Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2296~2302
e-ISSN: 2548-6861

Detection of Sugarcane Leaf Disease Using Pre-Trained Feature Extraction and SVM Method

Mufidatul Izza [1]*, Moch Lutfi [2]*
* Teknik Informatika, Universitas Yudharta Pasuruan
mufidahizza12@gmail.com [1], moch.lutfi@yudharta.ac.id [2]

Article Info

Article history:

Received 2025-08-01
Revised 2025-09-05
Accepted 2025-09-10

Keyword:

VGG16,
SVM,
Image classification,
Sugarcane leaf disease,
Feature extraction.

ABSTRACT

Sugarcane (Saccharum officinarum) is an important commodity in the sugar industry, but it is vulnerable to leaf diseases such as Red Rot, Rust, Yellow Leaf, and Mosaic, which can significantly reduce the quality and quantity of yields. Manual identification is time-consuming and prone to subjective errors, therefore an automatic detection method based on digital images is required. This study proposes a combination of VGG16 pre-trained as a feature extractor with Support Vector Machine (SVM) as a classifier. The dataset used is the Sugarcane Leaf Disease Dataset from Kaggle, consisting of 2,521 images of five classes, which were then balanced through augmentation in the form of rotation, zoom, and flipping to a total of 3,000 images (600 per class). The preprocessing stage includes resizing the images to 224×224 pixels and normalization using the preprocess_input function. Three model scenarios were tested, namely SVM, VGG16, and VGG16+SVM. Evaluation was carried out using two methods, namely an 80:20 train–test split and 10-fold cross-validation, with metrics of accuracy, precision, recall, F1-score, G-Mean, and AUC. The experimental results show that VGG16+SVM provides the best performance with an accuracy of 99.60% on the 80:20 scheme, while on 10-fold cross-validation the average accuracy is 80.76%. This value surpasses the baseline SVM and VGG16 + Softmax, proving that the integration of VGG16 feature extraction with SVM classification can produce stable and accurate performance. This research contributes to the development of image-based plant disease detection systems to support precision agriculture and fast decision-making.

I. INTRODUCTION

Sugarcane (Saccharum officinarum) is one of the most important agricultural commodities [1] and serves as the primary raw material in the sugar industry [2]. As a major sugarcane-producing country, Indonesia has a high demand for optimal sugarcane production to meet national sugar needs [3]. However, sugarcane productivity is often threatened by leaf diseases [4] such as Red Rot, Rust, Yellow Leaf, and Mosaic, which can significantly reduce harvest quality [5],[6]. These diseases are generally characterized by changes in color, shape, and specific patterns that can be observed through digital images [7]. The use of digital image processing technology has therefore become an important solution for early disease detection [8]. One of the commonly applied approaches is feature extraction, which aims to capture essential information from images to represent object characteristics. In the context of visual feature extraction, the VGG16 deep learning model has proven effective in extracting features from various image types [9],[10]. Research by [11]demonstrated that VGG16 achieved the highest accuracy of 89.5% in citrus leaf disease detection, outperforming models such as InceptionV3. In this study, however, VGG16 is employed solely as a feature extractor, utilizing the output from the convolutional base without the fully connected layer, to obtain feature representations more efficiently. The classification stage is performed using Support Vector Machine (SVM), which, according to [12], achieved the highest accuracy of 87% compared to other algorithms. Building on these findings, this study combines

the strength of VGG16 in feature extraction with the robustness of SVM in classification to develop a more efficient and accurate sugarcane leaf disease detection system. Model performance is evaluated using accuracy, precision, recall, F1-score, G-Mean, and ROC AUC metrics, with validation carried out through both an 80:20 train–test split and 10-fold cross-validation. This study is expected to make a significant contribution to the development of an automatic, fast, and accurate sugarcane leaf disease detection system, while also supporting the productivity of the national agricultural sector through the application of appropriate artificial intelligence technologies.

II. METHODS

This research was conducted following the stages shown in Figure 1.

