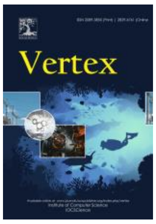

ISSN 2089-385X (Print) | 2829-6761 (Online), Volume 15, No. 2 (2026), pp. 34-41

Vertex

Published by: Institute of Computer Science (IOCS)

Unsupervised Machine Learning Based DSS for Land Profiling and Disease Risk Mitigation in Smart Farming

Embun Fajar Wati[1], Elvi Sunita[2], Andi Diah Kuswanto[3]

[1,2] Information Technology, Universitas Bina Sarana Informatika, Jakarta, Indonesia

Abstract

Decision Support Systems (DSS) in smart farming require methodologies capable of representing the holistic complexity of agricultural ecosystems. This study proposes a DSS framework based on unsupervised machine learning, specifically K-Means clustering, to automatically segment land profiles using IoT sensor records. The dataset consists of 500 global sensor data points covering seven essential environmental variables: soil moisture, pH, temperature, rainfall, humidity, sunlight duration, and the NDVI index. Through Principal Component Analysis (PCA) for dimensionality reduction and Silhouette Score evaluation, the system successfully identified and mapped seven land profiles with distinct microclimatic characteristics. Cross-tabulation analysis further demonstrates the principal novelty of this DSS, namely its ability to classify land into "Safe Zones" (Clusters 0, 3, and 4), which are characterized by Mild disease status and are suitable for Soybean, Cotton, and Maize, as well as "High-Risk Zones" (Clusters 1, 2, 5, and 6), which consistently correspond to Severe disease status. These findings indicate that a DSS based on environmental clustering is substantially more effective for crop selection recommendations and disease prevention than conventional predictive approaches. Ultimately, this framework provides farmers with actionable insights to optimize productivity and minimize agricultural risk.

Article Info

Article history:

Received : May 06,2026
Revised : May 25, 2026
Accepted : Jun 05, 2026

Keywords:

Crop disease risk;
Decision Support System;
IoT sensor data;
K-Means clustering;
Smart farming

Corresponding Author:

Embun Fajar Wati,
Information Technology,
Universitas Bina Sarana Informatika,
Jalan Kramat Raya No.98, DKI Jakarta 10450, Indonesia.
embun.efw@bsi.ac.id

Introduction

The adoption of the Internet of Things (IoT) in agriculture has enabled the continuous generation of sensor-based time-series data, which serves as the basis for the development of smart farming systems (Lestari et al., 2026)(Choudhary et al., 2025). Environmental variables such as soil moisture, air temperature, and vegetation indices including NDVI can now be monitored in real time to support smart farming practices (Wa ode denada mar ella et al., 2026)(Fathollahi et al., 2024). However, the abundance of such data is often not matched by the system's capacity to convert raw information into agronomic policies that are easily understood by field practitioners (Pratiwi et al., 2021). For this reason, a Decision Support System (DSS) is needed to uncover hidden patterns within the complex interactions among environmental variables without depending on crop yield labels, which are inherently influenced by multiple factors.

Most recent agricultural decision-support studies still rely heavily on supervised learning to predict crop yields in absolute terms (Jabed & Azmi Murad, 2024). This approach has an important limitation because crop yield is influenced by genetic characteristics and local microclimatic conditions that are often not captured by macro-scale IoT sensors (Nurwarsito et al., 2024)(Iriyanta et al., 2023). A significant research gap therefore remains in the use of IoT sensor data alone, without yield as a target variable, to gain a better understanding of dynamic land suitability (Syam et al., 2024). Specifically, the gaps in previous research regarding IoT-based clustering in smart farming are characterized by the scarcity of studies that leverage unsupervised learning to segment real-time environmental data for decision support. While traditional clustering methods often rely on static historical datasets or soil composition, there is a lack of frameworks that utilize continuous IoT sensor streams to dynamically identify land profiles and their direct correlation with crop health, specifically disease susceptibility. The main novelty of this research compared to conventional agricultural DSS is the transition from a supervised, yield-centric prediction model to an unsupervised clustering framework. Unlike traditional approaches that depend on historical yield labels to train algorithms, this study leverages pure IoT sensor data to autonomously discover land profiles and directly map these environmental patterns to crop disease susceptibility, providing a robust alternative that functions independently of complex yield-influencing variables.

