Volume 11 Issue 1 Year 2026 | Page 183-192 ISSN: 2527-9866
Received: ***28-12-2025*** | *Revised:* ***12-01-2026*** | *Accepted:****26-01-2026***

Color and Texture Feature Extraction for Disease Identification in Chili Leaves Using K-Nearest Neighbors

Andreyas[1], Derry Alamsyah[2]
[1,2] Multi Data Palembang University, Palembang, South Sumatra, Indonesia, 30113
e-mail: andreyas_2226250025@mhs.mdp.ac.id[1], derry@mdp.ac.id[2]
*Correspondence: andreyas_2226250025@mhs.mdp.ac.id, derry@mdp.ac.id

Abstract: *Manual identification of chili leaf diseases has the weakness of subjectivity which impacts the decline in harvest productivity. This study aims to build an accurate automatic classification system using a Machine Learning approach. The research methodology integrates the extraction of Hue, Saturation, Value (HSV) color features and Gray Level Co-occurrence Matrix (GLCM) texture on a dataset of 1,856 images divided with a ratio of 80:20. Hyperparameter optimization was performed using Grid Search on the K-Nearest Neighbors (K-NN) algorithm to find the best performance. The test results show that the optimal configuration is achieved at a value of K = 3 with the Manhattan distance metric, which produces a test accuracy of 92%. It is concluded that the integration of color and texture features with appropriate parameter optimization is proven to be effective as a reliable and efficient diagnostic solution.*

Keywords: *Chili, Classification, GLCM, Grid Search, HSV, K-Nearest Neighbors.*

1. Introduction

Chili peppers play a vital role in Indonesia's culinary and economic landscape, supporting the incomes of millions of farmers and influencing the stability of national food prices. Therefore, in addition to chili farmers, many people also cultivate chili peppers [1, 11]. Despite their high economic value, chili peppers are susceptible to leaf diseases, which can significantly reduce productivity. According to data from Data Indonesia, Indonesian chili production reached 1.39 million tons in 2021, a decrease of 8.09% compared to 2020, when chili production reached 1.5 million tons [2, 12]. The decline in chili production stems from chili plant diseases. The numerous types of diseases affecting chili plants make it difficult for farmers to identify the type of disease affecting their plants [3]. Currently, disease detection at the farmer level still relies heavily on subjective and error-prone visual observations, making the need for objective and accurate diagnostic methods urgent [14].

Computer Vision technology offers a promising solution to this problem. Amid the dominance of Deep Learning approaches, classic Machine Learning methods remain relevant. The combination of Gray Level Co-occurrence Matrix (GLCM) texture feature extraction and K-Nearest Neighbors (K-NN) classification has proven effective, as in the research of Miftahul Rizky Pulungan (2024) who successfully applied the GLCM and K-NN methods to identify chili leaf diseases with an accuracy of 88.46% [4]. However, previous studies often struggle with lighting variations and confusing background noise which degrades segmentation accuracy. This study aims to develop and optimize the Hue, Saturation, Value (HSV) - Gray Level Co-occurrence Matrix (GLCM) - K-NN workflow. By conducting holistic optimization using Grid Search, this study is expected to validate the limits and maximum accuracy of the classical method's performance and prove its feasibility as a reliable alternative diagnostic solution for chili farmers in Indonesia.

2. Literature Review

Based on literature studies from previous research. Pulungan et al. (2024) applied GLCM and KNN to chili leaf objects, however, Pulungan et al.'s research (2024) focused on the texture features of chili leaves using Gray Level Co-occurrence Matrix (GLCM) and K-Nearest Neighbors (K-NN) for classification with an accuracy of 88.46% at K = 3 [4]. On the other hand, previous research by Zuain et al (2022) applied HSV color features and effectively addressed lighting variations with an accuracy of 86%, even though it used the Raspberry Pi-based C4.5 algorithm [5]. Through literature studies from previous research, this research has a good opportunity to combine HSV color features and GLCM texture features, as well as apply K-NN optimization comprehensively to chili leaf diseases.

3.Methods

In this study, a process was implemented to minimize errors and ensure each process runs effectively. The research methodology stages are outlined in Figure 1. Research Methodology Stages.