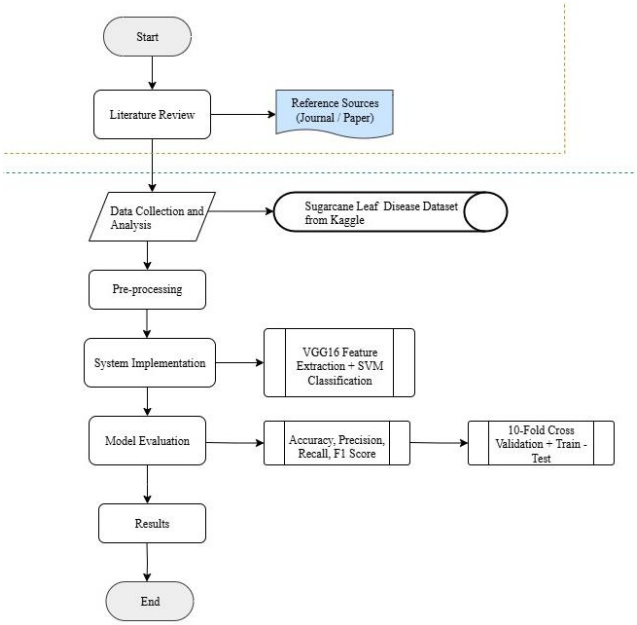

Figure 1. Research Methods

The research process began with a literature review of various references, both national and international journals, to gain an understanding of sugarcane leaf disease problems and relevant detection methods. This was followed by the collection of the Sugarcane Leaf Disease Dataset from Kaggle, which consists of five leaf image classes (Healthy, Red Rot, Rust, Yellow Leaf, and Mosaic). The dataset underwent preprocessing, including resizing the images to 224×224 pixels, normalization using the preprocess_input function, and augmentation techniques such as rotation, flipping, and zooming to balance class distribution. The system was then implemented by extracting features using the pre-trained VGG16 architecture on the convolutional base without the fully connected layer, after which the extracted features were classified using the Support Vector Machine (SVM) algorithm, known for its effectiveness in handling high-dimensional data. Model evaluation was carried out using two validation methods, namely train–test split (80:20) and 10-fold cross-validation, with performance metrics including accuracy, precision, recall, F1-score, G-Mean, and ROC AUC. This research workflow demonstrates the systematic relationship between stages, and is expected to produce an efficient, accurate, and reliable method for sugarcane leaf disease detection.

A. *Literature Review*

At this stage, references were collected from journals, articles, and other reliable sources. The purpose of this review was to gain a deeper understanding of the fundamental concepts of sugarcane leaf diseases, the transfer learning technique (VGG16), and the Support Vector Machine (SVM) classification algorithm.

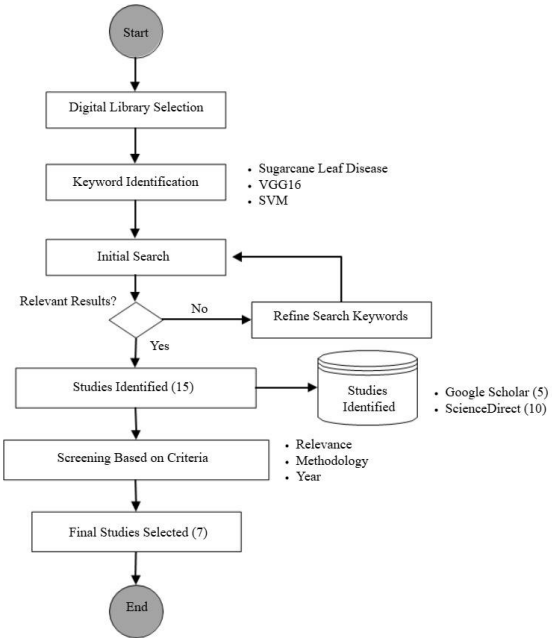

Figure 2. Find to References

B. *Dataset*

The dataset used in this study is the Sugarcane Leaf Disease Dataset, which was downloaded from Kaggle.com and consists of 2,521 sugarcane leaf images.

