Journal of Intelligent System and Telecommunications

Journal homepage: https://journal.unesa.ac.id/index.php/jistel/index

Prediction of Soil Organic Carbon Based on Soil Color Using Random Forest

Muhammad Afifi Andriansyah[1], Moch. Arifin[2]*

[1] Department of Electrical Engineering, Faculty of Engineering
Universitas Negeri Surabaya, Surabaya, Indonesia
[2,] Department of Agrotechnology, University of Pembangunan Nasional"Veteran" Jawa Timur, Surabaya, Indonesia
*Correspondence: E-mail: afifiandriansyah27@gmail.com

ARTICLE INFO

Article History:
Submitted/Received 22 May 2025
First Revised 23 June 2025
Accepted 28 June 2025
First Available Online 30 June 2025
Publication Date 30 June 2025

Keyword:
Soil Organic Carbon, Soil Color, Random Forest, Machine Learning, Regression, Soil Fertility.

ABSTRACT

This study aims to predict the soil organic carbon (C-organic) content based on soil color using the Random Forest algorithm. This prediction is essential as C-organic is a key indicator of soil fertility. The method used is regression with a machine learning approach. The dataset was obtained from soil color images and actual C-organic laboratory results. The model was evaluated using metrics such as Mean Squared Error (MSE), R-squared (R^2), and accuracy. Additionally, a classification was performed to categorize the fertility level of the soil to support the prediction interpretation. The results showed excellent performance of the Random Forest regression model, with an R^2 of 0.9988 and accuracy of 99.88%. The fertility classification showed perfect precision and recall in all classes. These findings demonstrate that soil color can be effectively used to predict C-organic content and support data-driven agricultural decisions.

.

1. INTRODUCTION

Soil organic carbon content is an important parameter in determining soil fertility levels. Conventional laboratory analysis to measure organic carbon content usually requires a long time and significant costs. Therefore, fast and affordable automatic prediction methods are highly needed. Soil color, which can be captured through digital imaging, has shown a correlation with organic carbon content and can be developed into a non-destructive predictive approach. Soil color reflects complex physical and chemical characteristics of the soil, including organic matter content. In this study, the Random Forest algorithm was chosen because it can handle non-linear data and provide stable prediction results[1]. Random Forest also has the advantage of reducing the risk of overfitting [2].

This study specifically focuses on predicting soil organic carbon content (C-organic) using the Random Forest Regression model, based on soil features such as color, moisture, and pH. Soil color is used as the primary indicator due to its strong correlation with organic matter content and pH[3] [4]. Color data is obtained from a color sensor that reads RGB (Red, Green, Blue) values, which are then converted into HSV (Hue, Saturation, Value) format—considered more stable for visual representation.

In addition to predicting numerical values of C-organic, the model is also used to classify prediction results into three soil fertility categories: *Not Fertile*, *Slightly Fertile*, and *Fertile*. The purpose of this classification is to simplify the predictive output into actionable information that can be easily used by farmers in the field, enabling them to make rapid decisions regarding soil treatments, such as the application of organic fertilizers or pH correction[5].

Soil moisture is incorporated as a key parameter in the model because of its known impact on the decomposition process of organic matter and microbial activity, both of which significantly influence organic carbon levels [6] Recent studies also show that variations in moisture levels significantly affect sensor color readings and the interpretation of soil fertility classifications, which necessitates testing the model's robustness across different moisture ranges [7]. The evaluation demonstrated that the model consistently maintains accuracy above 99%, even under dry or saturated soil conditions—indicating strong generalization capability.

The selected model, Random Forest Regressor, was chosen for its ability to handle multivariate and non-linear data with high performance, while also providing insight into the relative contribution of each feature to the prediction outcome[8]. Random Forest operates by building numerous decision trees and aggregating their results through bagging, thereby reducing overfitting and improving predictive accuracy[9]

The model's performance was validated using evaluation metrics such as Mean Squared Error (MSE) and R^2 (coefficient of determination) for regression, along with confusion matrix, precision, and recall for classification. Visualizations such as scatter plots, confusion matrices, and boxplots were employed to evaluate the model's effectiveness under varying moisture conditions. By combining soil color and moisture features, and supported by a robust and interpretable machine learning method, the proposed model offers a practical, cost-effective, and accurate solution for real-time soil fertility monitoring—making it highly relevant to precision agriculture and sustainable land management[10] [11].