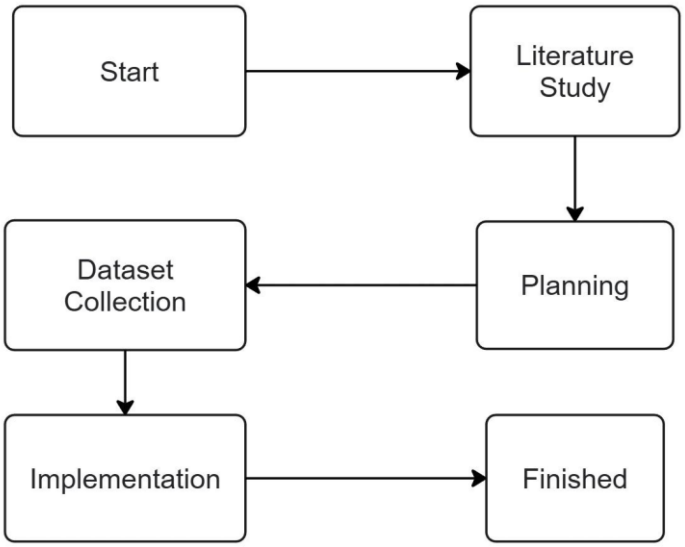

Figure 1. Research Methodology Stages

The dataset used in this study is the "Chili Plant Leaf Disease and Growth Stage Dataset from Bangladesh," published on the Mendeley Data platform by Nirob, M.A.S. et al. (2025)(https://data.mendeley.com/datasets/w9mr3vf56s/1). This dataset is specifically designed for research in the field of Computer Vision and Machine Learning in the agricultural context with a focus on chili plant disease detection. This dataset provides 1,856 original high-resolution images in .jpg format. These images were manually captured in a real plantation environment in Jamalpur, Bangladesh. Each image in the dataset has been labeled to enable precise classification tasks [6]. In this dataset, each class is unbalanced, therefore the division uses Stratified Random Sampling, with this division the proportion of each disease class in the training and test data is maintained consistently according to the proportion of the original dataset. This dataset is divided with a ratio of 80:20, namely 80% training data and 20% test data [15], and in addition during the training process, the K-Fold Cross-Validation method to maintain the validity of the model from unbalanced data. Table 1. The type of data collected is digital images divided into six different classes.

Table 1. Chili Leaf Disease Dataset

Disease Class	Picture
Bacterial Spot	
Curl Virus	
Cercospora Leaf Spot	
Nutrition Deficiency	

White Spot

Healthy Leaf

Image processing methods such as segmentation based on the Hue, Saturation, and Value (HSV) color space were chosen for their ability to separate color and light information, making them better suited to handling images with varying lighting conditions in the field compared to the Red, Green, and Blue (RGB) color space, where color and brightness are mixed together [7, 13]. Texture feature extraction using the Gray Level Co-occurrence Matrix (GLCM) was chosen for its ability to capture patterns on the surface of diseased chili leaves. The main features used are contrast, correlation, energy, and homogeneity [8]. The k-Nearest Neighbors (k-NN) classification algorithm was chosen for its working principle and application, as it only stores the entire training data [9, 10]. The optimization process will be carried out using Grid Search and K-Cross Validation with 5-Fold to ensure that the model is tested on all parts of the training data in turn, resulting in good and accurate results.

The first stage is the input dataset. The second stage is pre-processing, the first will be Resized to 256x256 pixels so that each image has the same dimensions, then Noise Reduction is performed to remove spots on the dataset image and the final stage, by increasing the local contrast in the image using Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2.0 and tile grid size of (8, 8) so that the details of the disease become clearer in the shadowed parts. The third stage is to segment leaf objects from the background by converting to HSV and making a mask on the leaf area, segmentation was performed using a custom algorithm named 'Mask V13' designed to handle lighting variations and soil background noise. The masking process involves two parallel streams: Healthy Area Mask: Defined in HSV space with a Green range of H=[20, 100], S=[25, 255], V=[25, 255] to capture chlorophyll-rich areas. Disease Spot Mask: Targeted bright/white lesions using HSV range H=[0, 180], S=[0, 50], V=[120, 255]. To prevent soil from being misclassified as disease spots (since soil can be bright), a LAB color space filter was applied, excluding pixels with A-channel > 127 (red/brown spectrum).The final binary mask is a bitwise OR combination of both streams, followed by morphological opening (3x3 kernel) and closing 7x7 kernel) to remove noise and fill holes within the leaf object.. The fourth stage is carried out by extracting color features by calculating