This study addresses the identified gap by proposing a decision support system framework based on unsupervised learning (Akhiril Anwar Harahap et al., 2023)(Noviardi & Rosda Syelly, 2026). Rather than estimating crop yields, the system automatically segments land profiles to map interactions among sensor parameters and directly associate them with crop disease risk. To implement this approach, the problem-solving procedure follows a structured pipeline: first, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the complex IoT variables and minimize multicollinearity; second, K-Means Clustering is used to group the land profiles; and finally, cross-tabulation analysis is conducted to empirically verify the relationship between the resulting clusters, crop types, and disease severity.

How can a DSS model based on K-Means clustering be developed to group agricultural land profiles using IoT sensor data and identify crop disease distribution patterns within each cluster? This study aims to classify land profile characteristics into several clusters based on IoT sensor parameters through PCA and K-Means, as well as to examine the relationship between the resulting land profiles, crop dominance, and disease severity as the foundation for DSS recommendations.

Ultimately, this research aims to produce a robust DSS model that classifies land into actionable "Safe" and "High-Risk" zones. The practical value of this system lies in equipping farmers with accurate, data-driven insights for making better crop selection decisions and preventing disease at an early stage, thereby reducing agricultural losses and minimizing resource inefficiency.

Method

This study employs a computational quantitative approach and utilizes an unsupervised machine learning-based Knowledge Discovery in Databases (KDD) framework (Shu & Ye, 2023)(Yu et al., 2024). The KDD framework was selected because it offers a systematic method for extracting implicit, previously unknown, and potentially valuable insights from large volumes of IoT data.

Data processing was carried out computationally in 2024. The data were obtained from a secondary source, specifically a global agronomic repository available in the Kaggle dataset. This dataset includes 500 farm observation points collected from various regions around the world. The population consists of all sensor log records contained in the dataset. As the primary inputs for the DSS, seven continuous environmental variables were selected as features: soil_moisture_%, soil_pH, temperature_C, rainfall_mm, humidity_%, sunlight_hours, and NDVI_index (Arif et al., 2024). Seven environmental variables were selected as the primary features for the DSS because they comprehensively represent the critical biotic and abiotic factors that directly influence plant physiological processes and disease development. By encompassing soil conditions (moisture and pH), atmospheric parameters (temperature, rainfall, humidity, and sunlight), and vegetation health (NDVI), these variables provide a holistic environmental fingerprint necessary to accurately distinguish between land profiles without relying on yield data. In contrast, the categorical variables, namely crop type and disease status, were intentionally excluded from the clustering stage to preserve the unsupervised nature of the model and were instead used in the post-clustering cross-tabulation analysis.

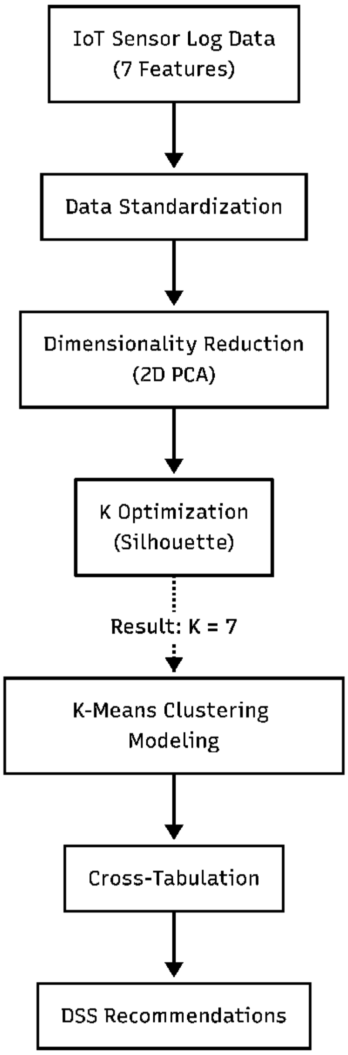

Figure 1. Flowchart of Research Stages

The procedure in this study, as shown in Figure 1, is presented chronologically and consists of five structured stages. First, pre-processing involves removing identification variables such as farm_id, timestamps, and coordinates, followed by data standardization using StandardScaler to ensure that all environmental features are given equal weight. Second, dimensionality reduction is performed by applying Principal Component Analysis (PCA) to transform the seven variables into two principal components, thereby reducing multicollinearity while preserving as much variance as possible (Sari, 2023). Third, the optimal number of clusters is determined by identifying the best k through the Silhouette Score method (Putri Vania & Betha Nurina Sari, 2023). Fourth, K-Means clustering is applied to group the data based on the optimal k value obtained in the previous step (Nugroho & Adhinata, 2022). Fifth, DSS formulation is conducted through cross-tabulation analysis to examine the relationship between the resulting clusters and the categorical variables, thereby generating decision rules (Giatma Dwijuna Ahadi & Chintyana, 2025).

