

Comparative study of extraction features and regression algorithms for predicting drought rates

Irza Hartiantio Rahmana^{1,2}, Amalia Rizki Febriyani¹, Indra Ranggadara¹, Suhendra¹, Inna Sabily Karima¹

¹Information Systems Department, Faculty of Computer Science, Mercu Buana University, Jakarta, Indonesia

²Drug and Food Data and Information Centre, Indonesian Food and Drug Authority, Jakarta, Indonesia

Article Info

Article history:

Received Mar 24, 2021

Revised Apr 06, 2022

Accepted Apr 14, 2022

Keywords:

Drought

Logistic regression

NDVI

NDWI

Random forest regression

ABSTRACT

Rice is the primary staple food source for Indonesian people, with consumption increasing so that rice production needs to be increased. Rice drought is one of the problems that can hamper rice production. This research aims to determine the best extraction feature between the normalized difference vegetation index (NDVI) and the normalized difference water index (NDWI) in describing rice fields' dryness. Moreover, using the random forest regression algorithm. This research compares NDVI with NDWI using data originating from Sentinel-2A and retrieved via the google earth engine. Regression algorithms are used in research to predict drought in paddy fields. This research shows that NDVI is better than NDWI in predicting drought using random forest regression algorithms and logistic regression algorithms. The random forest regression algorithm based on the results obtained shows that the average root mean square error (RMSE) on NDVI is 0.018, and NDWI is 0.012. Based on the logistic regression algorithm results, it was found that the average value of RMSE on NDVI was 0.346, and NDWI was 0.336. Based on the results of the RMSE, it shows that the forecasting ability of the random forest regression algorithm is better than the logistic regression.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Indra Ranggadara

Information Systems Department, Faculty of Computer Science, Mercu Buana University

Jl. Meruya Selatan, Kembangan, Jakarta Barat 11650, Indonesia

Email: indra.ranggadara@mercubuana.ac.id

1. INTRODUCTION

In Indonesia, rice drought can occur every year due to El Nino and could significantly impact the agricultural sector in several Indonesian regions [1]. To measure farmers' and communities' resilience is facing drought and identify the factors influencing it to summarize the policy implications with the various indicators produced. It is also obtained from the application of livelihoods in identifying determining factors. Strengthening farmers' resilience to drought can be strengthened by the ease of credit, easy equipment rental, and technical efficiency in rice production [2]. Drought can significantly impact crop yields when production is reduced, leading to price increases to consumers [3]. It also increases production costs which can have an impact on the economic sector [4]. Drought on agricultural land can significantly impact the economy, politics, and technology, especially in high severity that creates enormous losses [5]. In the rice-growing season period, an adequate irrigation system is required, but drought can occur at any time. Climate change is currently impacting different rainfall patterns every year, even in different regions [6].

In this research, a comparison of the moisture content in the vegetative and generative phases was carried out to be predicted in the ripening phase because the water content in each phase was different and

greatly affected rice growth and ultimately affected grain production. In the vegetative phase, growth has active tillers, a gradual increase in plant height, and leaves begin to increase periodically. Extension stems characterize the reproductive phase, decreasing the number of tillers, booting, the appearance of flag leaves, crowns, and flowering. In the reproduction phase, on average, it can be estimated at 30 days in most cultivars. The initial phase extends the internodes or the grafting stage and varies slightly according to cultivar and weather conditions. During this period, grains' size and weight will increase from starch and sugar sources released from the sheath of leaves and stems. The grain turned to gold, and the rice leaves began to age [7]. Drought conditions can decrease the quality of grain yields per clump, especially in chlorophyll content, the ratio of chlorophyll a/b, and increased proline and total sugar accumulation [8].

Remote sensing is the observation of an object using a device remotely [9]. Sentinel-2 consists of 13 spectral bands and has an orbital map with a width of 290 km. Each of the Sentinel-2 constellation satellites has a repeating cycle of 10 days, and with both satellites fully operational, a 5-day resolution can be achieved at the equator [10]. According to literature research using the Sentinel-2A vegetation value index, five classes were taken from these three main periods: land preparation, early vegetative, late vegetative, generative, and harvest/ripening [11]. Thus, to classify the cover of rice fields can use Sentinel-2 imagery [12]. Moreover, it supported using the google earth engine for image processing without using clouds. The imagery used comes from the Sentinel-2A satellite because it is more accessible using an earth engine than Google, as it costs nothing and rotates around the earth for ten days [13] so that monitoring is done faster.

The threat of drought has hit some areas in Kebumen Regency, farmers in collaboration with the district government, farmers have started looking for various alternative water sources that will eventually flow using a pumping system. So, it takes a systematic and efficient effort that will have a risk of loss due to these threats by using the normalized difference water index (NDWI) and normalized difference vegetation index (NDVI). NDVI results can describe results with specific cloud-free images that are not available when relying on sensors with a high spatial resolution for a certain period. NDWI is remote sensing, sensitive to water content changes [13]. Near-infrared (NIR) and short wave infrared (SWIR) combinations eliminate variations caused by the leaves' inner structure and the leaves' dry matter content, increasing vegetation moisture uptake accuracy [14].