the Mean, Standard Deviation on each H, S, V channel and texture feature extraction Texture by calculating Contrast, Correlation, Energy, Homogeneity. The fifth stage is feature normalization using the Min-Max Scaler calculation with a range of 0 to 1, this feature normalization is done so that no feature dominates the K-NN calculation. The sixth stage is Optimization, Optimization will be carried out using Grid Search and the K-Cross Validation method with 5-Fold. The process is in Figure 2.

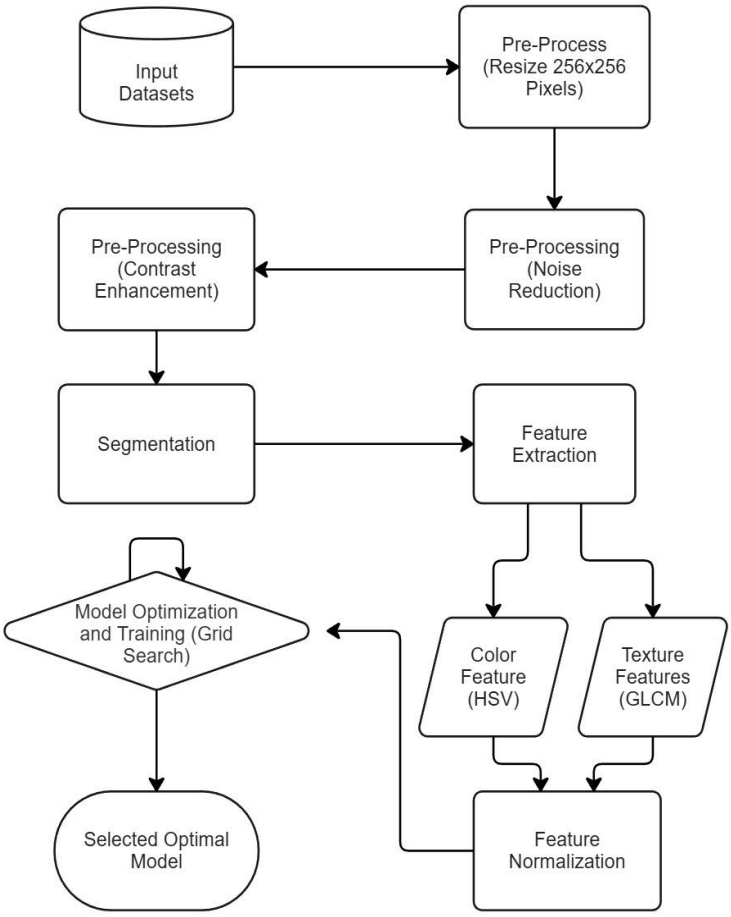

Figure 2. Model Development Workflow

4. Results and Discussion

The results of the pre-processing implementation can be seen in Figure 3. The first column is still in the image from the existing dataset. In the second column, contrast enhancement with CLAHE (Contrast Limited Adaptive Histogram Equalization) is used to even out the lighting and increase local contrast so that the texture of the disease spots is more visible when later used in the GLCM texture feature extraction input. In the third column, segmentation is carried out with Mask V13 which is useful for masking, namely by removing the background image and also capturing the leaf or disease area that is detected. In the third column (Mask V13), especially in Rows 1, 4, and 5, the binary mask effectively isolates the leaf area. Although the mask visualization appears dark due to the binary scaling display, the success of the segmentation is proven in the 'Final Result' column, which displays a perfectly intact leaf object with the background completely removed. However, the mask column V13 in rows 2 and 3 displays imperfect masking results, as can be seen in the 'Final Result' column, which shows black or hollow areas, caused by intense light reflection on the smooth leaf surface, causing the HSV value in that area to fall outside the established green leaf threshold range. However, most of the diseased texture areas can still be processed and are not a fatal problem.