Figure 3. Leaf Deases

The dataset includes five image categories representing different leaf conditions, including healthy leaves and leaves

infected with various diseases, namely Healthy, Red Rot, Rust, Yellow Leaf, and Mosaic. The number of sugarcane leaf images used for each class is presented in Table 1:

TABEL I
IMAGE CLASS

No.	Class	Image Count
1.	Healthy	522
2.	Red Rot	518
3.	Rust	514
4.	Yellow	505
5.	Mosaic	462

C. *Data Preprocessing*

The preprocessing steps include:

1. Resize:

All images were resized to 224×224 pixels.

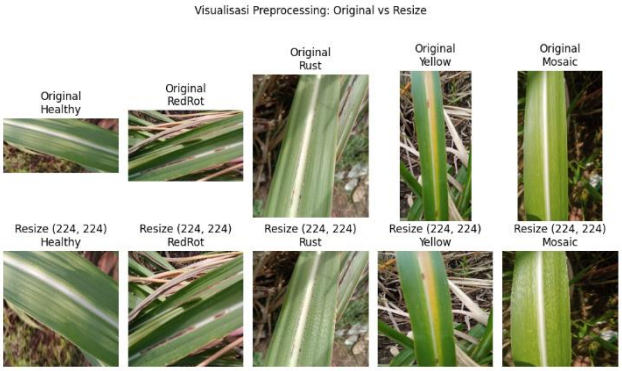

Figure 4. Preprocessing

2. Data Augmentation:

The images in the dataset were first resized to 224×224 pixels, converted to arrays using img_to_array, and then normalized using preprocess_input.

Figure 5. Preprocessing

Augmentation was performed using ImageDataGenerator with transformations including rotation (20°), width and height shifts, shear, zoom (0.1), horizontal flipping, brightness adjustment (darker/brighter), and channel shifting. The augmentation process was applied only to classes with fewer than 600 images to balance the data distribution across classes, preventing the model from being biased toward classes with more samples.

TABEL II
ORIGINAL DATASET

No.	Category	Image Count
1.	Healthy	522
2.	Red Rot	518
3.	Rust	514
4.	Yellow	505
5.	Mosaic	462
	Total	2.521

TABEL III
DATA AUGMENTATION

No.	Category	Image Count
1.	Healthy	600
2.	Red Rot	600
3.	Rust	600
4.	Yellow	600
5.	Mosaic	600
	Total	3.000

3. Normalization:

Normalization was performed using the preprocess_input() function from TensorFlow to ensure compatibility with the VGG16 format.

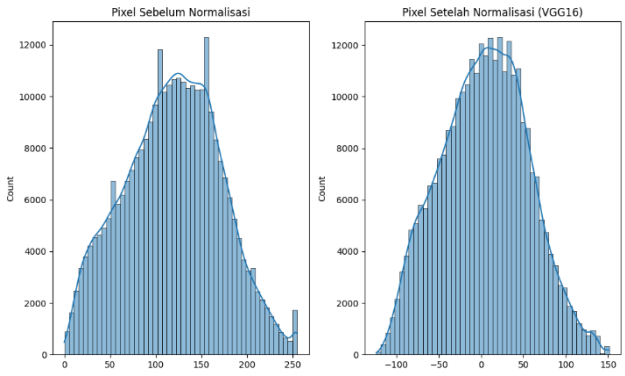

Figure 6. Normalization

D. *Data Splitting*

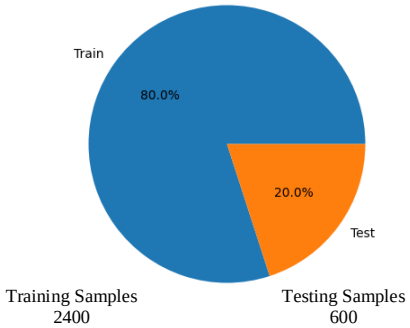

Figure 6. Data Splitting

TABEL IV
NUMBERS OF DATA PER CLASS

	Class	Train Count	Test Count
0	Healthy	480	120
1	RedRot	480	120
2	Rust	480	120
3	Yellow	480	120
4	Mosaic	480	120

TABEL V
10-FOLD CROSS VALIDATION

Fold	Accuracy (%)
1	0,84
2	0,81
3	0,78
4	0,79
5	0,80
6	0,78
7	0,85
8	0,77
9	0,80
10	0,80
Average	**0,80**

E. System Implementation

1. Feature Extraction with VGG16

VGG16 is a deep learning architecture pre-trained on the ImageNet dataset. In this study, the VGG16 model was employed as a feature extractor, where the fully connected layers were removed and only the convolutional base was used. Features were extracted from the last output layer (block5_pool) and then flattened into a fixed-dimensional vector that represents the visual characteristics of the images.

