Indonesian Journal of Data and Science

Volume 6 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v6i2.252

Research Article

Performance Analysis of Random Forest and Naive Bayes Methods for Classifying Tomato Leaf Disease Datasets

Rima Ananda Nasution [1,*], Lilis Nur Hayati [2], Irawati [3]

[1] Universitas Muslim Indonesia, Makassar, Indonesia, 13020210238@umi.ac.id
[2] Universitas Muslim Indonesia, Makassar, Indonesia, lilis.nurhayati@umi.ac.id
[3] Universitas Muslim Indonesia, Makassar, Indonesia, irawati@umi.ac.id

Correspondence should be addressed to Rima Ananda Nasution; 13020210238@umi.ac.id

Received 05 April 2025; Accepted 28 June 2025; Published 31 July 2025

© Authors 2025. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

Tomato productivity is often disrupted by diseases affecting tomato plants, such as early blight and late blight, which can significantly reduce crop yields. Early detection of these diseases is crucial to prevent greater losses. This study compares two machine learning-based classification methods, namely Random Forest and Naïve Bayes, in identifying diseases on tomato leaves. The dataset used consists of 1,255 images obtained from Kaggle, with the data divided into two classes: early blight with 627 images and late blight with 628 images, which were then subjected to preprocessing and data splitting with three ratio scenarios (70:30, 80:20, and 90:10) for training and testing. This study shows that it only achieved an accuracy of 76.98%, while the Random Forest method had the highest accuracy of 92.86% in the 90:10 data ratio scenario. Thus, the Random Forest method has proven to be more effective in classifying tomato leaf diseases compared to Naïve Bayes. The implementation of this model can help farmers detect diseases more quickly and accurately, thereby increasing agricultural productivity.

Keywords: Random Forest, Naïve Bayes, Tomato Leaf Disease.
Dataset link: https://kaggle.com/datasets/muhammadmasdar/tomato-disease-ready

1. Introduction

Tomato (Solanum lycopersicum) is a widely cultivated plant. Tomatoes are highly favored by the Indonesian people, and the demand for tomatoes continues to increase every year. However, tomato production is often disrupted by various diseases that can cause significant losses and reduce the quality of tomatoes [1]. The health of tomato plant leaves plays a crucial role in determining the quality of the harvest, so monitoring and early detection of leaf diseases become very important steps in tomato cultivation. Leaves play an important role in supporting plant growth, so disturbances in this part can directly impact the overall development of the plant. If the leaves are affected by disease, it can worsen the condition of the plant and reduce harvest productivity. Therefore, the identification of the types of diseases affecting tomato plants is essential so that farmers can take appropriate and effective management steps [2]. The current development of technology has the potential to assist farmers in reducing misidentification of diseases in tomato plants thru the application of artificial intelligence (AI) approaches [3], [4]. With the identification produced, farmers can control disease attacks on tomato leaves and the goals of tomato cultivation can be achieved. Diseases on tomato leaves have been detected and classified into two classes, namely Early Blight and Late Blight.

This research aims to improve the accuracy in classifying diseases on tomato leaves by combining the Naive Bayes and Random Forest algorithm models [5], [6], [7], [8]. By comparing two different classification methods, it is hoped that this research can make a significant contribution to early diagnosis and management of diseases on tomato leaves. The increase in accuracy achieved can serve as a foundation for the development of a more efficient and reliable

classification system in tomato cultivation practices. The approach of comparing two classification algorithms, such as Naive Bayes and Random Forest, is still rarely used in the classification of tomato leaf diseases [9], [10]. Most studies only use one algorithm, making it less optimal in handling data complexity. The combination of algorithms has the potential to improve the accuracy and reliability of classification by comparing how well both approaches work in categorizing tomato plant leaf diseases.

2. Method:

At this stage, a systematic and structured approach is undertaken, as illustrated in **Figure 1**.