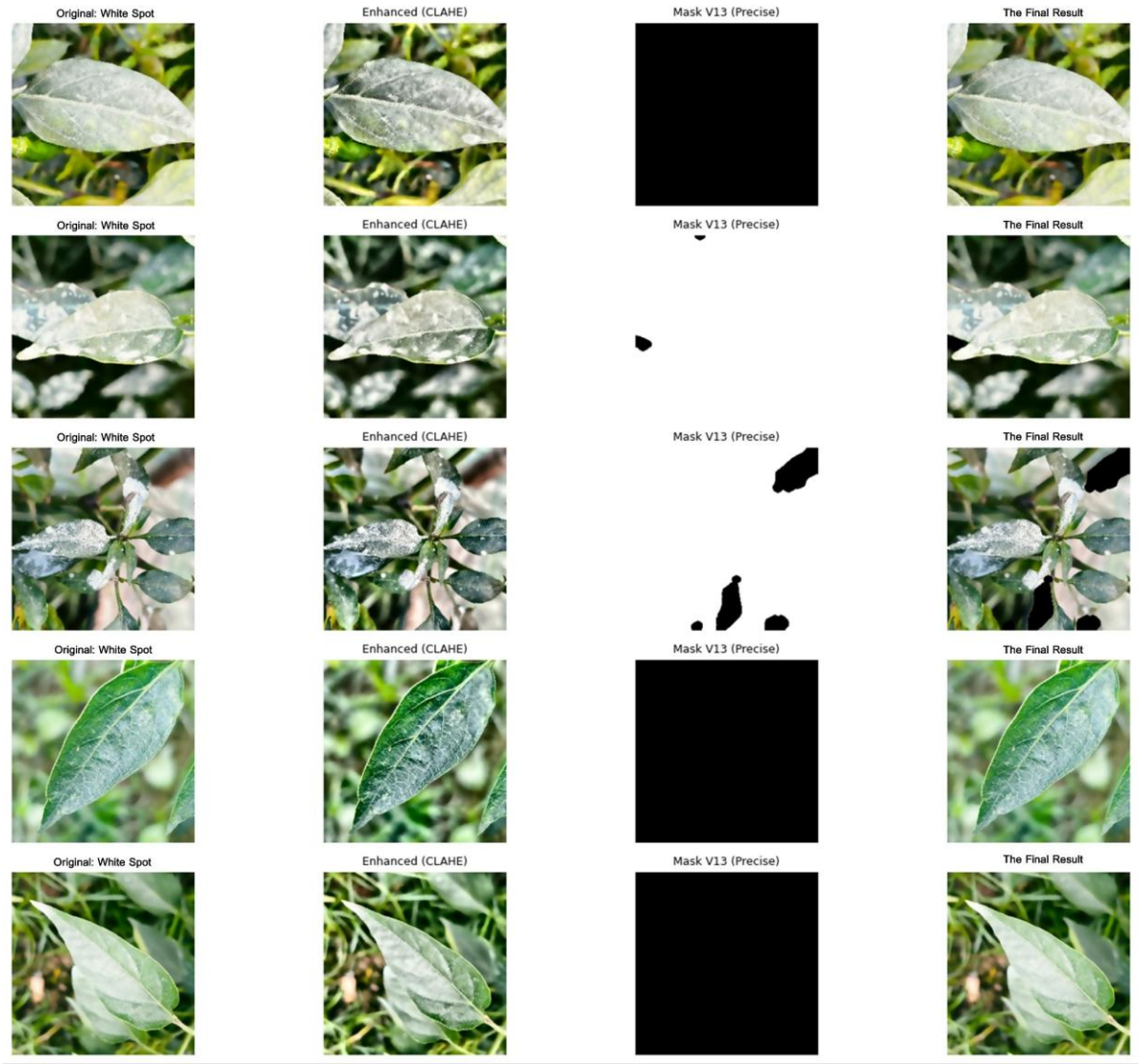

Figure 3. Image Pre-processing Results

Table 2. Grid Search Parameters

No.	Parameter	Tested Values
1	Number of Neighbors (k)	3, 5, 7, 9, 11, 13, 15
2	Distance Metric	Euclidean, Manhattan, Chebyshev
3	Weights	Uniform, Distance
4	GLCM Pixel Distance	1,2,3

To obtain the best model performance, this study applied Hyperparameter Tuning using the Grid Search method with parameters in Table 2. The tested odd value range from 3 to 15 was used because it is the equilibrium point. The three distance matrices were used to ensure unbiased pattern search. Both weight tests were conducted because there were several similar diseases and during the voting, no one was more dominant. GLCM pixel distance 1 was used to capture fine/small spot patterns on leaves and GLCM pixel distance 3 was used to capture block/wide area patterns on leaves.