In 1995, Tin Kam Ho proposed the random forest (RF) with his research entitled random decision forest [15], then in 2001, it was redeveloped by Leo Breiman, which was then patented [16]. Random forest regression algorithm is an ensemble learning that combines most regression trees. The regression tree can be represented by collecting hierarchically arranged conditions continuously from the root to the tree leaves [17]. Logistic regression algorithm (LR) is mathematical modeling with an approach that can describe several variables' relationships. So far, the logistic regression algorithm is the most widely used modeling procedure for epidemiological data analysis [18]. As a result, the random forest algorithm consists of trees that have been planted with user values. The result will be obtained from the average error in the numerical predictor results. The random forest predictor is formed by taking the generalization errors over k trees [19], while the logistic regression algorithm describes the relationship of multiple X s to a dichotomous dependent variable [18]. This research contribution compares the extraction features of NDVI and NDWI and compares random forest regression and logistic regression to predict drought in the ripening phase using the Sentinel-2 satellite.

2. PREVIOUS RESEARCH

Some of the research results can be described as follows: the application of the NDVI method using remote sensing in determining the density of vegetation is widely used as research material. This study aims to explain the phenology of rice using Sentinel 2-A imagery with the NDVI to determine the beginning and end of the rice planting period, making it easier to monitor rice field conditions to improve plant size predictions in a short time [12]. Another study that uses the NDVI method aims to estimate rice productivity based on NDVI wave characteristics and regression from NDVI and rice productivity [20]. Subsequent research aims to: i) develop a phenology-based Landsat develop a Landsat scheme based on phenology to identify paddy fields during two phenological phases (flooding/transplantation and ripening) at a regional scale; and ii) systematically evaluate the accuracy and resultant uncertainty of the Landsat-based rice field map [21].

Using Landsat 8, NDVI aims to map various irrigated crops, highly fragmented, small in size, and heterogeneous agricultural landscapes [22]. The NDVI method is also used to use the Landsat 8 time series variogram, namely operational land imager (OLI), NDVI, NIR, and red images, to model agricultural land's spatial heterogeneity at various stages of growth [23]. From related research, five research use the NDWI extraction feature. The first research used NDWI to monitor drought [24]. Another research used NDVI for mapping vegetation moisture content [25]. NDWI is also used for detecting changes in surface water [26]. Another NDWI research was used to evaluate vegetation cover types [27]. The NDWI method is also used for monitoring drought in vegetation [28]. Furthermore, the contribution of this research also uses the random forest regression algorithm to predict drought inland. NDVI and NDWI need to be compared to this

extraction feature to detect drought. This research predicts the moisture content of experiencing drought using a random forest regression [29].

3. RESEARCH METHOD

3.1. Research stages

We are estimating the productivity of the approach used to answer the research objectives. Figure 1 explains the identification of rice fields in Kebumen, Central Java. Furthermore, by using data collection from Sentinel-2A data, pre-processing was carried out starting with atmospheric correction followed by Sentinel 2 reflectance, followed by a sampling strategy on the rice field at zoning size 160×154 for sample acquisition. The following process is feature extraction using NDVI and NDWI to find out water indication and drought comparison. Then, the modeling process continues the results of feature extraction using random forest regression and logistic regression. Finally, the evaluation model uses root mean square error (RMSE) and out of bag (OOB) to see the level of accuracy, which results in prediction comparison. So that, in the end, it can show a comparison analysis.

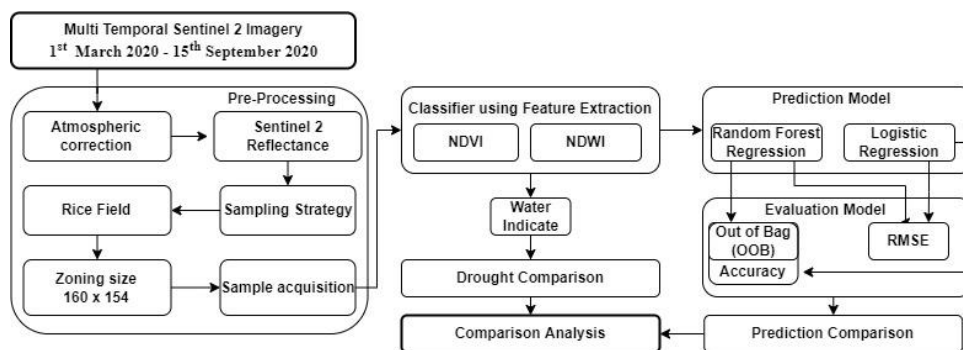


Figure 1. Research stages

3.2. Research area and data collection

The rice field area's research location is located in the Kebumen Regency, Central Java Province, the largest rice-producing area with 2174 hectares. Geographically this location or area of interest (AOI) is at coordinates 109.699456004, 109.745133512, -7.772345033, -7.728641145 [EPSG: 4326]. This research data and information were conducted with Sentinel 2 imagery from 1 March 2020 to 15 September 2020. From that period used imagery data with areas without cloud cover so that the land can be seen. In 1 period, obtained imagery is processed through the pre-processing stage that clips on land used for research sites. The data did not use in June 2020 and August 2020 because the research object is 80% covered by clouds, the research using an image that at least has a cloud tolerance of up to 10%.