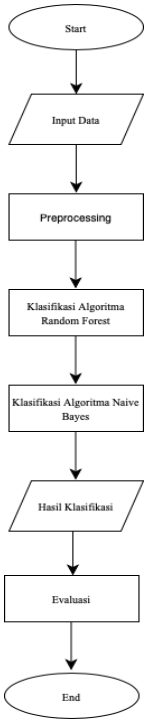


Figure 1. Research Design

Input Data

The research data was taken from the website www.kaggle.com. This dataset contains images in .jpg, .jpeg, .png, .bmp, .gif, and .tiff formats, with an RGB color scheme, totaling 1255 images, and a resolution of 255 x 255 pixels. This research focuses on the target class of tomato leaf diseases, with the data divided into two classes: late blight with 628 images and early blight with 627 images.

Pre-processing

Data preprocessing is the initial stage before conducting data mining, aimed at addressing potential issues during data processing due to inconsistencies in data format [11], [12]. The stages of data preprocessing in this study include data cleaning, data reduction, and normalization to prepare the data before further analysis. [13] In this study, the data preprocessing stage begins with the grouping of the dataset, which consists of a "train" folder for training the model and a "validation" folder for testing the model. Next, the dataset is cleaned by checking all the files in the folder to ensure that only valid image files will be processed. The accepted image extensions include .jpg, .jpeg, .png, .bmp, .gif, and .tiff. If files with other extensions are found, they will be considered irrelevant and removed from the dataset.

The next step is to resize the images to a specific target, which is 255x255 pixels. After that, normalization is performed by dividing the pixel values by 255, so that the pixel range originally from 0-255 can be normalized to 0-1. This stage aims to reduce the burden during the classification process.

Classification

Random Forest

Random Forest is a method commonly used for various purposes, such as regression, classification, and others. This method uses many decision trees. Random Forest combines the results from each decision tree that has been created during training to generate class predictions in classification problems, or thru averaging in regression models [14], [15]. This algorithm serves to address the overfitting problem that often arises during the training process, and this is also one of the main challenges in using Random Forest [16], [17]. There are several advantages of Random Forest, including its ability to improve accuracy even when there is missing data, its resistance to outliers, and its efficient use of data storage space. In addition, the feature selection mechanism by choosing the best features can improve the performance of the classification model. As a method that uses mechanisms like this, Random Forest easily operates on big data with complex parameters [18], [19].

Naïve Bayes

The Naïve Bayes algorithm is based on the theorem of probability according to Bayes. This method was designed by Thomas Bayes and aims to utilize historical data to estimate the likelihood of an event occurring in the future [20], [21]. This algorithm is one of the simpler forms of Bayesian classification, and its principle of operation uses Bayes' Theorem to calculate the probability of a set of events. Bayes' Theorem will explain how the probability of an event is evaluated [22], [23]. This method is generally used in the data classification process using the Bayes' Theorem equation for its calculations. Below is the formula for the calculation of Bayes' Theorem, which can be seen in Equation 1.

$$P(Y|X) = \frac{P(X|Y)xP(Y)}{P(X)} \quad (1)$$

Evaluation

This evaluation is to measure the performance of the algorithm used. The testing was conducted using a method to assess the extent to which the system successfully diagnoses tomato plant leaf diseases, known as the confusion matrix. The table in the confusion matrix serves to show the test data that is accurately classified and those that experience classification errors [17]. Confusion Matrix is a good way to evaluate performance in machine learning classification models. This matrix compares the model's prediction results with the actual values in the form of a matrix, which consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [24], [25].

In the classification process, there are several test matrices used to measure the model's performance, including accuracy, precision, recall, and f1-score. [5], [26] Accuracy measures the proportion of correct predictions against all the data analyzed. Recall is used to measure the ratio of the number of true positive predictions to the total actual positive data. Meanwhile, precision indicates how accurate those positive predictions are, by comparing the number of true predictions to all the data predicted as positive.