Table 3. Classification results on test data

No.	Disease Class	Precision	Recall	F1-Score	Support
1.	*Bacterial Spot*	0.78	0.81	0.79	31
2.	*Cercospora Leaf Spot*	0.96	0.61	0.75	36
3.	*Curl Virus*	0.91	0.94	0.92	85
4.	*Healthy Leaf*	0.95	0.98	0.96	92
5.	*Nutrition Deficiency*	0.91	0.97	0.94	89
6.	*White Spot*	0.95	0.97	0.96	39

The classification results on the test data are in Table 3. The Healthy Leaf and White Spot classes demonstrate the highest detection performance, both achieving an F1-Score of 0.96. Conversely, the Cercospora Leaf Spot class shows the lowest detection performance with an F1-Score of 0.75. The Healthy Leaf class has the best detection performance with a Recall of 0.98 and the Cercospora Leaf Spot class has the lowest detection performance of 0.61. Regarding precision, the Cercospora Leaf Spot class achieves the highest score of 0.96, whereas the Bacterial Spot class records the lowest precision at 0.78. In the Support column which presents the number of datasets for each class used in the testing process, a total of 372 test data are divided into several data per class. For the Healthy Leaf class, Recall (0.98) is higher than Precision (0.95). This indicates the behavior of the model with high sensitivity. The model successfully captured 98% of the actual healthy leaves, minimizing the risk of 'False Negatives' (where healthy plants are mistakenly accused of being diseased). This is beneficial for farmers because it avoids unnecessary pesticide use on healthy plants. In contrast, for the Cercospora Leaf Spot class, Precision (0.96) is significantly higher than Recall (0.61). This indicates conservative model behavior. When the model predicts Cercospora, it is 96% likely to be correct (high confidence). However, the low Recall means the model fails to detect 39% of true Cercospora cases, often misclassifying them as visually similar diseases. This poses a risk in the field, as many infected plants may go undetected.

Table 4. Grid Search Hyperparameter Optimization Results

No.	Parameter	Best Value	Information
1.	Distance Matrix	Manhattan	More effective than Euclidean on high-dimensional data
2.	Number of Neighbors (K)	3	Capturing local patterns of disease while minimizing bias
3.	GLCM corner	[0°, 45°, 90°, 135°]	Using a combination of all angle directions
4.	Validation Accuracy	91.44%	Average validation accuracy on 5-Fold Cross Validation

Based on the parameter combination testing using Grid Search in Table 2. The best configuration was obtained with a value of K = 3 and the Manhattan distance metric. This configuration produced a validation accuracy of 91.44%. From Table 2. The use of four angles, namely 0°, 45°, 90°, 135°, did not produce different outputs, but combined with mean pooling.

This was done to minimize texture orientation bias in leaf images. The selection of the hyperparameter value K = 3 as the optimal configuration indicates that the characteristics of chili leaf disease have a very local and specific pattern. A small K value makes the mode maintain the sensitivity of micro-texture details in disease spots, which may be lost or outvoted if using a large K value (for example, K = 15). In addition, the dominance of the Manhattan distance metric over Euclidean shows its effectiveness in handling high-dimensional HSV and GLCM feature spaces. Manhattan is proven to be more robust in measuring than Euclidean in handling high-dimensional data because it does not impose excessive penalties on features that have extreme value differences. This characteristic makes the distance calculation more stable and representative of the visual conditions of the disease in the field compared to the Euclidean metric which tends to be sensitive to data deviations.