2. SVM Classification Model Training

Classification was performed using the Support Vector Machine (SVM) algorithm from the scikit-learn library. SVM works by finding the optimal hyperplane that separates data from different classes. The SVM parameters, such as kernel, C, and gamma, significantly influence the model's performance.

3. Parameter Optimization with GridSearchCV

To achieve optimal performance, parameter tuning was conducted using GridSearchCV. GridSearchCV is a technique for finding the best parameters by testing all specified combinations of parameters.

F. Model Evaluation

The model performance was evaluated using accuracy, precision, recall, F1-score, G-Mean, and ROC AUC, calculated using the following formulas:

1. Accuracy: The percentage of correct predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: The model's capability to accurately predict positive instances.

$$precision = \frac{TP}{TP + FP}$$

3. Recall: The model's capability to detect all positive samples.

$$recall = \frac{TP}{TP + FN}$$

4. F1-Score: The harmonic average of precision and recall.

$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

5. Confusion Matrix: A matrix showing the distribution of predictions and actual labels for each class [13].
6. G-Mean: Measures the balance of model performance in multi-class classification based on the geometric mean of recall for each class[14].
7. ROC Curve and AUC: The ROC Curve illustrates classification performance across various thresholds, while the AUC (Area Under the Curve) quantifies the model's ability to distinguish between classes[15].

G. The Proposed Model

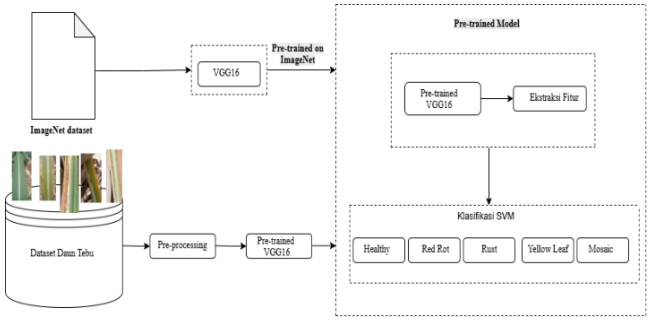

Figure 7. Proposed Model

The proposed model, as illustrated in the figure, represents the workflow of the sugarcane leaf disease detection system in this study. The process begins with dataset collection and image preprocessing, followed by feature extraction using the VGG16 architecture pre-trained on ImageNet. The extracted feature vectors are then processed by the Support Vector Machine (SVM) algorithm to classify the images into five categories: Healthy, Red Rot, Rust, Yellow Leaf, and Mosaic.

III. RESULTS AND DISCUSSION

A. Software and Hardware Requirements

All implementation processes were carried out on the Google Colaboratory (Colab) platform, which supports NVIDIA Tesla T4 GPU computing. The programming language used was Python 3.x, with supporting libraries such as TensorFlow, Keras, Scikit-Learn, NumPy, and Matplotlib. The operating system employed was Windows 10 (64-bit) for file management and integration with Google Colab. This

hardware and software configuration was selected to ensure efficient and stable computational performance.

B. Feature Extraction Results with VGG16

Feature extraction was performed using the pre-trained VGG16 model with the fully connected layers removed, leaving only the convolutional base. Each preprocessed image was extracted into a 25,088-dimensional feature vector, resulting in a total feature representation of size (3000, 25088) for the entire dataset.

C. SVM Model Training and Validation Results

Training was conducted using the SVM algorithm, chosen for its capability to separate classes in high-dimensional spaces. Parameter optimization (C, kernel, gamma) was performed using GridSearchCV, which tested combinations of parameters with 10-fold cross-validation to ensure stable model performance and prevent overfitting.