3. Results and Discussion

Results

This research uses 360 validated images, with 180 images of late blight and 180 images of early blight in separate folders. This research uses three data splitting ratios, namely 70:30, 80:20, and 90:10, where the larger percentage is used for training and the remainder for testing. Classification of tomato leaf diseases was performed using the Random Forest and Naive Bayes methods.

Random Forest

The Random Forest model was trained to classify two tomato leaf diseases, namely early blight and late blight, by examining its accuracy. In addition to accuracy, a confusion matrix is used to see the modeling results, and the results are then summarized or made into a classification report, which will produce several parameters and results, including precision, recall, f1-score, support, and accuracy. The results of the classification report will show the accuracy of each variable in each parameter.

```
Rasio 70:30 - Validation accuracy: 0.89
Akurasi secara manual: 88.59%

Classification Report:
              precision    recall  f1-score   support

early_blight       0.88      0.91      0.89       200
 late_blight       0.89      0.86      0.88       177

    accuracy                           0.89       377
   macro avg       0.89      0.88      0.89       377
weighted avg       0.89      0.89      0.89       377
```

Figure 2. Results of Random Forest Ratio 70:30

In **Figure 2**, the model successfully recognized tomato leaves affected by early blight with an accuracy of 91%, and late blight was recognized with an accuracy of 86%. Overall, the Random Forest model with a 70:30 data split scenario achieved an accuracy level of 88.59%, indicating that the model has a strong performance in classifying diseases on tomato leaves.

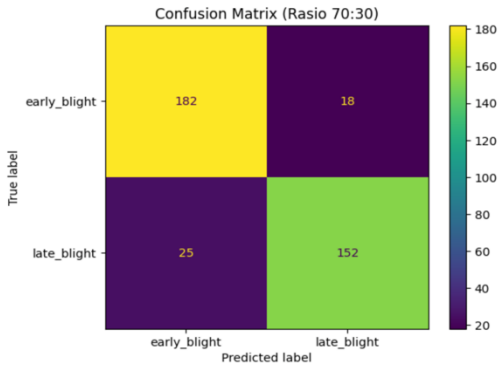


Figure 3. Confusion Matrix of Random Forest Ratio 70:30

Based on **Figure 3**, the confusion matrix shows the classification results with the data splitting scenario, where 70% of the data is for training and 30% for testing. The model successfully classified 182 early blight samples and 152 late blight samples correctly. However, there were 18 classification errors in early blight and 25 errors in late blight.

```
Rasio 80:20 - Validation accuracy: 0.91
Akurasi secara manual: 91.24%

Classification Report:
              precision    recall  f1-score   support

early_blight       0.90      0.95      0.92       142
 late_blight       0.93      0.86      0.90       109

    accuracy                           0.91       251
   macro avg       0.92      0.91      0.91       251
weighted avg       0.91      0.91      0.91       251
```

Figure 4. Results of Random Forest Ratio 80:20

In **Figure 4**, the model successfully recognized tomato leaves affected by early blight with an accuracy of 95%, while late blight was recognized with an accuracy of 86%. Overall, the Random Forest model with an 80:20 data split scenario achieved an accuracy level of 91.23%, indicating that the model has a fairly good performance in classifying diseases on tomato leaves.

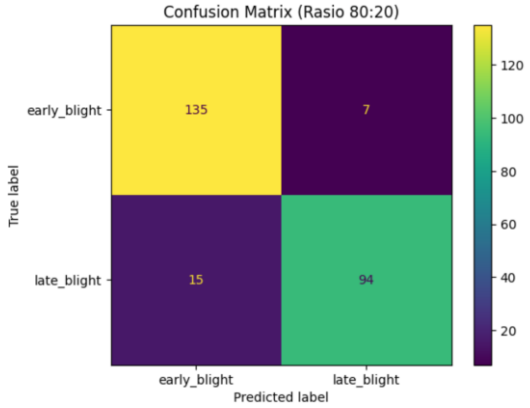


Figure 5. Confusion Matrix of Random Forest Ratio 80:20

Based on **Figure 5**, the confusion matrix shows the classification results with the data splitting scenario, with 80% of the data for training and 20% for testing. The model successfully classified 135 early blight samples and 94 late blight samples correctly. However, there were 7 classification errors for early blight and 15 for late blight.

```
Rasio 90:10 - Validation accuracy: 0.93
Akurasi secara manual: 92.86%

Classification Report:
              precision    recall  f1-score   support

early_blight       0.91      0.97      0.94        71
 late_blight       0.96      0.87      0.91        55

    accuracy                           0.93       126
   macro avg       0.93      0.92      0.93       126
weighted avg       0.93      0.93      0.93       126
```