Figure 2 explains the research location in Kebumen, Central Java. The characteristics of the area of Kebumen Regency can be distinguished into alluvial soil, latosol soil, podsollic soil, regosol soil, gray glei humus, and alluvial associations and the litosol and brown mediterranean associations, where the potential of the land can show that some of the areas are classified as fertile enough to be used as agricultural land. However, several sub-districts such as Sempor, Karanganyam, Sadang and Alian have soil characteristics that are less capable of being used as agricultural land [30].



Figure 2. Research area in Kebumen

3.3. Pre-processing

The pre-processing stage is the stage where data preparation is carried out before the data is processed. A raster data usually has an extensive area coverage not to reflect the area to be researched. It can describe the research area. It is essential to cut data or what is commonly known as clipping. A raster data usually has an extensive area coverage not to reflect the area to be researched. It can describe the research area. It is essential to cut data or what is commonly known as clipping [31]. The area of interest defines the clipped region, which can then be defined by points or shapes based on the coordinates. The shape of the defined area will follow the clipping procedure. The steps are carried out using the google earth engine in atmospheric correction, followed by Sentinel 2 reflectance so that a sampling strategy is obtained in the fields. After cutting, the array size at one became smaller, namely 160×54 for sample acquisition. The image is taken from medium satellite imagery because the land used at the research site contained thin clouds to see better results and minimize de-noising from an imbalance dataset. It increases accuracy by reducing errors, especially for predictive models. One challenge is developing a general auto exposure solution that includes a wide range of imaging sensors [32] with a camera's fast and powerful auto-exposure algorithm [33].

3.4. Extraction feature

3.4.1. NDVI

NDVI is a vegetation measurement that helps find vegetation density and see the level of plant health. NDVI is also used to measure the greenness of vegetation. NDVI is sensitive to photosynthetic activity by chlorophyll so the NDVI value can be used to make vegetation classifications. NDVI results are obtained from the ratio of red (RED) and NIR [34]:

$$NDVI = \frac{(Band\ 8A - Band\ 4)}{(Band\ 8A + Band\ 4)} \quad (1)$$

The (1) describes the NDVI calculated from bands 4 RED and 8 (NIR, resolution 10-m) or 8A (NIR, resolution 20-m) obtained from Sentinel-2A [35]. NDVI is also commonly used in drought monitoring, agricultural production forecasting, and fire-prone zone forecasts, as well as maps of desert attacks all over the world. The amount of historical data available can affect the forecasting results [36]. Since it is easier to adjust for changes in lighting conditions, surface slope, exposure, and other external factors, NDVI is becoming more commonly used in global vegetation monitoring.

3.4.2. NDWI

The NDWI method, which combines NIR and SWIR, is used to determine the water's condition. NDWI is used to determine water status by combining NIR and SWIR because both are located on a high reflectance and have a profound depth in the vegetation canopy [37]. NDWI can effectively improve water information in most cases. The (2) describes the RED band as band 4, the NIR band is band 8A, and the SWIR band is band 11 on Sentinel-2A [37].

$$NDWI = \frac{(Band\ 8A - Band\ 11)}{(Band\ 8A + Band\ 11)} \quad (2)$$

3.5. Modeling and evaluation prediction

The random forest regression algorithm combines many regression trees into an ensemble learning algorithm. A regression tree is a set of boundaries or conditions arranged hierarchically to be extended sequentially from tree roots to leaves [38]–[40]. The random forest is a solution to solve this problem. The random forest method is one of the methods in the decision tree. A decision tree is a flowchart shaped like a tree with a root node used to collect data, an inner node located on the root node containing questions about data, and a leaf node used to solve problems and make decisions. Which consists of various decision trees with (3) [41].

$$\{h(x, \theta_k), t = 1, 2, 3, \dots, N\} \quad (3)$$

The (3) explains that θ_k is a random variable distributed independently, x is the input variable, and N is the total of regression decision trees. The probability of generating a random forest is determined during the process extracted moment. The estimate of the total N of the unselected sample is referred to as the out-of-bag (OOB) result [41]. For regression, random forest constructs several K of regression trees and averages the results. After the K like a tree grows, the predictor of random forest regression is explained by the (4) [40].

$$\hat{f}_{rf}^K(x) = \frac{1}{K} \sum_{k=1}^K T(x) \quad (4)$$

The (4) explains that x is the input variable, T is the tree value (1,2,3,... N), and K is the total number of trees in the random forest (the size of the random forest) [40]. Furthermore, the previous stage's performance evaluation of the prediction results used the RMSE model to calculate the prediction error [42]. The RMSE has been used as a primary statistical metric to calculate model efficiency in meteorology, air quality, and climate science. Although both have been used to evaluate model efficiency for many people over the years, there is no agreement about model errors' most suitable metrics.

To make it easier, we will say we already have n sample model errors, counting e as ($e(i), i = 1,2,5 \dots, n$). Uncertainties resulting from observation errors or the methods used to compare models and observations are not considered in this research [43]. OOB is data that is not used to develop trees and represents data outside the sample used for cross-validation purposes. It will be easier to determine an indicator that indicates if the case is in the bag or OOB [44].