During the training process of a Support Vector Machine (SVM) model, several key parameters significantly influence its performance and classification results. The C parameter, ranging between [10, 100], serves as a regularization parameter that controls the balance between the separating margin and the training error. A higher C value tends to produce a narrower margin but may increase the risk of overfitting, while a lower value allows a wider margin with greater tolerance for misclassification. Next, the Kernel parameter determines the type of kernel function used to map the data into a higher-dimensional space, with options such as ‘rbf’, ‘poly’, and ‘sigmoid’. The choice of kernel greatly affects the model’s ability to handle non-linear data. Lastly, the Gamma parameter, which can take the values ‘scale’ or ‘auto’, controls how far the influence of a single training example reaches. A higher gamma value makes the model more focused on nearby data points (more sensitive to noise), whereas a lower gamma value allows a broader influence across the data space.

In this specific SVM configuration, the C parameter is set to 100, meaning the model places a strong emphasis on minimizing classification errors during training. A high C value drives the model to create a tighter separating margin, although it increases the risk of overfitting. The Kernel used is Sigmoid, a type of kernel that operates with an activation function similar to that used in neural networks, making it suitable for capturing non-linear relationships among features. Meanwhile, the Gamma parameter is set to Auto, which automatically computes its value based on the number of features in the dataset. This configuration helps the model adjust the influence of each data point on others proportionally to the complexity of the dataset.:

This combination of parameters achieved the highest validation score and was used to train the final model, which was subsequently evaluated.

D. Model Performance Comparison

TABEL VI
TRAIN TEST 80:20

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM	93,12%	0,932	0,931	0,901	0,905
VGG16 + Softmax	90,95%	0,949	0,909	0,959	0,943
VGG16 + SVM	99,60%	0,996	0,996	0,996	0,995

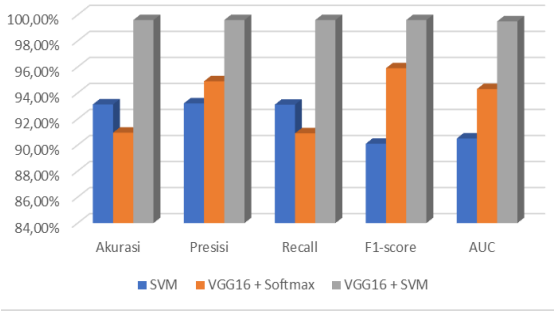

Figure 8. Train Test 80:20

TABEL VII
10-FOLD CROSS VALIDATION

Fold	SVM	VGG16 + Softmax	VGG16 + SVM
1	0,72	0,75	0,80
2	0,73	0,76	0,80
3	0,71	0,74	0,78
4	0,72	0,75	0,79
5	0,73	0,76	0,80
6	0,71	0,74	0,78
7	0,74	0,77	0,85
8	0,70	0,73	0,77
9	0,72	0,75	0,80
10	0,73	0,76	0,80
Average	**0,72**	**0,75**	**0,80**

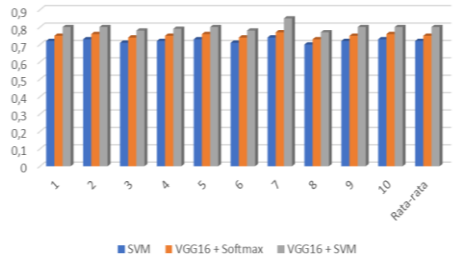

Figure 9. 10-Fold Cross Validation

E. Model Evaluation

The confusion matrix on the test data (80:20) illustrates the distribution of correct and incorrect predictions for each class. This visualization helps to assess the model’s performance in more detail, particularly in distinguishing between sugarcane leaf disease classes. In general, the confusion matrix is used to evaluate classification performance by showing the number of correct and incorrect predictions for each class.

