Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

# Analysis of Mango Leaf Condition Using the C-Means Method Based on Streaming Data

**Cinthya Agatha Sinaga[1], Herlina Br Nainggolan[2], Paska Marto Hasugian[3]**
[1, 2,3]Data Science Study Program, Santo Thomas Catholic University Medan, Jl.Setia Budi No.479F Tanjung Sari, Medan, Indonesia

**Article Info**
**Keywords:**
C-Means,
Data Streaming,
Leaf Condition Evaluation,
Computer Vision,
Mango Leaves

**ABSTRACT**
Efficiency in the agricultural sector is often hampered by conventional and manual plant identification processes. This study implements an automatic mango leaf condition evaluation system capable of live data acquisition, contour-based autocropping, and classification using the C-Means algorithm in a data streaming environment. The system extracts key features such as color (RGB), saturation, and contrast. Numerical data is normalized using Z-Score transformation before being grouped into three categories: Fresh, Sick, and Dry. The results show that the system is able to effectively distinguish biological conditions through automatic mapping. This research provides a responsive solution for real-time plant health monitoring.



**Corresponding Author:**
Cinthya Agatha Sinaga
Catholic University of Santo Thomas Medan, Jl. Setia Budi No. 479F Tanjung Sari, Medan, Indonesia
cinthyaagathaa@gmail.com

## INTRODUCTION

Digital transformation in Indonesia has triggered a massive surge in data volume, which has now become a strategic asset for organizations to improve operational efficiency. The use of Big Data has proven to be a vital instrument in the strategic decision-making process, where the integration of advanced information systems enables in-depth analysis of market behavior and community needs (Prasetyo and Wijaya 2024) This dynamic requires a solid framework so that the data can be managed appropriately to support sustainable digital economic growth (Handoko 2023).

However, the large amount of data that must be managed often poses a challenge for conventional physical infrastructure, which has scalability limitations. As a solution, Cloud Computing technology provides flexible storage and processing without the need for costly independent hardware investments (Prasetyo and Wijaya 2024). In Indonesia, cloud technology adoption has reached a point where companies and government agencies are beginning to rely on this infrastructure as the backbone of more responsive and transparent public services.

Several previous studies have explored the use of computer vision for plant health detection, such as research by Hidayat, Saputra, and Wijaya(2023), which found that color space analysis is very effective for detecting chlorosis in leaves through changes in color pigments. Meanwhile, Pratama and Santoso(2024) emphasize that color saturation and

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

intensity parameters are able to distinguish the severity of pathogen infection in leaf tissue in more detail.

On the other hand, Suryana and Putra(2024) found that texture feature extraction using leaf surface roughness patterns provides additional accuracy in recognizing physical damage caused by fungal attacks. Although these studies have produced significant results, most of them still rely on static data processing. Therefore, this study aims to integrate these color and texture parameters into a streaming data-based system using the C-Means Clustering method. This is done to provide a more responsive and automated identification solution, so that healthy and diseased leaves can be categorized with precision at the time of data collection.

The synergy between Cloud Computing and Big Data Analytics (BDA) opens up new opportunities in the implementation of Smart Cities in various urban areas in Indonesia. A strong cloud infrastructure is capable of supporting inter-system connectivity and real-time data processing necessary for smart and adaptive city management (Juroihan and Susanto 2024). Through integrated cloud sensors, local governments can now directly monitor various city parameters to improve the quality of life for citizens amid rapid urbanization (A. Widiastuti and Kusuma 2023).

In the context of data analysis, the use of distributed frameworks such as Apache Spark is now increasingly common for handling complex and varied datasets. The integration between cloud platforms and big data analysis tools enables batch and streaming processing at speeds far higher than traditional methods. This capability is crucial for the development of machine learning models that require large amounts of historical data to generate accurate and relevant predictions for the industrial sector.

Although it offers many conveniences, migration to the digital ecosystem also brings new challenges, especially related to information security and personal data privacy. Regulations such as the Personal Data Protection Law (PDP Law) in Indonesia are very important as a legal basis for mitigating the risk of data leaks in the era of Big Data (Gonzalez and Woods 2018) . Therefore, the implementation of high-level security protocols such as data encryption and multi-factor authentication is a mandatory standard for cloud service providers to ensure the integrity of user information.