To measure, test, and evaluate the model's results, its predictive performance will be rigorously assessed using Root Mean Square Error (RMSE) to quantify the magnitude of absolute error in Kg/ha, and the R-Squared (R2) metric to determine the proportion of variance in rice yield explained by the model. These evaluation metrics will be computed after inverse transformation to preserve real-world interpretability, and the outcomes will be compared with baseline historical averages to statistically confirm the added value of the multivariate exogenous features and the LSTM architecture.

Results and Discussions

The presentation of the data begins with the clustering optimization results, where the Silhouette Score analysis confirms that k=7 is the optimal number for segmenting the IoT sensor data. The factor that led to k=7 being selected as the optimal number of clusters is that it yielded the highest average Silhouette Score compared to other k values, indicating that the data points within each cluster are very close to each other (cohesive) while being distinctly far from points in other clusters (separated). This quantitative validation ensures that the seven clusters represent the most natural and statistically significant grouping of the complex environmental variables without forcing arbitrary divisions.

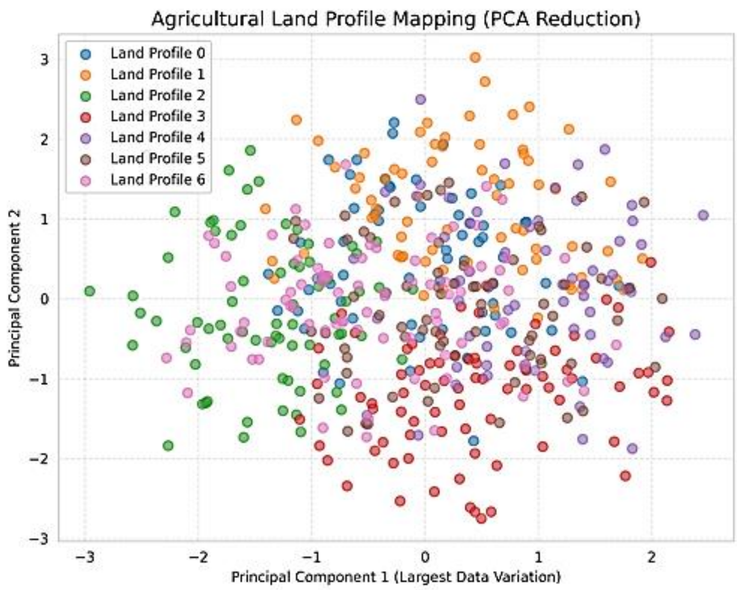

Figure 2. PCA Mapping Visualization

The effectiveness of the dimensionality reduction technique is clearly shown in Figure 2, where the PCA visualization reveals distinct boundaries among the seven land profile clusters with minimal overlap. This visual result strongly suggests that the combination of seven sensor parameters creates a unique environmental "fingerprint".

Table 1. Sensor Centroid Characteristics of 7 Land Profiles

Cluster	Soil Moisture (%)	Temp (°C)	Rainfall (mm)	Humidity (%)	Sunlight (hrs)	NDVI
0	23,34	24,12	133,01	77,42	8,6	0,75
1	22,97	27,11	139,51	50,33	5,97	0,68
2	31,6	23,31	108,42	73,72	6,4	0,42
3	35,66	25,01	206,57	53,07	8,4	0,55
4	33,24	20,08	237,06	68,88	6	0,77
5	25,81	28,33	247,38	75,41	6,79	0,55
6	15,2	24,49	196,87	61,25	7,02	0,55

To provide a quantitative explanation of these profiles, Table 1 presents the centroid output of the algorithm, which describes the average environmental conditions for each cluster. The interpretation of Table 1 is highly plausible because it reflects the complex inverse relationships among the variables; for example, Land Profile 4 shows the highest rainfall (237.06 mm) together with the lowest temperature (20.08°C) and the highest NDVI value (0.77). In contrast, Land Profile 6 indicates a critical agronomic irregularity: although rainfall remains relatively high (196.87 mm), soil moisture is extremely low (15.20%). This pattern suggests sandy soil conditions with excessive drainage, which is consistent with the theoretical explanation (Yonce Melyanus Killa et al., 2024) and demonstrates that the clustering results are not only statistically valid but also agronomically meaningful.