Figure 6. Results of Random Forest Ratio 90:10

In **Figure 6**, the model successfully recognized tomato leaves affected by early blight with an accuracy of 97%, while late blight was recognized with an accuracy of 87%. Overall, the Random Forest model with a 90:10 data split scenario achieved an accuracy level of 92.86%, indicating that the model has good performance in classifying diseases on tomato leaves.

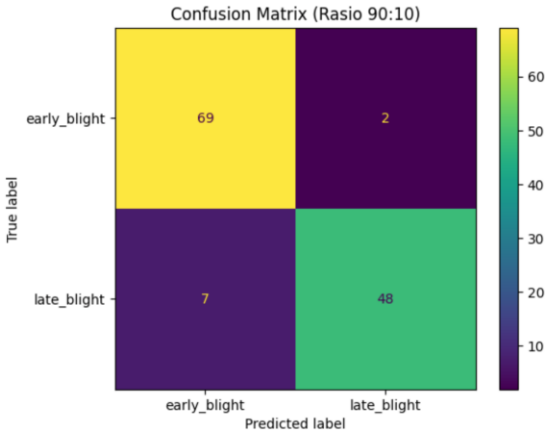


Figure 7. Confusion Matrix of Random Forest Ratio 90:10

Based on **Figure 7**, the confusion matrix shows the classification results with the data splitting scenario, where 90% of the data is for training and 10% for testing. The model successfully classified 69 early blight samples and 48

late blight samples correctly. However, there were 2 misclassifications in early blight and 7 misclassifications in late blight.

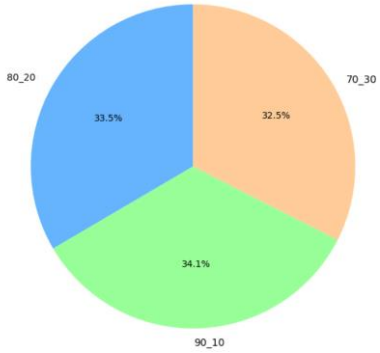


Figure 8. Comparison results of the accuracy of three data ratios

In **Figure 8**, a comparison of the accuracy of the Random Forest method is shown. The accuracy results of the Random Forest for each scenario are 32.5%, 33.5%, and 34.1%. The data was divided into three scenarios, namely 70:30, 80:20, and 90:10.

Naïve Bayes

	precision	recall	f1-score	support
early_blight	0.70	0.78	0.73	200
late_blight	0.71	0.62	0.66	177
accuracy			0.70	377
macro avg	0.70	0.70	0.70	377
weighted avg	0.70	0.70	0.70	377

Akurasi untuk rasio 70_30: 70.29%

Figure 9. Results of Naive Bayes Ratio 70:30

In **Figure 9**, the naive Bayes model was able to classify tomato leaves affected by early blight with an accuracy of 78% and late blight with an accuracy of 62%. Overall, the 70:30 data split scenario resulted in an accuracy rate of 70.29%, indicating that the model performed quite well in identifying both types of diseases in tomato plants.

	precision	recall	f1-score	support
early_blight	0.75	0.77	0.76	142
late_blight	0.69	0.67	0.68	109
accuracy			0.73	251
macro avg	0.72	0.72	0.72	251
weighted avg	0.72	0.73	0.72	251