In conclusion, the integration of Big Data and Cloud Computing is an inevitable strategic step in driving digital transformation in Indonesia. A balance is needed between aggressive technological innovation and the readiness of human resources and adaptive regulations to deal with the complexity of data in the future (Khirade and Patil 2015). Understanding the determining factors in the adoption of this technology will be the key for organizations to remain competitive and provide optimal services to the public (Atmadja et al., 2023).

## METHODS

This section outlines the systematic stages in the development of a leaf condition analysis system, from streaming data acquisition to the final classification process using a cloud computing and big data approach.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

## Clustering Architecture

The system workflow is comprehensively designed to ensure the efficiency of image data processing from acquisition to visualization. The integration between image capture hardware and processing algorithms in the cloud enables real-time and automatic analysis (Fauzi and Nugroho 2025) .

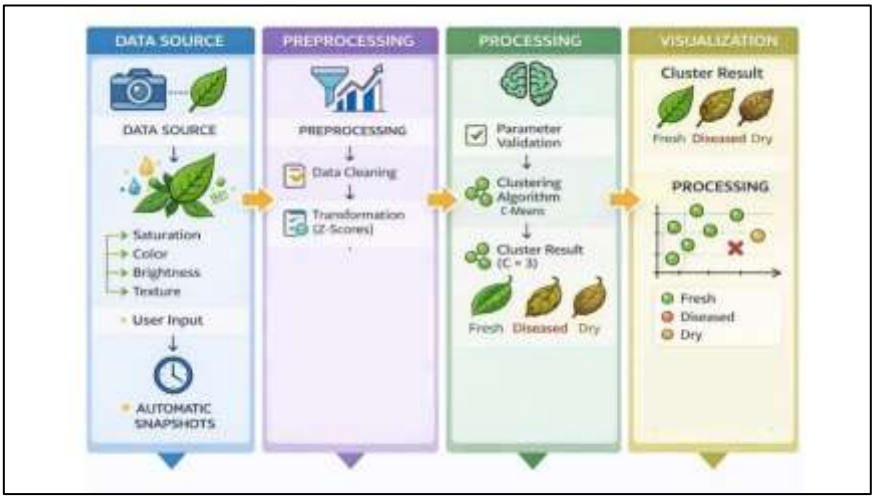


**Figure 1** Clustering architecture

This system architecture is designed as an integrated linear workflow, consisting of four main stages:

1. Data Sourced (Data Acquisition): The process begins with the acquisition of leaf images directly through a camera. The system extracts key parameters such as saturation, color, brightness, and texture (Malathi and Sinthia 2019) . After the Automatic Snapshots mechanism captures the images, the system performs the Selection stage by detecting leaf contours using precision red markers, which is followed by Autocropping to isolate the pure leaf object from the background (Sari and Ramadhan 2025) .

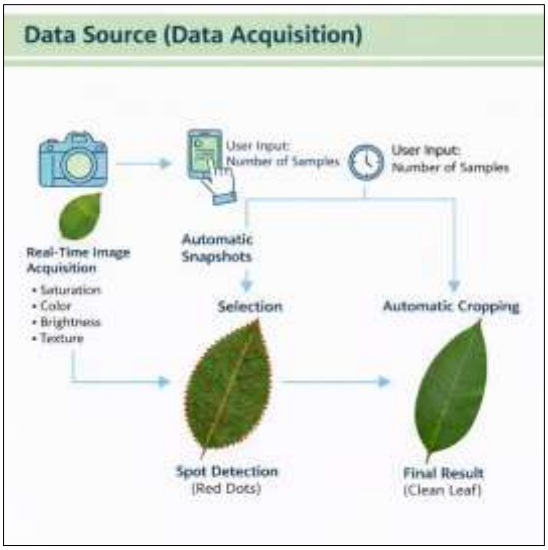


**Figure 2** Data Sourced Mechanism

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

After the Automatic Snapshots mechanism captures the image, the system automatically performs the Selection stage by detecting leaf contours using precise red dot markers. This stage ends with the Autocropping process, which removes the entire background area outside the red dot detection to ensure that only pure leaf object data is forwarded to the next processing stage.

2. Preprocessing: The raw dataset that has been collected enters the cleaning stage to select images that are suitable for processing. After the dataset is cleaned, a transformation is performed using the Z-Scores method .