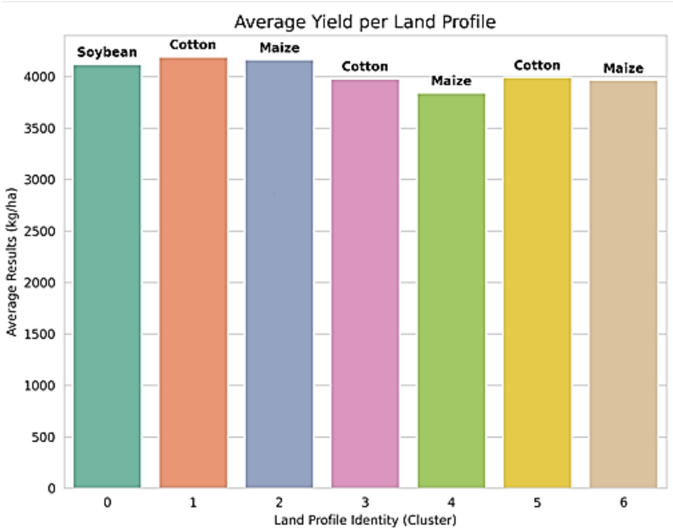

Figure 3. The Average Crop Yield

Although Figure 3 offers an initial explanation by presenting a bar chart of the average crop yield across clusters along with the dominant crop types, the main scientific contribution of this study is more clearly demonstrated through the cross-tabulation analysis in Table 2. This table is essential to the interpretation because it shifts attention from simple productivity values to the underlying relationship between environmental profiles and disease susceptibility.

Table 2. DSS Recommendations: Cross-Tabulation Analysis of Land, Crops, and Diseases

Cluster	Average Yield (kg/ha)	Dominant Crop	Dominant Disease Status	Land Area	DSS Zone Classification
0	4117,48	Soybean	Mild	61	Safe Zone
3	3976,18	Cotton	Mild	79	Safe Zone
4	3839,39	Maize	Mild	65	Safe Zone
1	4189,38	Cotton	Severe	70	Risk Zone
2	4161,77	Maize	Severe	74	Risk Zone

5	3990,25	Cotton	Severe	69	Risk Zone
6	3964,18	Maize	Moderate	82	Risk Zone

The discussion and analysis of Table 2 are directly aligned with the research objective, as they reveal an important novelty by successfully classifying the land into two distinct policy categories based on disease risk. The "Safe Zones" (Clusters 0, 3, and 4) are consistently associated with Mild disease status. The data reasonably indicate that Profile 0, which is marked by a balance of high humidity (77.42%) and optimal sunlight (8.60 hours), provides a suitable environment for Soybean, while Clusters 3 and 4 support the disease resistance of Cotton and Maize. This interpretation is highly consistent with the findings and supports the view (YUNI RESTI et al., 2022) that the balance between light exposure and humidity is a key indicator of crop resilience. In contrast, the "High-Risk Zones" (Clusters 1, 2, 5, and 6) offer a clear explanation of crop-land mismatch conditions. Land Profile 1, with high temperature (27.11°C), low humidity (50.33%), and low NDVI (0.68), logically creates environmental stress that leads to Severe disease in Cotton. Likewise, Land Profile 2's very low NDVI value (0.42) is directly associated with the highest disease severity in Maize. Overall, this analysis shows that the proposed DSS is capable of identifying hazardous environmental interactions that cannot be detected from yield values alone (Fathir et al., 2025)(Herfia Rhomadhona et al., 2021)(Fani Arinda et al., 2025).

The contribution of real-time clustering to the development of smart farming systems lies in its ability to transform static land management into dynamic, responsive interventions. By continuously segmenting land based on live sensor data, the system enables immediate detection of environmental shifts—such as sudden moisture drops or temperature spikes—that signal emerging disease risks before visual symptoms appear, thereby facilitating proactive rather than reactive agricultural practices. Furthermore, the implications of this research for the development of AI-driven agricultural DSS are profound, as it demonstrates that unsupervised learning models can serve as a robust alternative to traditional supervised systems in data-scarce environments. This study establishes a framework where the DSS does not merely predict future yields but actively diagnoses current land health and suitability, paving the way for autonomous decision-making systems that rely on environmental pattern recognition rather than extensive historical labeling.