Akurasi untuk rasio 80_20: 72.51%

Figure 10. Confusion Matrix of Naive Bayes Ratio 70:30

In **Figure 10**, the naive Bayes model is able to classify tomato leaves affected by early blight with an accuracy of 75% and late blight with an accuracy of 67%. Overall, the 80:20 data split scenario results in an accuracy level of 72.51%, indicating that the model performs quite well in identifying both types of diseases in tomato plants.

	precision	recall	f1-score	support
early_blight	0.80	0.79	0.79	71
late_blight	0.73	0.75	0.74	55
accuracy			0.77	126
macro avg	0.77	0.77	0.77	126
weighted avg	0.77	0.77	0.77	126

Akurasi untuk rasio 90_10: 76.98%

Figure 11. Results of Naive Bayes Ratio 80:20

In **Figure 11**, the naive bayes model was able to classify tomato leaves affected by early blight with an accuracy of 79% and late blight with an accuracy of 75%. Overall, the 90:10 data split scenario resulted in an accuracy rate of 76.98%, indicating that the model performed quite well in identifying both types of diseases in tomato plants.

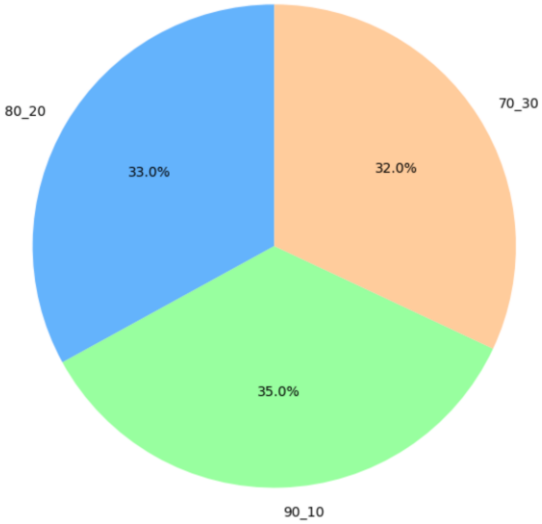


Figure 12. Comparison results of the accuracy of three data ratios

In **Figure 12**, a comparison of the accuracy of the Naive Bayes method is shown. The Naive Bayes accuracy results for each scenario are 32.0%, 33.0%, and 35.0%. The data was divided into three scenarios: 70:30, 80:20, and 90:10.

4. Conclusion

This study compares two methods, Random Forest and Naive Bayes, to classify diseases on tomato plant leaves, namely early blight and late blight. The data used in this study was obtained from Kaggle, consisting of 1,255 images. The data was processed thru several stages, namely preprocessing, normalization, and division into three data ratios (70:30, 80:20, 90:10) for training and testing. Based on the evaluation, the Random Forest method has higher accuracy compared to Naive Bayes. The evaluation results show that Random Forest with a (90:10 data ratio) achieved an accuracy of 92.86%, while Naive Bayes only reached 76.98%. In conclusion, the Random Forest method demonstrates more effective performance in improving disease classification in tomato plants compared to Naive Bayes. The accuracy achieved is certainly not perfect, so additional efforts are needed to improve the accuracy.

References:

[1] A. Verma, "Classifying Tomato Leaf Diseases Using Diverse Deep Learning Architectures: AlexNet, DenseNet, Inception, Xception," *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation Iatmsi 2025*. 2025, doi: 10.1109/IATMSI64286.2025.10984581.

[2] S. Batra, "An Improved Diagnostic Approach for Classifying Tomato Leaf Diseases using Ensemble Deep Learning based Technique," *8th International Conference on Electronics Communication and Aerospace Technology Iceca 2024 Proceedings*. pp. 820–825, 2024, doi: 10.1109/ICECA63461.2024.10800833.

[3] A. Vinothini, "Transfer learning based deep learning model for classifying tomato plant leaf diseases," *Eng. Res. Express*, vol. 7, no. 2, 2025, doi: 10.1088/2631-8695/add6f5.