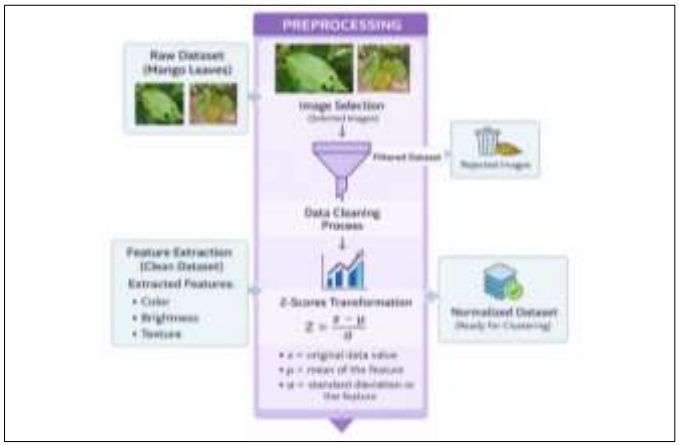


**Figure 3** Data Preprocessing Mechanism

This transformation stage is crucial for standardizing the scale of numerical features from various parameters so that they have balanced weights before being processed by the algorithm.

$$Z = \frac{(X - \mu)}{\sigma} \quad (1)$$

An explanation of the components of the formula is as follows (B. Setiawan and Pratama 2023):

Z: Standardized value of the transformation result.

x: Original value of the measured feature (saturation, color, brightness, or texture).

μ: The mean value of the entire data sample.

σ: Standard deviation indicating the spread of data from the mean.

Through this transformation, each parameter with different units is converted to the same scale with a mean of 0 and a variance of 1, thereby preventing the dominance of certain features in the distance calculation in the C-Means algorithm.

3. Processing: At this stage, Parameter Valorization is performed as a final validation of features before entering the clustering engine. The system uses the C-Means algorithm to automatically group the data. Based on the validated parameters, the algorithm divides the data into two main clusters (C=3), which represent the biological conditions of the plants (Kusuma and Rahayu 2024).

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

4. Visualization (System Output): The processing results are presented in two formats: Cluster Results (visual categories of Fresh or Sick) and Scatter Plot, which maps the spatial distribution of data to show the scientific separation of data groups (R. Pratama and Santoso 2024).

## Dataset Analysis

The dataset analysis stage is a fundamental initial step to ensure data quality before clustering. Leaf morphological features are extracted into a numerical format that can be recognized by the system to accurately represent plant conditions (Khirade and Patil 2015).

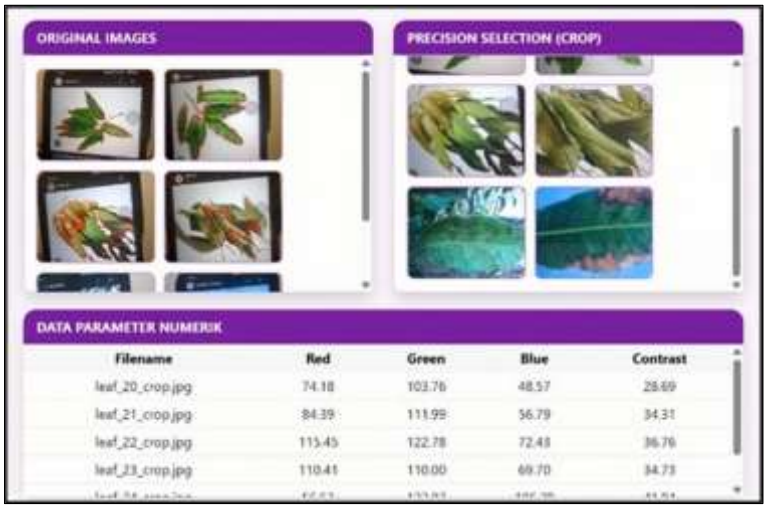


**Figure 4** Camera Capture Dataset

The system demonstrates excellent adaptability in separating leaf objects from varied backgrounds through the Precision Selection mechanism. The red spot detection process precisely follows the outer contours of the leaves, which are then automatically saved to the Static/captures directory. This database becomes the main reference for the system in classifying plant health in the subsequent stages (Fauzi and Nugroho 2025).