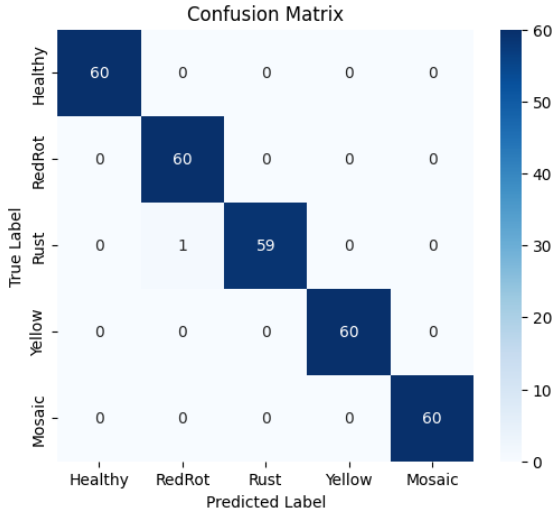

Figure 10. 10-fold cross-validation

Based on the 10-fold cross-validation results, the highest accuracy was obtained in Fold 7, with a value of 0.85. This score was the highest among all folds, indicating that Fold 7 can be considered the best representation of the SVM model's performance in classifying sugarcane leaf images.

F. Comparison with Previous Studies

TABEL VIII
PREVIOUS STUDIES

Framework	Dataset	Model	Accuracy	Evaluation	Splitting
Sujatha et al., 2021	Daun Citrus	VGG16 (CNN)	89%	Akurasi, Precision, F1 Score	-
Chiatra & Sabita, 2025	Daun Tebu	HOG + SVM	96%	Akurasi, Precision, F1 Score, 10-Fold CV	10-Fold CV
Proposed Method	Daun Tebu	VGG16 + SVM	99,60% (Train–Test 80:20) 80% (10-Fold CV)	Akurasi, Precision, Recall, F1 Score, AUC	Train - Test 80:20 10-Fold CV

The table presents a comparison between previous studies and the proposed method. Sujatha et al. (2021) applied end-to-end VGG16 on citrus leaves, achieving an accuracy of 89%, while Chiatra & Sabita (2025) used HOG + SVM on sugarcane leaves, obtaining 96% accuracy. The proposed method in this study, VGG16 + SVM, achieved an accuracy of 99.60% using an 80:20 train–test split and 80% with 10-fold cross-validation.

G. Discussion of Results

The evaluation using the 80:20 train–test split scheme showed that the combination of VGG16 as a feature extractor and SVM as a classifier achieved the best performance compared to the baseline models. The VGG16 + SVM model attained an accuracy of 99.60%, with average precision, recall, and F1-score values of 0.996. The confusion matrix visualization indicated that most predictions were on the main diagonal, demonstrating the model's strong ability to distinguish between Healthy, Red Rot, Rust, Yellow Leaf, and Mosaic classes. An AUC value of 0.995 further supports the high consistency of the model in separating the classes.

For comparison, the pure SVM baseline model achieved only 93.12% accuracy, while VGG16 + Softmax reached 90.95%. This difference highlights that integrating a pre-trained CNN with SVM is superior in detecting complex visual patterns in sugarcane leaf images compared to single models. The application of data augmentation also played a crucial role in increasing image variation and balancing class distribution, preventing the model from being biased toward certain classes.

These findings are consistent with Sujatha et al. (2021), who reported that VGG16 has strong capability in feature extraction for plant disease detection. However, unlike previous studies using end-to-end VGG16, this study utilized only the convolutional base as a feature extractor, resulting in a more efficient approach without involving fully connected layers.

IV. CONCLUSION

This study proposed a digital image-based method for detecting sugarcane leaf diseases by combining VGG16 as a feature extractor and SVM as a classifier. Using the 80:20 train–test split scheme, the proposed model achieved an accuracy of 99.60%, with average precision, recall, and F1-score of 0.996, and an AUC of 0.995. These results demonstrate that the integration of VGG16 and SVM can deliver very high performance in classifying sugarcane leaf diseases, outperforming both the pure SVM baseline and VGG16 + Softmax models. The main contribution of this study is to validate the effectiveness of combining a pre-trained CNN with SVM to support an automated sugarcane leaf disease detection system. This method has the potential to serve as a foundation for developing fast, efficient, and accurate AI-based applications to support precision agriculture practices. However, this study has limitations due to the dataset being limited in size and sourced from a single origin. Therefore, future research should test the model on field data with greater environmental variation and consider integrating it into mobile or IoT applications so that it can be directly utilized by farmers.

