prediction of winter wheat," *J. Res. Appl. Agric. Eng.*, vol. 63, no. 4, p. 125, 2018.

[11] V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016, doi: 10.17485/ijst/2016/v9i38/91714.

[12] A. Assa, Rosniati, and M. R. Yunus, "Effects of cocoa clones and fermentation times on physical and chemical characteristics of cocoa beans (Theobroma cacao L.)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 528, no. 1, pp. 1–7, 2019, doi: 10.1088/1757-899X/528/1/012079.

[13] A. Wahyu Soesilo, I. A. Sari, and . I., "Yield Performance of Locally Selected Cocoa Clones in North Luwu," *Pelita Perkeb. (a Coffee Cocoa Res. Journal)*, vol. 31, no. 3, pp. 152–162, 2015, doi: 10.22302/iccri.jur.pelitaperkebunan.v31i3.172.

[14] G. Civeira, "Introductory Chapter: Soil Moisture. We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %," *Intech*, pp. 1–3, 2019, [Online]. Available: http://dx.doi.org/10.1039/C7RA00172J%0Ahttps://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics%0Ahttp://dx.doi.org/10.1016/j.colsurfa.2011.12.014

[15] A. Herawati, Rahayu, G. Herdiansyah, Supriyadi, and R. Wijayanti, "The impact of climate change on land suitability and cocoa productivity in Tulakan District, Pacitan Regency," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 824, no. 1, 2021, doi: 10.1088/1755-1315/824/1/012028.
[16] S. Baja, Harli, L. Asrul, R. Padjung, and R. Neswati, "The Effect of Soil Chemicals on Cocoa Productivity in West Sulawesi," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 921, no. 1, 2021, doi: 10.1088/1755-1315/921/1/012046.
[17] H. Wijayati and H. Haqqi, "The Indonesian Global Cocoa Chain's Position in the Pandemic Era," *Int. J. Soc. Sci. Econ. Art*, vol. 12, no. 1, pp. 10–21, 2022, doi: 10.35335/ijosea.v12i1.75.
[18] D. Arianto, Z. Basri, and M. Bustami, "Induksi kalus dua klon kakao (Theobroma cacao L.) unggul Sulawesi pada berbagai konsentrasi 2,4 dichlorophenoxy acetic acid secara in vitro," *e-J. Agrotekbis*, vol. 1, no. 3, pp. 211–220, 2013.
[19] A. W. Susilo, B. Setyawan, and I. A. Sari, "Yield Performance of Some Promising Cocoa Clones (Theobroma cacao L.) at Dry Climate Condition," *Pelita Perkeb. (a Coffee Cocoa Res. Journal)*, vol. 36, no. 1, pp. 24–31, 2020, doi: 10.22302/iccri.jur.pelitaperkebunan.v36i1.372.
[20] M. Sundari and P. R. Sihombing, "The Influence of Climate Factors on Cocoa Productivity in Sulawesi, 2019," *Param. J. Stat.*, vol. 1, no. 1, pp. 21–30, 2021, doi: 10.22487/27765660.2021.v1.i1.15444.
[21] W. Niether, L. Armengot, C. Andres, M. Schneider, and G. Gerold, "Shade trees and tree pruning alter throughfall and microclimate in cocoa (Theobroma cacao L.) production systems," *Ann. For. Sci.*, vol. 75, no. 2, Jun. 2018, doi: 10.1007/s13595-018-0723-9.
[22] L. S. Fraga Junior, L. M. Vellame, A. S. de Oliveira, and V. P. da Silva Paz, "Transpiration of young cocoa trees under soil water restriction," *Sci. Agric.*, vol. 78, no. 2, 2020, doi: 10.1590/1678-992x-2019-0093.
[23] F. Yoroba *et al.*, "Evaluation of Rainfall and Temperature Conditions for a Perennial Crop in Tropical Wetland: A Case Study of Cocoa in Côte d'Ivoire," *Adv. Meteorol.*, vol. 2019, 2019, doi: 10.1155/2019/9405939.
[24] B. W. Farhanandi and N. K. Indah, "Karakteristik Morfologi dan Anatomi Tanaman Kakao (Theobroma cacao L.) yang Tumbuh pada Ketinggian Berbeda," *LenteraBio Berk. Ilm. Biol.*, vol. 11, no. 2, pp. 310–325, 2022, doi: 10.26740/lenterabio.v11n2.p310-325.
[25] Z. Qiu *et al.*, "Analysis of the accuracy of using ERA5 reanalysis data for diagnosis of evaporation ducts in the East China Sea," *Front. Mar. Sci.*, vol. 9, no. January, pp. 1–14, 2023, doi: 10.3389/fmars.2022.1108600.
[26] H. Liang, "Influence mechanism of development effect of forest ecotourism in China," *For. Stud.*, vol. 64, no. 2010, pp. 93–100, 2016, doi: 10.1515/fsmu-2016-0006.
[27] G. M. Tinungki, "The analysis of partial autocorrelation function in predicting maximum wind speed," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 235, no. 1, 2019, doi: 10.1088/1755-1315/235/1/012097.
[28] H. M. Maheshwarappa and S. Majumder, "Interpretation of p-value: The Correct Way!," *Indian J. Respir. Care*, vol. 12, no. 1, pp. 1–2, 2023, doi: 10.5005/jp-journals-11010-1026.
[29] J. I. Daoud, "Multicollinearity and Regression Analysis," *J. Phys. Conf. Ser.*, vol. 949, no. 1, 2018, doi: 10.1088/1742-6596/949/1/012009.
[30] R. Mopuri, S. G. Kakarla, S. R. Mutheneni, M. R. Kadiri, and S. Kumaraswamy, "Climate based malaria forecasting system for Andhra Pradesh, India," *J. Parasit. Dis.*, vol. 44, no. 3, pp. 497–510, 2020, doi: 10.1007/s12639-020-01216-6.
[31] M. K. Wisyaldin, G. M. Luciana, and H. Pariaman, "Pendekatan LSTM untuk Memprediksi Kondisi Motor 10 kV pada PLTU Batubara," *Kilat*, vol. 9, no. 2, pp. 311–318, 2020, [Online]. Available: http://jurnal.itpln.ac.id/kilat/article/view/997%0Ahttps://jurnal.itpln.ac.id/kilat/article/download/997/775
[32] J. Hernández, D. López, and N. Vera, "Primary user characterization for cognitive radio wireless networks using long short-term memory," *Int. J. Distrib. Sens. Networks*, vol. 14, no. 11, 2018, doi: 10.1177/1550147718811828.
[33] C. Jiang *et al.*, "A mixed deep recurrent neural network for MEMS gyroscope noise suppressing," *Electron.*, vol. 8, no. 2, pp. 1–14, 2019, doi: 10.3390/electronics8020181.
[34] C. C. Liu, T. Wu, and C. He, "State of health prediction of medical lithium batteries based on multi-scale decomposition and deep learning," *Adv. Mech. Eng.*, vol. 12, no. 5, 2020, doi: 10.1177/1687814020923202.
[35] J. Brownlee, "Long Short-Term Memory Networks With Python," *Mach. Learn. Mastery With Python*, vol. 1, no. 1, p. 228, 2017.