## Parameter Analysis

In determining leaf condition, the system performs feature extraction based on visual characteristics that reflect the physiological condition of the plant. These features are quantified into four main parameters summarized in the following table:

**Table 1** Object Parameters

| Parameter | Description | Significance in Leaf Condition |
|---|---|---|
| Color | Representation of chromatic values (often in RGB or HSV color space). | Detecting chlorosis or pigment changes; diseased leaves tend to show yellow or brownish colors (R. Hidayat, Saputra, and Wijaya 2023). |
| Saturasi | A measure of the intensity or purity of a color. | Distinguishing the severity of infection; areas affected by pathogens usually have different saturation compared to healthy tissue (B. Pratama and Santoso 2024). |

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

| Parameter | Description | Significance in Leaf Condition |
|---|---|---|
| Brightness | The level of light reflection or luminance from the leaf surface. | Identifying necrosis or spots; dried or dead leaves often have a brightness level that contrasts with fresh leaves (N. Widiastuti, Kusuma, and Rahayu 2023). |
| Texture | The degree of smoothness, roughness, or regularity of patterns on the surface. | Recognizing physical damage; fungal or bacterial attacks often alter leaf texture, making it rough or perforated (M. Suryana and Putra 2024). |

The table details four main parameters in leaf image analysis, namely color, saturation, brightness, and texture, each of which serves to detect specific biological indications such as chlorosis, necrosis, or physical damage caused by pathogens. This explanation provides a technical basis for the system to objectively distinguish between healthy, diseased, and dry leaves through accurate numerical feature extraction.

## RESULTS AND DISCUSSION

### Results of System Testing and Feature Extraction Implementation

Before the visual grouping results are displayed, the system first performs numerical feature extraction as mathematical indicators to distinguish the biological conditions of mango leaves. This process produces an objective representation of data, where each visual parameter is converted into measurable statistical values (Prasetyo and Wijaya 2024).

**Table 2** Feature Extraction Results Image

| filename | mean_r | mean_g | mean_b | mean_h | mean_s | mean_v | contrast |
|---|---|---|---|---|---|---|---|
| leaf_1_crop.jpg | 3.84 | 73.89 | 1.73 | 0.93 | 6.01 | 3.96 | 1.31 |
| leaf_2_crop.jpg | 98.67 | 3.84 | 61.79 | 26.72 | 4.92 | 103.69 | 1.77 |
| leaf_3_crop.jpg | 103.77 | 5.63 | 113.82 | 2.90 | 65.93 | 135.92 | 29.95 |
| leaf_4_crop.jpg | 3.05 | 4.89 | 1.90 | 1.98 | 6.52 | 4.90 | 0.92 |
| leaf_5_crop.jpg | 3.88 | 4.13 | 2.92 | 37.74 | 100.82 | 107.84 | 48.95 |
| leaf_6_crop.jpg | 91.66 | 115.77 | 2.88 | 1.95 | 105.83 | 4.84 | 0.99 |
| leaf_7_crop.jpg | 5.05 | 5.20 | 3.24 | 1.47 | 4.15 | 5.40 | 1.26 |
| leaf_8_crop.jpg | 3.81 | 4.75 | 4.01 | 65.62 | 2.59 | 113.65 | 1.44 |

The statistical values in the table are then visually mapped into an RGB line graph and a saturation versus contrast scatter plot to see the spatial distribution pattern of the data.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