Conclusions

This research successfully demonstrates that a Decision Support System (DSS) architecture based on unsupervised machine learning, specifically K-Means clustering, can map 500 land points into seven distinct microclimatic profiles using only seven IoT sensor parameters. The main findings indicate that patterns of environmental interaction are the key determinants of crop conditions, enabling the system to identify "Safe Zones" (Clusters 0, 3, and 4), which are resistant to severe disease, as well as "Risk Zones" (Clusters 1, 2, 5, and 6), which consistently correspond to severe disease status. The implications of using K-Means clustering for mapping agricultural land profiles are profound, as this approach allows for the objective segmentation of heterogeneous farmland into distinct management zones based on quantitative environmental similarities rather than arbitrary administrative boundaries. This granular mapping facilitates precision agriculture by enabling farmers to apply targeted inputs and specific crop varieties to each zone, optimizing resource allocation and minimizing waste while maximizing yield potential. The principal contribution of this study lies in the development of automatic, data-driven decision rules for disease risk mitigation, eliminating the need for predefined yield labels. The practical significance of these findings is substantial, as agricultural management can shift from reactive disease treatment to proactive, environment-based prevention by using sensor-driven zonal classifications to improve crop selection and reduce agricultural losses. However, this study also has several limitations. First, the current clustering model is based on static cross-sectional data and has not yet captured temporal dynamics or real-time fluctuations in microclimatic conditions. Second, the model does not include underlying soil physical properties, such as texture or structure, which may help explain anomalies such as excessive drainage. These limitations open a clear direction for future research, which is recommended to develop a dynamic clustering-based DSS capable of updating land profiles in real time as hourly sensor data are received. The opportunities for further research using dynamic clustering and streaming IoT data are significant, particularly in enabling the system to capture temporal dependencies and evolutionary patterns in microclimatic conditions. By leveraging continuous data streams, future DSS frameworks can evolve into adaptive early-warning mechanisms that detect environmental drift and sudden changes, allowing for immediate agronomic interventions rather than relying on static, historical snapshots. Furthermore, future research is recommended to incorporate deeper soil physical attributes into the feature set.

Reference

Akhiril Anwar Harahap, Muhammad Raihan, Nailul Aman, & Putri Risma Andini. (2023). Perbandingan Teknik Unsupervised Learninguntuk Pengelompokan Data Jumlah DesaDi Indonesia. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 163–170.

Arif, N., Nugraheni, D. R., & Husna, D. M. (2024). Exploring The Correlation Between NDVI, LST, And Soil Moisture in The Context of Climate Change. *International Journal for Disaster and Development Interface*, *4*(2), 123–139. https://doi.org/10.53824/ijddi.v4i2.85

Choudhary, V., Guha, P., Pau, G., & Mishra, S. (2025). An overview of smart agriculture using internet of things (IoT) and web services. *Environmental and Sustainability Indicators*, *26*, 100607. https://doi.org/10.1016/j.indic.2025.100607

Fani Arinda, Yani Maulita, & I Gusti Prahmana. (2025). SISTEM PENDUKUNG KEPUTUSAN UNTUK MENENTUKAN PRIORITAS PROGRAM PENGENDALIAN PENCEMARAN LINGKUNGAN MENGGUNAKAN METODE TOPSIS. *Global Research and Innovation Journal (GREAT)*, *1*(3), 954–960.

Fathir, Tri Stiyo Famuji, Erin Eka Citra, & Siti Mutmainah. (2025). Analisis Perancangan Sistem Informasi Pendukung Keputusan untuk Mitigasi Bencana Alam Berbasis Data Real-Time. *Scientific : Journal of Computer Science and Informatics*, *2*(1), 23–29. https://doi.org/10.34304/scientific.v2i1.339

Fathollahi, L., Wu, F., Melaki, R., Jamshidi, P., & Sarwar, S. (2024). Global Normalized Difference Vegetation Index forecasting from air temperature, soil moisture and precipitation using a deep neural network. *Applied Computing and Geosciences*, *23*, 100174. https://doi.org/10.1016/j.acags.2024.100174

Giatma Dwijuna Ahadi, & Chintyana, A. (2025). Analisis Data Kategorik dengan Crosstabulation dan Correspondence Analysis (Studi Retrospektif Data Puskesmas Pringgarata Triwulan I). *Jurnal Eksbar*, *2*(2), 1–9. https://doi.org/10.29408/eksbar.v2i2.33410

Herfia Rhomadhona, Winda Aprianti, Jaka Permadi, & Muhammad Khoirul Anam. (2021). Sistem Pendukung Keputusan Distribusi Bantuan Pertanian Menggunakan Simple Additive Weighting. *Journal of System and Computer Engineering (JSCE)*, *2*(2), 141–152.