In this research used logistic regression (LR) algorithm is a derivative of the natural algorithm as a regression function of the predictors compared with random forest. Logistic regression is an approach to making predictive models such as linear regression, commonly referred to as ordinary least squares (OLS) regression. The difference is that researchers predict bound variables that scale dichotomy in logistic regression. With one predictor, X , this takes the form of equations [45].

$$\ln[\text{odds}(Y = 1)] = \beta_0 + \beta_1 X \quad (5)$$

The (5) explains that \ln stands for the natural algorithm, Y is the result, and $Y = 1$ when the event occurs (versus $Y = 0$ if it does not), β_0 is the intercept term and β_1 represents the regression coefficient, change in the event probability algorithm with 1 unit change in predictor X [45]. If OLS requires the condition or assumption that residual errors are distributed normally. Conversely, in this regression there is no need for these assumptions because in this type of logistic regression follows the distribution of logistics. Whereas if the dependent variable used consists of more than two categories, then the right logistic regression model is multinomial logistic regression.

4. RESULT AND DISCUSSION

Based on the visualization of NDVI shown in Figure 3 Result of visualization, the drought occurred in March 2020. According to the area of interest (AOI) related to drought in the location of Kebumen, Central Java. To make it easier to explain the results of preprocessing carried out with the NDVI and NDWI indexes, it is seen in Figure 3.

Figure 3 shows preprocessing for March, April, May, July, and September 2020, describes preprocessing by clipping according to the research location. Figure 3 is divided into two figures, namely preprocessing NDVI in figure 3(a) and NDWI in figure 3(b). This figure uses band 4, band 8A, and band 11 for the extraction feature and does not use cloud data to better value.

Figure 3(a) shows using NDVI. NDVI is divided into six class categories: non-vegetation, lowest dense, lower dense, dense, higher densities, and highest dense. Vegetation has the potential to store biomass and carbon. So the presence of vegetation can show how much carbon and biomass stocks are [46]. Staining on NDVI has a sensitivity index value that tends to be less good for detecting water content.

While Figure 3(b) shows using NDWI. NDWI uses the same categories as Figure 3(b) to obtain preprocessing results, which are compared in parallel to monitor the tested land. From the visualization, it can be seen the results of the comparison between NDVI and NDWI in Figure (4).

The results of preprocessing the vegetation index used are based on the NDVI index with a range of 0 to 1. This index describes the greenish level of a plant. The vegetation index is a mathematical combination of the red band and the NIR band as an indicator of the presence and condition of vegetation; in this case, the index range is used to determine the moisture content at the location being tested and then depicted with graphics to get the actual value in the results of data processing, seen in Figure 4.

Figure 4 describes the comparison of the NDVI and NDWI vegetation index values. The higher the water content, the closer the extraction feature value approaches 1, and vice versa: the lower the water content, the closer the feature extraction value approaches 0. It seems that NDVI is better at predicting the level of dryness in rice fields. The results showed that NDVI did best in drought compared to NDWI. According to Table 1, NDVI is divided into six class categories: non-vegetation, lowest dense, lower dense, dense, higher densities, and highest dense [46].

After get the index value, it needs to evaluate based on statistic to monitoring the drought and show in Table 2. It shows the evaluation results using RMSE to evaluate the error comparison to detect the de-noising value in the dataset used, then use an RF (OOB) and LR to see the percentage of predictions. The scaling factor cannot change the value adaptively after training, but it can learn model patterns and averages in the training set [47].

Share training data and testing data with a percentage of 80% and 20% to guide modeling to meet local optimal points better [48]. In NDVI, the average value of RF (OOB) is 0.988 with RMSE 0.018, while the average value is LR 0.952 with RMSE 0.346. In NDWI, the average value of RF (OOB) is 0.99 with RMSE 0.012, while the average value is LR 0.946 with RMSE 0.336. Based on these data results, the prediction evaluation results on NDVI are better than NDWI. From the results of the vegetation index and the algorithm that has been made, it can be seen that NDVI is better than high vegetation levels with blue coloring. Furthermore, the algorithm's results indicate that the RF and LR algorithms' average values will be higher with a high index. The RMSE value for NDVI is 0.018, indicating that NDVI is better in terms of evaluation than NDWI, which has an RMSE value of 0.012.

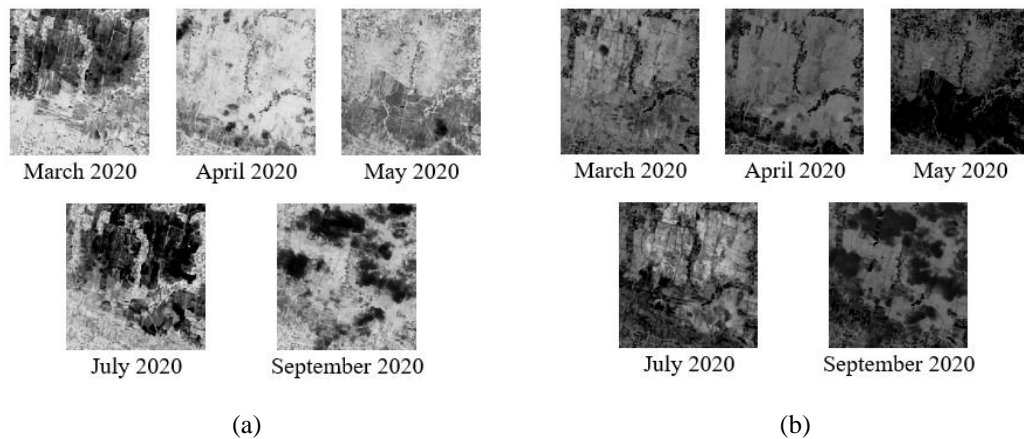


Figure 3. Preprocessing for March, April, May, July, and September 2020: (a) NDVI and (b) NDWI