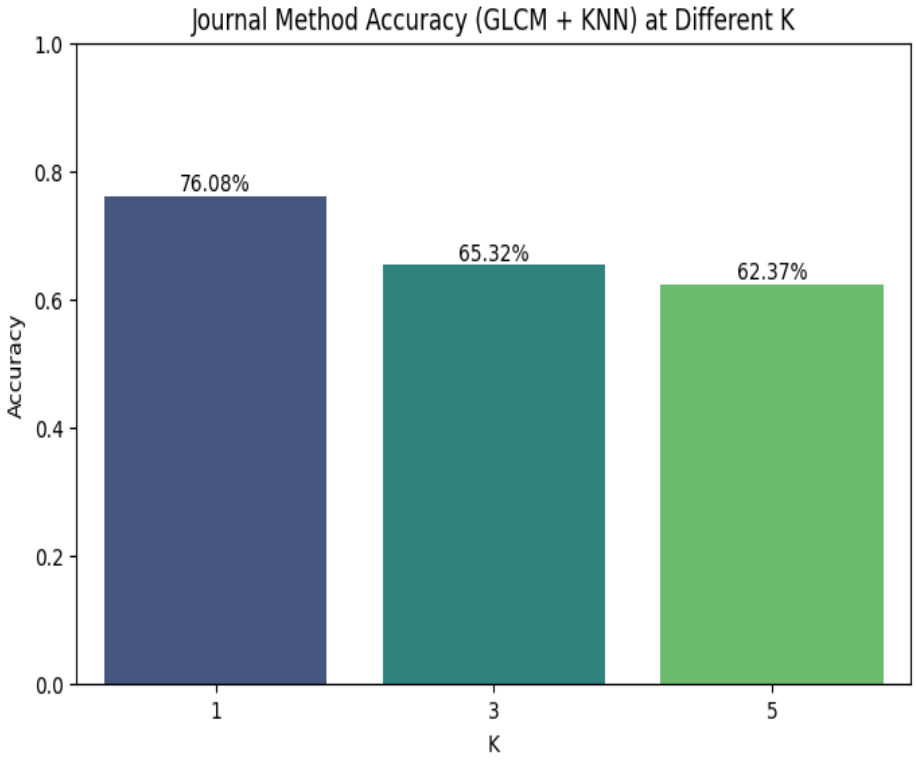

Figure 4. Accuracy results of GLCM and KNN models (Standard Method) [4].

As a comparison, this study implements a standard method referring to the research of Pulungan et al. (2024) applied to the field dataset. The stage begins by converting the image to a 400x400 pixel grayscale without color segmentation and GLCM angles [0°, 45°, 90°, 135°]. Classification is carried out using k-NN with Euclidean distance at k = 1, 3, 5. This step aims to validate whether the addition of HSV color features to the method applied in this study improves the accuracy of chili leaf disease objects. The results in Figure 4 explicitly show that the proposed method (91.44%) significantly outperforms the standard GLCM+KNN method (76.08%). This confirms that the integration of HSV color features provides an accuracy increase of +15.36%, proving that color information is very important for distinguishing chili leaf diseases that have similar textures.

5. Conclusions

Overall, the best accuracy obtained from the HSV and GLCM and KNN models is 0.9144 (91.44%) with the best parameter combination, namely classification matrix Manhattan, KNN = 3, GLCM angle = [0°, 45°, 90°, 135°], 5-Fold Cross-Validation. The results obtained from this method show more dominant results than the results from the GLCM and KNN models, namely with an accuracy of 0.7608 (76.08%) with K = 1 and a Euclidean matrix, 0.6532 (65.32%) with K = 3 and a Euclidean matrix, 0.6237 (62.37%) with K = 5 and a Euclidean matrix. The achievement of 92% test accuracy answers the fundamental problem of research regarding the limitations of subjective manual visual observation. The integration of HSV color features has been shown to handle lighting variability in plantations, while GLCM features with K = 3 effectively capture specific disease symptom patterns. Thus, the developed model has been validated as an objective and reliable diagnostic solution to minimize disease identification errors frequently encountered by chili farmers. The optimal configuration (K = 3 with Manhattan distance) yielded a validation accuracy of 91.44% during training and achieved a testing accuracy of 92% on a previously unseen dataset. However, this study is limited to the specific dataset from Bangladesh. Future work should verify the model on chili varieties common in Indonesia to handle domain shift.