REFERENCES

[1] S. Thite, Y. Suryawanshi, K. Patil, and P. Chumchu, "Sugarcane leaf dataset: A dataset for disease detection and classification for machine learning applications," *Data Br.*, vol. 53, p. 110268, 2024, doi: 10.1016/j.dib.2024.110268.

[2] W. L. Pratitis, K. Kurniasari, and H. Al Fata, "Classification of Spotted Disease on Sugarcane Leaf Image Using Convolutional Neural Network Algorithm," *JTECS J. Sist. Telekomun. Elektron. Sist. Kontrol Power Sist. dan Komput.*, vol. 3, no. 2, p. 117, 2023, doi: 10.32503/jtecs.v3i2.3433.

[3] A. Arifin, A. Arrasyid, M. A. Firlata, and S. A. Putra, "Klasifikasi Penyakit Tanaman Tebu dengan Pendekatan Support Vector

Machine," vol. 3, no. 10, pp. 2613–2617, 2024.

[4] I. Ordine Pires da Silva Simões, R. G. de Freitas, D. E. Cursi, R. G. Chapola, and L. R. do Amaral, "Recognition of sugarcane orange and brown rust through leaf image processing," *Smart Agric. Technol.*, vol. 4, no. January, pp. 0–6, 2023, doi: 10.1016/j.atech.2023.100185.

[5] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 153–160, 2023, doi: 10.57152/malcom.v3i2.897.

[6] N. Amarasingam, F. Gonzalez, A. S. A. Salgadoe, J. Sandino, and K. Powell, "Detection of White Leaf Disease in Sugarcane Crops Using UAV-Derived RGB Imagery with Existing Deep Learning Models," *Remote Sens.*, vol. 14, no. 23, 2022, doi: 10.3390/rs14236137.

[7] A. Petchiammal and D. Murugan, "Automated Paddy Leaf Disease Identification using Visual Leaf Images based on Nine Pre-trained Models Approach," *Procedia Comput. Sci.*, vol. 252, pp. 118–126, 2025, doi: 10.1016/j.procs.2024.12.013.

[8] T. Huang, R. Yang, W. Huang, Y. Huang, and X. Qiao, "Detecting sugarcane borer diseases using support vector machine," *Inf. Process. Agric.*, vol. 5, no. 1, pp. 74–82, 2018, doi: 10.1016/j.inpa.2017.11.001.

[9] P. K. Mannepalli, A. Pathre, G. Chhabra, P. A. Ujjainkar, and S. Wanjari, "Diagnosis of bacterial leaf blight, leaf smut, and brown spot in rice leafs using VGG16," *Procedia Comput. Sci.*, vol. 235, pp. 193–200, 2024, doi: 10.1016/j.procs.2024.04.022.

[10] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, "A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans," *Meas. Sensors*, vol. 24, no. September, p. 100506, 2022, doi: 10.1016/j.measen.2022.100506.

[11] R. Sujatha, J. M. Chatterjee, N. Z. Jhanjhi, and S. N. Brohi, "Performance of deep learning vs machine learning in plant leaf disease detection," *Microprocess. Microsyst.*, vol. 80, no. October 2020, p. 103615, 2021, doi: 10.1016/j.micpro.2020.103615.

[12] S. D. Chiatra and H. Sabita, "Deteksi Objek Daun Tebu Dengan Menggunakan Metode Klasifikasi Pada Machine Learning," *Semin. Nas. Has. Penelit. dan Pengabdi. Masy. 2025*, pp. 113–124, 2025.

[13] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. December 2022, 2023, doi: 10.1016/j.cor.2022.106131.

[14] Anjna, M. Sood, and P. K. Singh, "Hybrid System for Detection and Classification of Plant Disease Using Qualitative Texture Features Analysis," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 1056–1065, 2020, doi: 10.1016/j.procs.2020.03.404.

[15] F. L. P. de Souza, M. A. Dias, T. D. Setiyono, S. Campos, L. S. Shiratsuchi, and H. Tao, "Identification of soybean planting gaps using machine learning," *Smart Agric. Technol.*, vol. 10, no. December 2024, 2025, doi: 10.1016/j.atech.2025.100779.