The performance evaluation of the model is carried out using the Mean Squared Error (MSE) metric, coefficient of determination (R-squared), confusion matrix, as well as precision and recall from the classification results. Visualizations in the form of scatter plots, bar charts, heatmaps, and boxplots are used to facilitate understanding of the relationship between actual data and prediction results. One additional analysis conducted is an observation of prediction accuracy at each humidity range, to determine how consistently the model works under different humidity conditions[12]. The use of the Random Forest method in this research also

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

provides advantages in terms of interpretability, as it can show the relative contribution of each feature to the prediction result.[13]

The color and soil moisture-based organic carbon prediction method using Random Forest has the potential to be an efficient tool in soil monitoring systems. With this technology, it is expected that soil management can be carried out more accurately, sustainably, and data-driven, in order to support food security and optimal agricultural productivity.

2. METHODS

2.1.Material

The dataset used in this study consisted of numerical features extracted from digital soil images, specifically Red (R), Green (G), and Blue (B) color channels that had undergone calibration. Additionally, Hue (H), Saturation (S), and Value (V) were derived through color space conversion. Complementary features such as soil pH and moisture content were also included. The target variable, soil organic carbon (C-Organik), was obtained through standardized laboratory tests. The data were collected from seven sampling points.

2.2.Method

The method employed in this study was the Random Forest Regressor, a machine learning algorithm capable of handling non-linear relationships and robust against overfitting. The workflow of this research included several key stages.

In the data preprocessing stage, important steps were taken to ensure the quality and relevance of the input data. These steps included handling any missing values, converting RGB (Red, Green, Blue) color values into the HSV (Hue, Saturation, Value) color space, and selecting the appropriate features for model training. The features used consisted of soil pH, moisture content, and color attributes (R, G, B, H, S, V), all of which have been shown to correlate with soil organic carbon (C-Organik) levels.

Following preprocessing, the Random Forest model was employed to predict continuous C-Organik values. These predictions were then categorized into three soil fertility classes to enhance interpretability and field applicability. The classification thresholds were defined as follows: soils with C-Organik values less than 1.5 were categorized as Unfertile (Tidak Subur), values between 1.5 and 2.5 as Less Fertile (Kurang Subur), and values equal to or above 2.5 as Fertile (Subur). This classification allows farmers to quickly interpret model output and make informed decisions regarding soil management.

2.3. Evaluation

Evaluation of regression performance was performed using MSE and R^2 metrics. Classification accuracy was assessed using confusion matrix, precision, and recall. The performance evaluation of the model was carried out using statistical indicators such as Mean Squared Error (MSE) and the coefficient of determination (R^2) to assess the accuracy of regression predictions. For classification tasks, confusion matrix analysis was used along with precision and recall values to measure how well the predicted labels matched the actual classes[14].

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

2.3.1. Mean Squared Error (MSE)

In the regression process, Mean Squared Error (MSE) is often used to measure the quality of the model by calculating the average of the squared differences between the actual and predicted values. The residuals mentioned in Regression-Kriging (RK) are part of this error, which is then further processed using kriging to model the spatial variation unexplained by the regression [15].

Mean Squared Error (MSE) measures the average of the squares of the differences between the actual and predicted values. It penalizes large errors more heavily, making it sensitive to outliers and large deviations[11].

MSE measures the average of the squares of the differences between the actual and predicted values. It penalizes large errors more heavily. The formula is:

Equation (1)

$$MSE = (1/n) * \Sigma (yi - \hat{y}i)^2$$

Where:

- yi is the actual value of C-Organik
- $\hat{y}i$ is the predicted value
- n is the number of observations

2.3.2. Coefficient of Determination (R^2 Score)

R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

Equation (2)

$$R^2 = 1 - [\Sigma (yi - \hat{y}i)^2 / \Sigma (yi - \bar{y})^2]$$

Where:

- $\bar{y}$ is the mean of the actual C-Organik values.