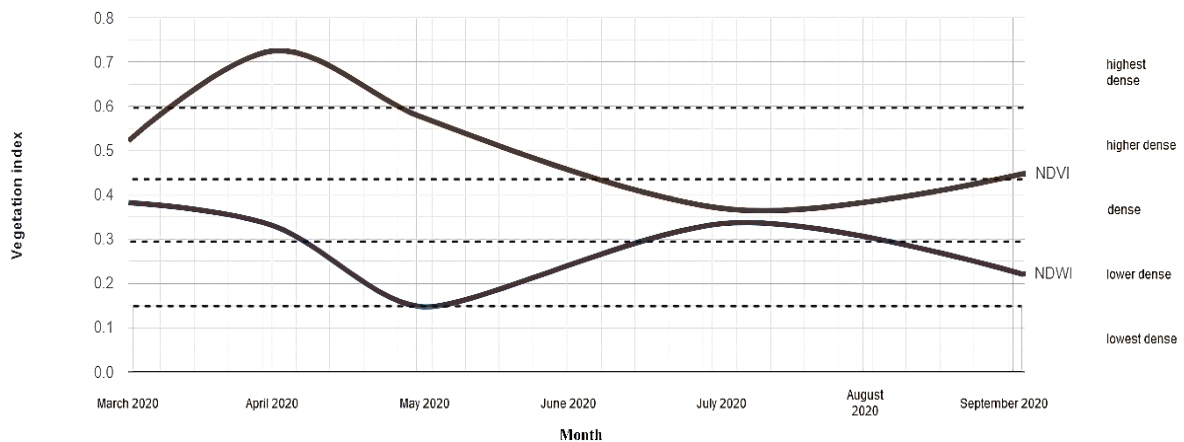


Figure 4. Comparison vegetation index value

In Figure 4, NDVI from March 2020 to September 2020 experienced a decrease in the vegetation index level, while NDWI from March 2020 to September 2020 also decreased, only experiencing a slight increase in July 2020. Furthermore, to clarify the level of vegetation is explained in Table 1.

Table 1. The index value of vegetation

No.	Dense class	NDVI	Hex code
1	Non vegetation	< 0	#ffffff
2	Lowest dense	0-0.15	#d1e3f3
3	Lower dense	0.15-0.3	#9ac8e1
4	Dense	0.3-0.45	#529dcc
5	Higher dense	0.45-0.6	#1c6cb1
6	Highest dense	> 0.6	#08306b

Table 2. Evaluation prediction of NDVI and NDWI

Months	NDVI				NDWI			
	RMSE	RF(OOB)	RMSE	LR	RMSE	RF(OOB)	RMSE	LR
March	0.02	0.99	0.35	0.96	0.01	0.98	0.39	0.96
April	0.01	0.99	0.39	0.93	0.01	0.99	0.43	0.92
May	0.01	0.98	0.30	0.92	0.01	0.99	0.13	0.95
July	0.02	0.99	0.30	0.95	0.02	0.99	0.43	0.94
September	0.03	0.99	0.39	0.95	0.01	1	0.30	0.96
Average	0.018	0.988	0.346	0.952	0.012	0.99	0.336	0.946

5. CONCLUSION

In this research, it can be concluded that the NDVI extraction feature is better than the NDWI extraction feature in predicting drought. Drought prediction is carried out by implementing the feature extraction value on the Sentinel-2 satellite image data. The data that has been feature extracted is then processed using the random forest regression algorithm and logistic regression algorithm to predict the drought of rice fields. Furthermore, the data was tested using RMSE, RF(OOB), and LR accuracy. The results obtained by NDVI have an average RF value (OOB) of 0.988 with an RMSE of 0.018, while the average value of LR is 0.952 with an RMSE of 0.346, while the NDWI average value of RF (OOB) is 0.99 with an RMSE of 0.012, while the average value of LR is 0.99, 0.946 with RMSE 0.336. Based on these data results, the evaluation of NDVI is better than NDWI. For further research, it is necessary to compare with other extraction features such as enhanced vegetation index (EVI), NDMI, soil adjusted vegetation index (SAVI), and other extraction features that are related to the level of the greenness of vegetation and to strengthen the prediction results, and further prediction evaluation is needed, using explained variance score (EVS), R squared (R^2), mean squared error (MSE), and mean absolute error (MAE).