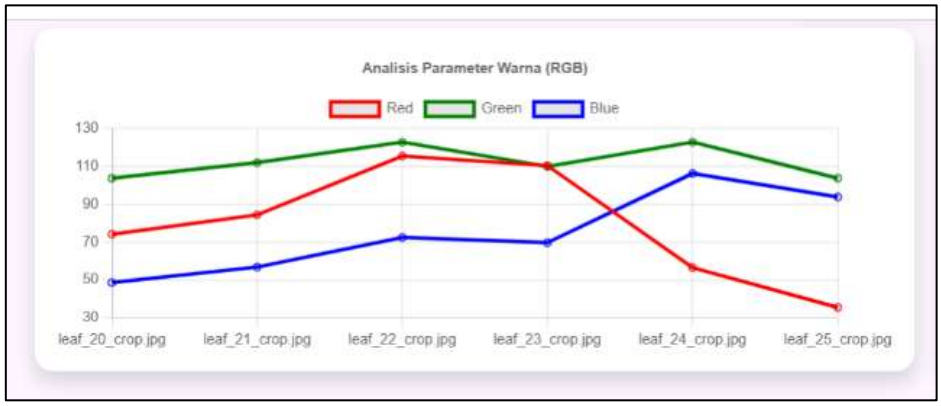


**Figure 5** Visual Results of Color Parameter Analysis

This visualization presents the average value trends of the Red, Green, and Blue (RGB) color channels from six processed mango leaf image samples. Visually, this graph shows the dominance of green color intensity in most samples, which is an indicator of healthy chlorophyll content. However, significant fluctuations are seen in the leaf_24_crop.jpg and leaf_25_crop.jpg samples, where the red and blue intensity values begin to overlap or decrease dramatically, indicating a change in color pigments due to pathogens or drying. Furthermore, Figure 7 displays a scatter plot that maps the spatial relationship between the Saturation and Contrast (Texture) parameters.

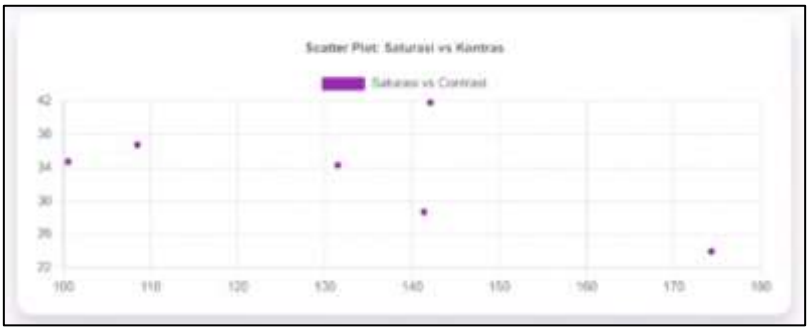

**Figure 6** Visual Analysis Results of Contrast and Saturation Parameters

This mapping is crucial because it shows data density in certain areas. It can be seen that most samples are clustered in the contrast value range between 100 and 140 with a saturation level above 30. However, there are outliers in the contrast values above 170 with low saturation (around 24), which mathematically indicates leaves with rough surface textures and pale colors, consistent with the biological characteristics of leaves in sick or dry conditions.

### C-Means Clustering Results

The final stage of this system is the visualization of the clustering results, which divides the entire mango leaf image dataset into relevant biological categories. The implementation of the C-Means algorithm successfully identified patterns of similarity in numerical features and distributed the samples into three functional clusters, namely "Fresh," "Sick," and "Dry" (Wicaksono and Utami 2025). This visualization output presents an intuitive classification through the division of boxes based on the physical characteristics of each object.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization


**Figure 7** Fresh Condition Cluster Results

Figure 7 shows the classification results for the "Fresh" cluster. Images in this category are dominated by leaves with high chlorophyll intensity, mathematically represented by stable mean_g (green) values and low contrast values. The system accurately maintains the visual integrity of the leaves through clean autocropping results, ensuring that only information from the leaf surface is used in the validation process .


**Figure 8** Cluster Results for Sick Condition

Figure 8 shows samples classified into the "Sick" cluster. This grouping is based on the detection of anomalous color points or spots on the leaf surface. Technically, the system recognizes fluctuations in saturation and brightness parameters that are not uniform compared to healthy leaves (Sari and Ramadhan 2025). This separation proves the sensitivity of the algorithm in objectively capturing the early symptoms of pathogen infection.


**Figure 9** Dry Condition Cluster Results

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Figure 9 presents the classification results for the "Dry" category. Leaves in this cluster have physical characteristics of pigment changes to a dominant brown color (necrosis) and a rough surface texture. This is consistent with feature extraction data showing very high contrast values and a drastic decrease in the green color channel (I. Suryana and Putra 2024). The system's success in separating dry conditions from diseased conditions indicates a good level of accuracy in recognizing the final phase of biological degradation in plants.