An R^2 value closer to 1 indicates that the model explains most of the variability in the target variable.

2.3.3. Classification Evaluation

Classification evaluation is the process of assessing how effectively a model distinguishes between different classes or categories. This process involves various metrics, such as accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC), each

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

offering insights based on the characteristics and goals of the classification task. Accuracy, while commonly used, can be misleading when dealing with imbalanced datasets; thus, measures like precision and recall are preferred in such contexts[16][17]. F1-score provides a balanced measure that considers both precision and recall, especially valuable when false positives and false negatives carry different consequences[18]. AUC, on the other hand, evaluates the model's ability to discriminate between classes regardless of threshold, making it suitable for comparing classifiers in imbalanced scenarios[19]. As noted by [20]. Selecting the right evaluation metric depends heavily on the specific objectives of the classification problem, the cost of misclassifications, and the distribution of the classes. After converting continuous predictions into categorical fertility classes, classification metrics were computed:

• Confusion Matrix: A table showing true vs. predicted class distributions.

• Precision: Measures the proportion of positive identifications that were actually correct:

Equation (3)

$$\text{Precision} = TP / (TP + FP)$$

• Recall (Sensitivity)

Recall (Sensitivity) is an evaluation metric in classification that shows a model's ability to detect all actual positive data, calculated as the ratio of true positives to the sum of true positives and false negatives. Recall is particularly important in applications where mistakes in overlooking positive cases can have serious consequences, such as in medical diagnosis or fraud detection. Recall is very useful for assessing how well a model detects all relevant cases, but it needs to be balanced with precision to avoid generating too many false positive predictions. [21] also emphasizes that recall is highly relevant in cases of imbalanced data, as it can describe the model's performance concerning the important minority class. Recent research has also shown the use of recall in various domains, such as disease detection[22] and hardware security, where positive cases become a priority.[23]

Measures the proportion of actual positives that were correctly identified:

Equation (4)

$$\text{Recall} = TP / (TP + FN)$$

Where:

• TP = True Positives

• FP = False Positives

• FN = False Negatives

These metrics were calculated for each class: Unfertile, Less Fertile, and Fertile.

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

3. RESULTS AND DISCUSSION

3.1. Regression Results

The updated evaluation shows excellent predictive performance with MSE of 0.0015 and R^2 of 0.9988, which corresponds to a prediction accuracy of 99.88%. The regression plot (Figure 1) shows near-perfect alignment with the ideal line (y = x).

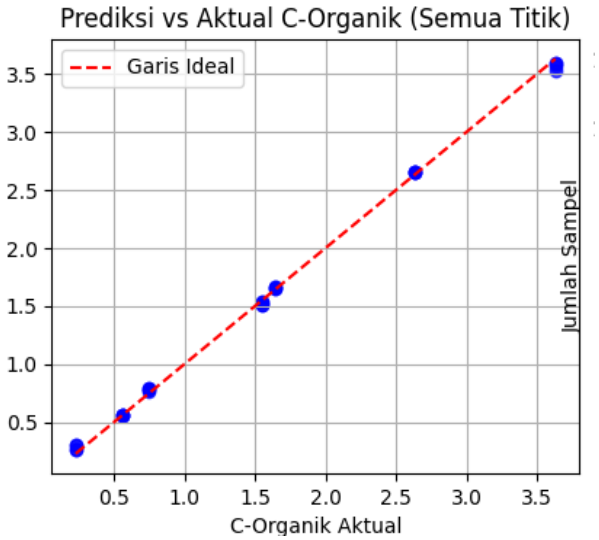

Figure 1. Scatter plot of predicted vs actual C-organic values.