REFERENCES

- [1] E. Surmaini, T. W. Hadi, K. Subagyono, and N. T. Puspito, "Early detection of drought impact on rice paddies in Indonesia by means of Niño 3.4 index," *Theoretical and Applied Climatology*, vol. 121, no. 3-4, pp. 669-684, Sep. 2015, doi: 10.1007/s00704-014-1258-0.
- [2] A. Keil, M. Zeller, A. Wida, B. Sanim, and R. Birner, "What determines farmers' resilience towards ENSO-related drought? An empirical assessment in Central Sulawesi, Indonesia," *Climatic Change*, vol. 86, no. 3, pp. 291-307, 2008, doi: 10.1007/s10584-007-9326-4.
- [3] R. Lal, J. A. Delgado, J. Gulliford, D. Nielsen, C. W. Rice, and R. S. V. Pelt, "Adapting agriculture to drought and extreme events," *Journal Soil Water Conservation*, vol. 67, no. 6, pp. 162-166, Nov. 2012, doi: 10.2489/jswc.67.6.162A.
- [4] P. H. Gleick, *Impacts of California's Ongoing Drought: Hydroelectricity Generation*, Oakland, California: Pacific Institute, pp. 1-14, 2016. [Online]. Available: <https://pacinst.org/wp-content/uploads/2016/02/Impacts-Californias-Ongoing-Drought-Hydroelectricity-Generation-2015-Update.pdf>
- [5] D. M. Liverman, "Drought Impacts in Mexico: Climate, Agriculture, Technology, and Land Tenure in Sonora and Puebla," *Annals of the Association of American Geographers*, vol. 80, no. 1, pp. 49-72, 1990, doi: 10.1111/j.1467-8306.1990.tb00003.x.
- [6] X. Ding, X. Li, and L. Xiong, "Insight into differential responses of upland and paddy rice to drought stress by comparative expression profiling analysis," *International Journal of Molecular Science*, vol. 14, no. 3, pp. 5214-5238, 2013, doi: 10.3390/ijms14035214.
- [7] J. T. Hardke, "Rice Growth and Development," in *Rice Production Handbook*, University of Arkansas Division of Agriculture, vol. 66, pp. 9-20, 2013, [Online]. Available: <https://www.uaex.uada.edu/publications/pdf/mp192/mp192.pdf>
- [8] Maisura, M. A. Chozin, I. Lubis, A. Junaedi, and H. Ehara, "Some physiological character responses of rice under drought conditions in a paddy system," *Journal of ISSAAS (International Society for Southeast Asian Agricultural Sciences)*, vol. 20, no. 1, pp. 104-114, 2014, [Online]. Available: [https://repository.unimal.ac.id/1285/1/ISSAAS%20Journal%20Vol%20%20%20%282014%292.pdf](https://repository.unimal.ac.id/1285/1/ISSAAS%20Journal%20Vol%20%20%20%20%282014%292.pdf)
- [9] S. K. Lin, "Introduction to Remote Sensing. Fifth Edition. By James B. Campbell and Randolph H. Wynne, The Guilford Press, 2011; 662 pages. Price: £80.75, ISBN 978-1-60918-176-5," *Remote Sensing*, vol. 5, no. 1, pp. 282-283, Jan. 2013, doi: 10.3390/rs5010282.
- [10] L. R. Mansaray, F. Wang, J. Huang, L. Yang, and A. S. Kanu, "Accuracies of support vector machine and random forest in rice mapping with Sentinel-1A, Landsat-8 and Sentinel-2A datasets," *Geocarto International*, vol. 35, no. 10, pp. 1088-1108, 2020, doi: 10.1080/10106049.2019.1568586.
- [11] Supriatna, Rokhmatuloh, A. Wibowo, and I. P. A. Shidiq, "Rice productivity estimation by Sentinel-2A imagery in Karawang Regency, West Java, Indonesia," *International Journal of GEOMATE*, vol. 19, no. 72, pp. 49-53, 2020, doi: 10.21660/2020.72.5622.
- [12] E. A. P. Lestari, Supriatna, and A. Damayanti, "Model of paddy rice phenology using Sentinel 2-A imagery with NDVI algorithm in Subang Regency," *IOP Conference Series: Earth and Environmental Science*, vol. 481, 2020, doi: 10.1088/1755-1315/481/1/012069.
- [13] B. -C. Gao, "NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sensing of Environment*, vol. 58, no. 3, pp. 257-266, 1996, doi: 10.1016/S0034-4257(96)00067-3.
- [14] P. Ceccato, S. Flasse, S. Tarantola, S. Jacquemoud, and J. M. Grégoire, "Detecting vegetation leaf water content using reflectance in the optical domain," *Remote Sensing of Environment*, vol. 77, no. 1, pp. 22-33, Jul. 2001, doi: 10.1016/S0034-4257(01)00191-2.
- [15] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.
- [16] L. Breiman, "Random forests," *Machine Learning*, pp. 5-32, 2001. [Online]. Available: <https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf>