[36] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
[37] A. Jadon, A. Patil, and S. Jadon, "A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting," 2022, [Online]. Available: http://arxiv.org/abs/2211.02989
[38] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.
[39] R. L. Sapra, "Using R2 with caution," *Curr. Med. Res. Pract.*, vol. 4, no. 3, pp. 130–134, 2014, doi: 10.1016/j.cmrp.2014.06.002.
[40] R. J. Hyndman and G. Athanasopoulos, "Athanasopoulos, George_ Hyndman, Rob J. - Forecasting_ Principles and Practice (2018).pdf".
[41] E. S. Ahmed, E. Raheem, and S. Hossain, "Absolute Penalty Estimation," *Int. Encycl. Stat. Sci.*, pp. 1–3, 2011, doi: 10.1007/978-3-642-04898-2_102.
[42] A. Field, "Discovering Statistic Using IBM SPSS Statistic 5th," *Dk*, vol. 53, no. 9, pp. 1689–1699, 2017.
[43] T. K. Igawa, P. M. de Toledo, and L. J. S. Anjos, "Climate change could reduce and spatially reconfigure cocoa cultivation in the Brazilian Amazon by 2050," *PLoS One*, vol. 17, no. 1 January, pp. 1–14, 2022, doi: 10.1371/journal.pone.0262729.
[44] E. Santosa, G. P. Sakti, M. Z. Fattah, S. Zaman, and A. Wahjar, "Cocoa Production Stability in Relation to Changing Rainfall and Temperature in East Java, Indonesia," *J. Trop. Crop Sci.*, vol. 5, no. 1, pp. 6–17, 2018, doi: 10.29244/jtcs.5.1.6-17.
[45] W. Niether *et al.*, "The effect of short-term vs. long-term soil moisture stress on the physiological response of three cocoa (Theobroma cacao L.) cultivars," *Plant Growth Regul.*, vol. 92, no. 2, pp. 295–306, 2020, doi: 10.1007/s10725-020-00638-9.
[46] J. Richetti, F. I. Diakogianis, A. Bender, A. F. Colaço, and R. A. Lawes, "A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield," *Comput. Electron. Agric.*, vol. 205, no. December 2022, 2023, doi: 10.1016/j.compag.2023.107642.
[47] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," *Front. Plant Sci.*, vol. 10, no. January, pp. 1–14, 2020, doi: 10.3389/fpls.2019.01750.
[48] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, "Corn Yield Prediction With Ensemble CNN-DNN," *Front. Plant Sci.*, vol. 12, no. August, pp. 1–13, 2021, doi: 10.3389/fpls.2021.709008.
[49] W. W. Guo and H. Xue, "Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models," *Math. Probl. Eng.*, vol. 2014, 2014, doi: 10.1155/2014/857865.