Figure 1 presents a scatter plot comparing predicted values of soil organic carbon (C-organic) with those obtained from laboratory tests. Each point represents a sample, with the X-axis showing the actual lab values and the Y-axis indicating the model predictions. The dashed diagonal line (y = x) indicates ideal predictions. The close alignment of most points to this line demonstrates the model's high accuracy, further confirmed by a Mean Squared Error (MSE) of 0.0015 and a Coefficient of Determination (R^2) of 0.9988, implying that the model explains 99.88% of the variance. These results validate the strength of the model in regression tasks and support earlier findings by [2] that Random Forests are effective for nonlinear and complex datasets. The importance of features like soil color in RGB and HSV space, moisture content, and pH in estimating C-organic is also evident. These are strongly correlated to soil health and can serve as effective proxies for direct laboratory tests.

3.2. Fertility Classification

Classification was performed based on predicted C-organic values. The confusion matrix (Figure 2) shows that the model correctly classified all data points. The classification report showed perfect precision and recall (1.00) across all classes, indicating outstanding classification capability.

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

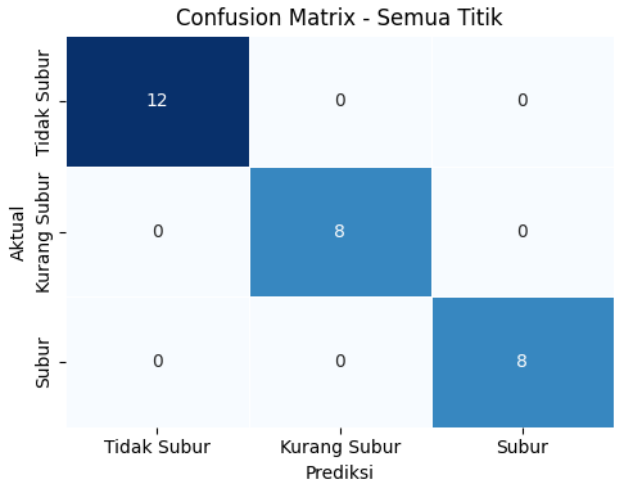

Figure 2. Confusion matrix of soil fertility classification.

This matrix compares the actual fertility categories from laboratory results with the categories predicted by the model.

- Y-Axis (Actual): Represents the true class labels from the dataset:
  - Unfertile
  - Less Fertile
  - Fertile
- X-Axis (Predicted): Represents the class labels as predicted by the model.

Table 1. Comparison of Actual vs. Predicted Soil Fertility Classes

Actual \ Predicted	Unfertile	Less Fertile	Fertile
Unfertile	12	0	0
Less Fertile	0	8	0
Fertile	0	0	8

The model achieved perfect classification performance. All predictions match the actual labels, with no misclassifications across any of the three fertility categories. This is indicated by all values falling along the diagonal of the confusion matrix. Such results demonstrate the model's exceptional ability to generalize and distinguish between fertility levels based on soil color features. This result underscores the model's exceptional discriminative power and its ability to generalize across varying sample conditions, particularly when using features such as soil color in RGB and HSV formats—parameters known to correlate strongly with key soil fertility indicators like organic carbon and pH[24]

The matrix reveals perfect classification, with all 28 samples correctly identified. The diagonal-only values and zero off-diagonal entries mean that there are no false positives or false negatives, resulting in precision and recall scores of 1.00 across all classes. Such performance underscores the robustness of the Random Forest classifier, as also shown in prior

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

research [1]. Table 1 provides a numerical summary of the classification results. All entries lie along the diagonal, reconfirming the zero-error performance. In real-world applications, especially in agriculture where misclassification may lead to over- or under-treatment of soil, this level of accuracy is critical. It demonstrates excellent feature separability, which is particularly important in multi-class problems and is a known strength of ensemble methods like Random Forest [25].

3.3. Accuracy by Moisture Range

To assess model robustness under various soil moisture conditions, the model was tested across different moisture ranges. Figure 4 shows consistent high accuracy (~99–100%) across all ranges, demonstrating the model's generalizability regardless of moisture content. This finding aligns with the results by [7], which showed that a soil sensor system maintained high predictive accuracy across different levels of soil moisture, from dry to wet conditions, with differences in C-organic and pH values remaining below 1% compared to laboratory results.