- [17] L. Wang, X. Zhou, X. Zhu, Z. Dong, and W. Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data," *The Crop Journal*, vol. 4, no. 3, pp. 212-219, Jun. 2016, doi: 10.1016/j.cj.2016.01.008.
- [18] D. G. Kleinbaum and M. Klein, "Modeling Strategy Guidelines," *Logistic Regression*, pp. 165-202, 2010, doi: 10.1007/978-1-4419-1742-3_6.
- [19] B. Singh, P. Sihag, and K. Singh, "Modelling of impact of water quality on infiltration rate of soil by random forest regression," *Modeling Earth Systems and Environment*, vol. 3, pp. 999-1004, 2017, doi: 10.1007/s40808-017-0347-3.
- [20] Liyantono, Y. Almadani, Y. Adillah, M. M. Yusuf, M. N. R. Mahbub, and A. Fatikhunnada, "Analysis of Paddy Productivity Using NDVI and K-means Clustering in Cibusah Jaya, Bekasi Regency," *IOP Conference Series: Materials Science and Engineering*, vol. 557, 2019, doi: 10.1088/1757-899X/557/1/012085.
- [21] C. Jin, X. Xiao, J. Dong, Y. Qin, and Z. Wang, "Mapping paddy rice distribution using multi-temporal Landsat imagery in the Sanjiang Plain, northeast China," *Frontiers of Earth Science*, vol. 10, pp. 49-62, 2016, doi: 10.1007/s11707-015-0518-3.
- [22] O. J. Eddine *et al.*, "Crop type mapping from pansharpened Landsat 8 NDVI data: A case of a highly fragmented and intensive agricultural system," *Remote Sensing Applications: Society and Environment*, vol. 11, pp. 94-103, 2018, doi: 10.1016/j.rsase.2018.05.002.
- [23] Y. Ding, K. Zhao, X. Zheng, and T. Jiang, "Temporal dynamics of spatial heterogeneity over cropland quantified by time-series NDVI, near infrared and red reflectance of Landsat 8 OLI imagery," *International Journal of Applied Earth Observations and Geoinformation*, vol. 30, pp. 139-145, 2014, doi: 10.1016/j.jag.2014.01.009.
- [24] L. F. Amalo, U. Ma'Rufah, and P. A. Permatasari, "Monitoring 2015 drought in West Java using Normalized Difference Water Index (NDWI)," *IOP Conference Series: Earth and Environmental Science*, vol. 149, 2018, doi: 10.1088/1755-1315/149/1/012007.
- [25] T. J. Jackson *et al.*, "Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans," *Remote Sensing of Environment*, vol. 92, no. 4, pp. 475-482, Sep. 2004, doi: 10.1016/j.rse.2003.10.021.
- [26] M. I. Ali, G. D. Dirawan, A. H. Hasim, and M. R. Abidin, "Detection of changes in surface water bodies urban area with NDWI and MNDWI methods," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 3, pp. 946-951, 2019, doi: 10.18517/ijaseit.9.3.8692.
- [27] V. S. da Silva, G. Salami, M. I. O. da Silva, E. A. Silva, J. J. M. Junior, and E. Alba, "Methodological evaluation of vegetation indexes in land use and land cover (LULC) classification," *Geology, Ecology, and Landscapes*, vol. 4, no. 2, pp. 159-169, 2020, doi: 10.1080/24749508.2019.1608409.
- [28] Y. Gu, E. Hunt, B. Wardlow, J. B. Basara, J. F. Brown, and J. P. Verdin, "Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data," *Geophysical Research Letters*, vol. 35, no. 22, pp. 1-5, 2008, doi: 10.1029/2008GL035772.
- [29] U. Saeed, J. Dempewolf, I. B-Reshef, A. Khan, A. Ahmad, and S. A. Wajid, "Forecasting wheat yield from weather data and modis ndvi using random forests for punjab province, Pakistan," *International Journal of Remote Sensing*, vol. 38, no. 17, pp. 4831-4854, 2017, doi: 10.1080/01431161.2017.1323282.
- [30] Kebumen District Government, "A technocratic draft of the Kebumen district mid-term development plan for 2021-2025," *Kebumen District Government 2020*, pp. 1-495, 2020, [Online]. Available: https://bappeda.kebumenkab.go.id/index.php/web/view_file/196
- [31] M. Wegmann, J. Schwalb-Willman, and S. Dech, *An Introduction to Spatial Data Analysis: Remote Sensing and GIS with Open Source Software*, UK: Pelagic Publishing, 2020.
- [32] Y. Su, J. Y. Lin, and C. -C. J. Kuo, "A model-based approach to camera's auto exposure control," *Journal of Visual Communication and Image Representation*, vol. 36, pp. 122-129, Apr. 2016, doi: 10.1016/j.jvcir.2016.01.011.
- [33] Y. Su and C. -C. J. Kuo, "Fast and robust camera's auto exposure control using convex or concave model," *2015 IEEE International Conference on Consumer Electronics (ICCE)*, 2015, pp. 13-14, doi: 10.1109/ICCE.2015.7066300.
- [34] USGS, "Landsat Normalized Difference Vegetation Index," *United States Geological Survey (USGS)*, 2020. [Online]. Available: https://www.usgs.gov/core-science-systems/nli/landsat/landsat-normalized-difference-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con (accessed Apr. 21, 2021).
- [35] M. Lange, B. Dechant, C. Rebmann, M. Vohland, M. Cuntz, and D. Doktor, "Validating MODIS and sentinel-2 NDVI products at a temperate deciduous forest site using two independent ground-based sensors," *Sensors*, vol. 17, no. 8, 2017, doi: 10.3390/s17081855.
- [36] Y. S. Triana and A. Retnowardhani, "Enhance interval width of crime forecasting with ARIMA model-fuzzy alpha cut," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 3, pp. 1193-1201, 2019, doi: 10.12928/TELKOMNIKA.v17i3.12233.
- [37] T. Zhang, J. Su, C. Liu, W. -H. Chen, H. Liu, and G. Liu, "Band selection in sentinel-2 satellite for agriculture applications," *2017 23rd International Conference on Automation and Computing (ICAC)*, 2017, pp. 1-6, doi: 10.23919/ICoNAC.2017.8081990.
- [38] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, New York, Washington D.C.: Taylor & Francis Group, 1984, doi: 10.1201/9781315139470.
- [39] J. R. Quinlan, "C4.5 Programs for Machine Learning," *Elsevier Inc.*, pp. 1-302, 1993, doi: 10.1016/C2009-0-27846-9.
- [40] V. R-Galiano, M. P. Mendes, M. J. G-Soldado, M. C-Olmo, and L. Ribeiro, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain)," *Science of The Total Environment*, vol. 476-477, pp. 189-206, 2014, doi: 10.1016/j.scitotenv.2014.01.001.
- [41] D. Liu *et al.*, "Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm," *Journal of Cleaner Production*, vol. 250, 2020, doi: 10.1016/j.jclepro.2019.119468.
- [42] I. Nurhaida *et al.*, "Implementation of Deep Learning Predictor (LSTM) Algorithm for Human Mobility Prediction," *International Journal of Interactive Mobile Technologies*, vol. 14, no. 18, pp. 132-144, 2020, doi: 10.3991/ijim.v14i18.16867.
- [43] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," *Geoscientific Model Development Discussions*, vol. 7, no. 1, pp. 1525-1534, 2014, doi: 10.5194/gmdd-7-1525-2014.
- [44] H. Ishwaran and M. Lu, "Random Survival Forests," *Wiley StatsRef Statistics Reference Online*, pp. 1-13, 2019, doi: 10.1002/9781118445112.stat08188.
- [45] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395-2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.
- [46] A. Zaitunah, S. Samsuri, A. G. Ahmad, and R. A. Safitri, "Normalized difference vegetation index (ndvi) analysis for land cover types using landsat 8 oli in besitang watershed, Indonesia," *IOP Conference Series: Earth and Environmental Science*, vol. 126, 2018, doi: 10.1088/1755-1315/126/1/012112.
- [47] Y. Su and C. C. J. Kuo, "On extended long short-term memory and dependent bidirectional recurrent neural network,"