[4] A. Sembiring, "The Performance of Various Concise Convolutional Neural Network Configurations in Classifying Tomato Diseases Based on Leaf Images," *Lecture Notes in Electrical Engineering*, vol. 1008. pp. 373–389, 2023, doi: 10.1007/978-981-99-0248-4_26.

[5] D. Kim, "Classification of surface settlement levels induced by TBM driving in urban areas using random forest with data-driven feature selection," *Autom. Constr.*, vol. 135, 2022, doi: 10.1016/j.autcon.2021.104109.

[6] O. S. Djandja, "Random forest-based modeling for insights on phosphorus content in hydrochar produced from hydrothermal carbonization of sewage sludge," *Energy*, vol. 245, 2022, doi: 10.1016/j.energy.2022.123295.

[7] K. Sen, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170399.

[8] I. Sulistiani, "Breast Cancer Prediction Using Random Forest and Gaussian Naïve Bayes Algorithms," *2022 1st Int. Conf. Inf. Syst. Inf. Technol. ICISIT 2022*, pp. 170–175, 2022, doi: 10.1109/ICISIT54091.2022.9872808.

[9] H. Nhat-Duc, "Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification," *Autom. Constr.*, vol. 148, 2023, doi: 10.1016/j.autcon.2023.104767.

[10] R. A. Disha, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, vol. 5, no. 1, 2022, doi: 10.1186/s42400-021-00103-8.

[11] K. N. Myint and Y. Y. Hlaing, "Predictive Analytics System for Stock Data: methodology, data pre-processing and case studies," *2023 IEEE Conf. Comput. …*, 2023.

[12] N. Rezova, L. Kazakovtsev, G. Shkaberina, and ..., "Data Pre-Processing for Ecosystem Behavior Analysis," *2022 Int. …*, 2022.

[13] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. …*, 2023.

[14] A. Balaram, "Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm," *Autom. Softw. Eng.*, vol. 29, no. 1, 2022, doi: 10.1007/s10515-021-00311-z.

[15] M. Salem, "Random Forest modelling and evaluation of the performance of a full-scale subsurface constructed wetland plant in Egypt," *Ain Shams Eng. J.*, vol. 13, no. 6, 2022, doi: 10.1016/j.asej.2022.101778.

[16] L. Zhang, "Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest)," *BMC Emerg. Med.*, vol. 22, no. 1, 2022, doi: 10.1186/s12873-022-00582-z.

[17] X. Yu, "Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes," *Aquat. Toxicol.*, vol. 251, 2022, doi: 10.1016/j.aquatox.2022.106265.

[18] Y. Shen, "Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms," *Comput. Geosci.*, vol. 159, 2022, doi: 10.1016/j.cageo.2021.105019.

[19] Y. Gu, "Predicting intersection crash frequency using connected vehicle data: A framework for geographical random forest," *Accid. Anal. Prev.*, vol. 179, 2023, doi: 10.1016/j.aap.2022.106880.

[20] M. V Anand, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[21] P. Venkata, "Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks," *Mater. Today Proc.*, vol. 62, pp. 7156–7161, 2022, doi: 10.1016/j.matpr.2022.03.035.

[22] E. Tieppo, "Classifying Potentially Unbounded Hierarchical Data Streams with Incremental Gaussian Naive Bayes," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13073, pp. 421–436, 2021, doi: 10.1007/978-3-030-91702-9_28.

[23] I. F. Hanbal, "Classifying Wastes Using Random Forests, Gaussian Naïve Bayes, Support Vector Machine and Multilayer Perceptron," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 803, no. 1, 2020, doi: 10.1088/1757-899X/803/1/012017.

[24] H. Azis, M. Abdullah, S. Ismail, and ..., "A Comparative Study of YOLO Models for Enhanced Vehicle Detection in Complex Aerial Scenarios," *2025 19th Int. …*, 2025.

[25] H. Azis, L. Syafie, F. Fattah, and ..., “Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques,” *2024*.

[26] Y. Zhao, “Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period,” *Ecol. Indic.*, vol. 135, 2022, doi: 10.1016/j.ecolind.2021.108529.