Iriyanta, K., Putranto, B. P. D., & Andriyani, W. (2023). IOT BASED SOIL MOISTURE MONITORING AND SOIL MOISTURE PREDICTION USING LINEAR REGRESSION (CASE STUDY OF VINCA PLANTS). *Journal of Intelligent Software Systems*, *2*(1), 1. https://doi.org/10.26798/jiss.v2i1.929

Jabed, Md. A., & Azmi Murad, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

Lestari, N. R., Permana, R. R., Rotty, Y. T., & Fitrian, H. P. (2026). Analisis Penerapan Internet Of Things (IoT) dalam Pertanian Cerdas di Era 4.0. *Infotek: Jurnal Informatika Dan Teknologi*, *9*(1), 231–239. https://doi.org/10.29408/jit.v9i1.33209

Noviardi, & Rosda Syelly. (2026). Unsupervised Autoencoder untuk Deteksi Anomali Cerdas pada Perangkat Edge Computing Berbasis TinyML. *Jurnal Technologica*, *5*(1), 165–184.

Nugroho, N., & Adhinata, F. D. (2022). Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa. *Teknika*, *11*(3), 170–179. https://doi.org/10.34148/teknika.v11i3.502

Nurwarsito, H., Suprayogo, D., Sakti, S. P., Prayogo, C., Oakley, S., Wibawa, A. P., & Adaby, R. W. (2024). Development of Microclimate Data Recorder on Coffee-Pine Agroforestry Using LoRaWAN and IoT Technology. *Journal of Robotics and Control (JRC)*, *5*(1), 271–286. https://doi.org/10.18196/jrc.v5i1.20991

Putri Vania, & Betha Nurina Sari. (2023). Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means. *Jurnal Ilmiah Wahana Pendidikan* , *9*(21), 547–558.

Sari, D. R. P. (2023). METODE PRINCIPAL COMPONENT ANALYSIS (PCA) SEBAGAI PENANGANAN ASUMSI MULTIKOLINEARITAS. *PARAMETER: Jurnal Matematika, Statistika Dan Terapannya*, *2*(02), 115–124. https://doi.org/10.30598/parameterv2i02pp115-124

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

Syam, K. K. S., Manju, G., Thomas, S., & Binson, V. A. (2024). Precision crop prediction using IoT-enabled soil sensors and machine learning. *ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA*, *10*(49). https://doi.org/10.5935/jetia.v10i49.1219

Wa ode denada mar ella, Fajerin Biabdillah, Agusma Wajiansyah, & Abbizar Mulia. (2026). SMARTSOIL: SISTEM MONITORING KELEMBABAN TANAH BERBASIS INTERNET OF THINGS (IOT) MENGGUNAKAN ESP32 DAN SENSOR SOIL MOISTURE. *Jurnal Informatika Dan Teknik Elektro Terapan*, *14*(1). https://doi.org/10.23960/jitet.v14i1.8788

Yonce Melyanus Killa, Melycorianda Hubi Ndapamuri, Edmilson Umbu Ratu, & Matias Umbu Teul. (2024). Kajian Sifat Fisik Tanah pada Lahan Kering Beriklim Kering di Kecamatan Wulla Waijelu Kabupaten Sumba Timur. *JURNAL GALUNG TROPIKA*, *13*(1), 19–26. https://doi.org/10.31850/jgt.v13i1.1161

Yu, Y., Xu, P., Liu, S., He, T., Yang, L., & Zhang, J. (2024). An unsupervised machine learning-based profile system of Chinese researchers. *Journal of Infrastructure Policy and Development*, *8*(11), 7281. https://doi.org/10.24294/jipd.v8i11.7281

YUNI RESTI, RATIH KEMALA DEWI, & TERA FIT RAYANI. (2022). SUHU, KELEMBABAN DAN INTENSITAS CAHAYA PADA PENANAMAN GREEN FOODER MENGGUNAKAN SISTEM SMART HIDROPONIK. *Jurnal Sains Terapan : Wahana Informasi Dan Alih Teknologi Pertanian*, *12*(2), 77–85.