Neurocomputing, vol. 356, pp. 151-161, Sep. 2019, doi: 10.1016/j.neucom.2019.04.044.




- [48] Y. Su, K. Fan, N. Bach, C. -C. J. Kuo and F. Huang, "Unsupervised Multi-Modal Neural Machine Translation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10474-10483, doi: 10.1109/CVPR.2019.01073.

BIOGRAPHIES OF AUTHORS






Irza Hartiantio Rahmana    He got a bachelor's degree in Information Systems in 2021. He works as an employee at the Drug and Food Data and Information Center, Indonesian Food and Drug Authority, and focuses his research on Business Intelligence, Artificial Intelligence, Geographic Information Systems, Enterprise Architecture, IT Governance and Assessment, Cybersecurity, Business Process Modeling, and Programming. He can be contacted at email: irza.rahmana@pom.go.id.






Amalia Rizki Febriani    a bachelor's degree in Information Systems in 2021. She works as an employee at the PT Sigma Cipta Caraka (Telkomsigma), and focuses his research on Information Systems and Software Development. She can be contacted at email: 41817110185@student.mercubuana.ac.id.






Indra Ranggadara    He got a bachelor's degree in Informatics Engineering in 2014 and a master's degree in Industrial Engineering in 2015, and he got another master's degree in Master Information Systems in 2020. He works as a lecturer in the computer science department, and his research focuses on Business Intelligence, Artificial Intelligence, Geographic Information Systems, E-Commerce, Enterprise architecture, IT Governance and Valuation, Business Process Modelling, and Programming. He can be contacted at email: indra.ranggadara@mercubuana.ac.id.



Suhendra    received a bachelor's degree in information technology from Mercu Buana University, Jakarta, in 2011. And the master's degree in Computer Science Budi Luhur University, Jakarta, in 2015. He is currently a lecturer in Mercu Buana University Department of Computer Science. He is also a practitioner in information technology, and his current research interest includes data mining, e-business, and information system. He can be contacted at email: suhendra.mercu@mercubuana.ac.id.



Inna Sabily Karima    she got a bachelor's degree in information technology from State Islamic University Syarif Hidayatullah Jakarta, Jakarta, in 2010. Moreover, she got a master's degree in Computer Science at IPB University, Bogor, in 2014. She is currently a lecturer in Mercu Buana University Department of Computer Science. Her current research interest includes e-business, software Engineering, and information system. She can be contacted at email: inna.sabily@mercubuana.ac.id.