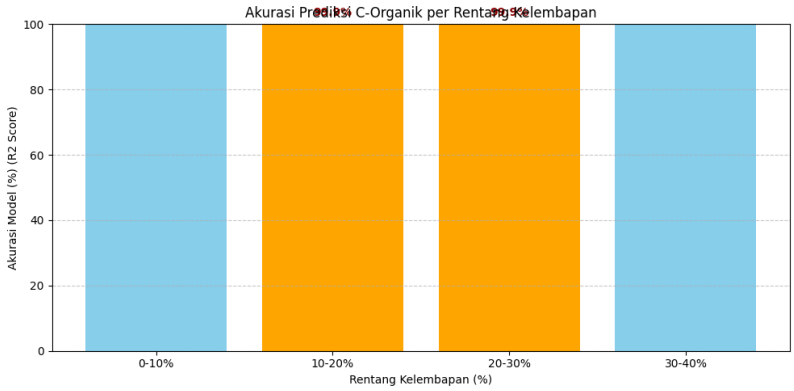

Figure 4. Prediction accuracy across different soil moisture ranges.

Figure 4 investigates model performance across different soil moisture levels, which can affect microbial activity and organic matter decomposition[7]. The model maintains an accuracy range of 99%–100%, showing excellent generalization under varying environmental conditions. This is consistent with previous findings that Random Forests maintain high performance in heterogeneous data environments [16]. This stability is crucial for real-time field applications. The ability to function accurately across moisture conditions suggests that the model is well-suited for integration into mobile apps or IoT-based soil sensors, enabling rapid, in-situ soil fertility analysis.

In summary, this study highlights how machine learning, particularly Random Forests, can transform traditional soil analysis. Using inexpensive sensors and accessible features like soil color and moisture, accurate fertility classification and C-organic prediction are not only feasible but field-deployable. This supports the move toward digital agriculture with real-time, data-driven decision tools for sustainable land management.

The proposed system offers high accuracy and real-time usability, especially in field conditions. However, it still relies on color sensor stability and may be influenced by external lighting or sensor calibration inconsistencies. Further testing on larger and more diverse datasets is recommended.

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X

4. CONCLUSION

This research confirms that soil color features can be effectively used to predict C-organic content using Random Forest regression. The model achieved a Mean Squared Error of 0.0015 and an R^2 of 0.9988, with a classification accuracy of 100%. These results support its integration into precision agriculture systems.

5. ACKNOWLEDGMENT

This section aims to express gratitude to all people/parties who helped with the research. Acknowledgements may also be written to anyone who provided intellectual contributions, technical assistance (including in writing and editing), or special equipment or materials.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

7. AUTHORS CONSTRIBUTION

Muhammad Afifi Andriansyah was responsible for the conceptualization, methodology design, software development, data collection, data analysis, visualization, and drafting the original manuscript. Moch. Arifin contributed as the data supervisor, overseeing technical aspects of data processing and participating in the review and editing of the manuscript.

8. REFERENCES

[1] A. Liaw and M. Wiener, "The R Journal: Classification and regression by randomForest," *R J.*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: http://www.stat.berkeley.edu/

[2] J. M. Klusowski, "Complete Analysis of a Random Forest Model," *arXiv*, vol. 13, pp. 1063–1095, 2018.

[3] S. S. Nodi, M. Paul, N. Robinson, L. Wang, and S. ur Rehman, "Determination of Munsell Soil Colour Using Smartphones," *Sensors*, vol. 23, no. 6, pp. 1–15, 2023, doi: 10.3390/s23063181.

[4] V. Ćirić *et al.*, "THE IMPLICATION OF CATION EXCHANGE CAPACITY (CEC) ASSESSMENT FOR SOIL QUALITY MANAGEMENT AND IMPROVEMENT," *Agric. For.*, vol. 69, no. 4, pp. 113–134, 2023, doi: 10.17707/AgricultForest.69.4.08.