## CONCLUSION

This study successfully implemented a plant health condition classification system based on digital images by integrating numerical feature extraction and the C-Means clustering algorithm. The test results show that the use of color (RGB), saturation, brightness, and texture (contrast) parameters combined with Z-Score transformation can produce highly accurate data segmentation. The system can effectively distinguish the biological condition of leaves into three main categories, namely fresh, diseased, and dry, by minimizing human subjectivity in the field diagnosis process. The integration of a linear system architecture from real-time data acquisition and automatic pre-processing to scatter plot visualization proves that Big Data and Cloud Computing technologies can be practically applied in the agricultural sector. The successful storage of the extracted dataset in the system directory also provides great opportunities for the development of a more extensive plant health database in the future. Overall, this system offers an efficient and objective smart solution to support sustainable plantation health monitoring.

## REFERENCE

Fauzi, A, and S Nugroho. 2025. "Data Pipeline Architecture for an Automated *Jurnal Resti (Engineering Systems and Information Technology)* 9(1).

Gonzalez, R C, and R E Woods. 2018. *Digital Image Processing*. 4th ed. Pearson.

Handoko, D. 2023. "Digital Transformation in the Agricultural Sector: Opportunities and Challenges in Indonesia." *Journal of Digital Economics and Business* 2(4).

Hidayat, R, A Saputra, and K Wijaya. 2023. "Color Feature Extraction for Early Detection of Mango Leaf Anthracnose Using RGB-HSV Transformation." *Journal of Agricultural Technology and Information Systems* 6(1): 12–25.

Juroihan, M, and A Susanto. 2024. "Implementation of Real-Time Streaming Data in Smart Agricultural Monitoring Systems." *Journal of* 15(1).

Khirade, S D, and A B Patil. 2015. "Plant Disease Detection Using Image Processing." *2015 International Conference on Computing Communication Control and Automation*: 768–71.

Kusuma, W A, and S Rahayu. 2024. "Development of Distributed C-Means Algorithm for Big Data Analytics." *Scientific Journal of* 10(1).

Malathi, M, and P Sinthia. 2019. "Color and Texture Feature Extraction of Plant Leaf Disease." *International Journal of Applied Engineering Research* 14(2): 450–55.

Prasetyo, E, and A Wijaya. 2024. "Integration of Cloud Computing and Big Data for Real-Time Agricultural Image Analysis." *Journal of Information Technology and Computer*

*Science (JTIIK)* 11(2).

Pratama, B, and I Santoso. 2024. "Analysis of Saturation and Chromaticity in Digital Images for Plant Disease Severity Grading." *International Journal of Computing and Digital Systems* 15(2): 301–15.

Pratama, R, and B Santoso. 2024. "Identification of Leaf Pathogen Infection Severity Based on Saturation and Color Intensity Parameters." *Journal of Data Science and Agrotechnology* 6(1): 12–20.

Sari, M P, and F Ramadhan. 2025. "Optimization of Autocropping in Leaf Image Segmentation Using Contour Detection." *Journal of Informatics Education and Research (JEPIN)* 11(3).

Setiawan, A, and M R Pratama. 2023. "Performance Analysis of Z-Score Normalization in Plant Disease Feature Extraction." *Journal of Computer Science and Agricultural Informatics* 5(2): 112–20.

Suryana, I, and D Putra. 2024. "Texture Feature Extraction for Classification of Physical Damage to Mango Leaves Due to Fungal Attack." *Journal of Computing and Artificial Intelligence* 4(3): 101–15.

Suryana, M, and E Putra. 2024. "Texture Features and Gray Level Co-Occurrence Matrix (GLCM) for Classification of Mango Leaf Health." *Applied Engineering in Agriculture* 40(1): 45–58.

Suryana, T, and R E Putra. 2024. "Analysis of Texture and Contrast for the Identification of Necrosis in Tropical Plants." *Journal of Science and Technology* 13(2).

Wicaksono, G, and S Utami. 2025. "Implementation of the C-Means Clustering Algorithm for Web-Based Plant Health Monitoring." *Journal of* 6(1).

Widiastuti, A, and H Kusuma. 2023. "Comparative Analysis of Cloud-Based Processing Performance*Journal of Telematics* 18(2).

Widiastuti, N, H Kusuma, and S Rahayu. 2023. "Luminance and Brightness Analysis for Identification of Necrotic Spots on Tropical Fruit Leaves." *Journal of Soft Computing and Data Science* 4(3): 88–99.