[5] A. Kumala, S. Supriatna, and A. Damayanti, "Model assessment of soil organic matter content by remote sensing in Bayah, Indonesia," in *AIP Conference Proceedings*, American Institute of Physics Inc., Oct. 2018. doi: 10.1063/1.5064189.

[6] M. Lamsani, R. A. Pangestika, M. Cahyanti, and E. R. Swedia, "SISTEM IDENTIFIKASI WARNA TANAH MUNSELL MENGGUNAKAN SENSOR WARNA TCS3200 DAN KELEMBABAN YL-69," *Sebatik*, vol. 27, no. 1, pp. 379–389, Jun. 2023, doi: 10.46984/sebatik.v27i1.2249.

[7] M. A. Andriansyah and M. Arifin, "Designing Soil Color Sensors to Determine Soil Characteristics Based on Internet of Things (IoT)," vol. 14, no. 1, pp. 83–91, 2025.

[8] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds

of classifiers to solve real world classification problems?," *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, 2014.

[9] R. Tibshirani, "Lasso Tibshirani.pdf," 1996.

[10] J. N. Quinton and P. Fiener, "Soil erosion on arable land: An unresolved global environmental threat," Feb. 01, 2024, *SAGE Publications Ltd*. doi: 10.1177/03091333231216595.

[11] F. Sohil, M. U. Sohali, and J. Shabbir, "An introduction to statistical learning with applications in R," *Stat. Theory Relat. Fields*, vol. 6, no. 1, pp. 87–87, 2022, doi: 10.1080/24754269.2021.1980261.

[12] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[13] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[14] D. Pranesti, S. Fitri, and M. A. Hanafiah, "Journal of Intelligent System and Application of Principal Component Analysis (PCA) for Identifying Dominant Factors Affecting Energy Efficiency in a House," vol. 1, pp. 96–104, 2024.

[15] T. Hengl, G. B. M. Heuvelink, and D. G. Rossiter, "About regression-kriging: From equations to case studies," *Comput. Geosci.*, vol. 33, no. 10, pp. 1301–1315, 2007, doi: 10.1016/j.cageo.2007.05.001.

[16] B. Heung, C. E. Bulmer, and M. G. Schmidt, "Predictive soil parent material mapping at a regional-scale: A Random Forest approach," *Geoderma*, vol. 214–215, pp. 141–154, 2014, doi: 10.1016/j.geoderma.2013.09.016.

[17] S. Daskalaki, I. Kopanas, and N. Avouris, "Evaluation of classifiers for an uneven class distribution problem," *Appl. Artif. Intell.*, vol. 20, no. 5, pp. 381–417, 2006, doi: 10.1080/08839510500313653.

[18] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[19] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005, doi: 10.1109/TKDE.2005.50.

[20] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[21] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLoS One*, vol. 18, no. 10 October, pp. 1–27, 2023, doi: 10.1371/journal.pone.0291908.

[22] S. Noteboom *et al.*, "Evaluation of machine learning-based classification of clinical impairment and prediction of clinical worsening in multiple sclerosis," *J. Neurol.*, vol. 271, no. 8, pp. 5577–5589, 2024, doi: 10.1007/s00415-024-12507-w.

[23] M. Harju and A. Mesaros, "Evaluating Classification Systems Against Soft Labels with Fuzzy Precision and Recall," no. September, 2023, [Online]. Available: http://arxiv.org/abs/2309.13938

[24] V. Kautsar, K. Faizah, and A. I. Uktoro, "Soil Color Comparison Using Munsell Soil Color Chart and Calibrated Smartphone Camera," *J. Teknotan*, vol. 18, no. 1, p. 13, 2024, doi: 10.24198/jt.vol18n1.3.

[25] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, doi: 10.1890/07-0539.1.

DOI: https://doi.org/10.26740/jistel.v1n2.p189-199
e- ISSN 3090-613X