References

[1] Hafidhoh, N. (2022). Identifikasi Penyakit Daun Tanaman Cabai Merah Dengan Ekstraksi Fitur Dan Klasifikasi Support Vector Machine. Prosiding Seminar Hasil Penelitian Dan Pengabdian Kepada Masyarakat (SEHATI ABDIMAS) POLTESA, 5(1), 64–74. https://ojs.poltesa.ac.id/index.php/SEHATI_ABDIMAS/article/view/434

[2] Rizaty, M. A. (2022, Juli 19). Produksi Cabai Rawit di Indonesia Turun 8,09% pada 2021(Webpage). Retrieved from DataIndonesia.id https://dataindonesia.id/sektor-riil/detail/produksi-cabai-rawit-di-indonesia-turun-809-pada-2021

[3] Irene Oktaviani Duka, Huan Arthur Ado, & Yampi R.Kaesmetan. (2024). Identifikasi Penyakit Tanaman Citra Daun Cabe Menggunakan Gray Level Co-Occurrencce Matrix Dan Support Vector Machine. Neptunus: Jurnal Ilmu Komputer Dan Teknologi Informasi, 2(2), 41–52. https://doi.org/10.61132/neptunus.v2i2.86

[4] Pulungan, M. R., Furqan, M., & Rifki, M. I. (2024). Klasifikasi Penyakit Pada Daun Cabai Menggunakan Gray Level Co-Occurrence Matrix Dan K-Nearest Neighbor. Syntax : Journal of Software Engineering, Computer Science and Information Technology, 5(2), 549–554. https://doi.org/10.46576/syntax.v5i2.5386

[5] S. S. Zuain, H. Fitriyah, & R. Maulana. (2021). Deteksi Penyakit pada Daun Cabai berdasarkan Fitur HSV dan GLCM menggunakan Algoritma C4.5 berbasis Raspberry Pi. Pengembangan Teknologi Informasi Dan Ilmu Komputer, 5(9), 3934–3940. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/9791

[6] Nirob, M. A. S., SIAM, A. K. M. F. K., Bishshash, P., & Assaduzzaman, M. (2025). Chili Plant Leaf Disease and Growth Stage Dataset from Bangladesh (Version 1). Mendeley Data. https://doi.org/10.17632/w9mr3vf56s.1

[7] Nurmadinah, N., Wajidi, F., & Arifin, N. (2025). Deteksi Penyakit Daun Cabai Menggunakan Kombinasi Glcm Dan Hsv Dengan Klasifikasi Svm. Insect (Informatics and Security): Jurnal Teknik Informatika, 11(2), 178–189. https://doi.org/10.33506/insect.v11i2.4820

[8] Patil, A., & Lad, K. (2022). Feature Selection for Chili Leaf Disease Identification Using GLCM Algorithm. In IOT with Smart Systems (pp. 555–564). Springer. https://doi.org/10.1007/978-981-16-3942-8_53

[9] Chauhan Pareshbhai Mansangbhai. (2024). Chili Disease Detection Using HOG with Euclidean Distance. Journal of Electrical Systems, 20(3), 1577–1584. https://doi.org/10.52783/jes.3654

[10] Ni'mah, F. S., Sutojo, T., & Setiadi, D. R. I. M. (2018). Identification of Herbal Medicinal Plants Based on Leaf Image Using Gray Level Co-occurence Matrix and K-Nearest Neighbor Algorithms. Jurnal Teknologi Dan Sistem Komputer, 6(2), 51–56. https://doi.org/10.14710/jtsiskom.6.2.2018.51-56

[11] Royun Nuha, M., Andita Putri, T., & Dwi Utami, A. (2023). Pendapatan Usahatani Cabai Merah Berdasarkan Musim di Provinsi Jawa Tengah. Jurnal Ilmu Pertanian Indonesia, 28(2), 323–334. https://doi.org/10.18343/jipi.28.2.323

[12] Pusdatin Kementan, 2023. (2023). OUTLOOK Cabai Pusat Data dan Sistem Informasi Pertanian. Pusat Data Dan Sistem Informasi Pertanian Sekretariat Jenderal-Kementerian Pertanian 2022, i–62.https://satudata.pertanian.go.id/assets/docs/publikasi/OUTLOOK_CABAI_2023_berbarcode_.pdf

[13] Rahmadewi, R., Sari, G. L., & Firmansyah, H. (2019). Pendeteksian Kematangan Buah Jeruk Dengan Fitur Citra Kulit Buah Menggunakan Transformasi Ruang Warna HSV. JTEV (Jurnal Teknik Elektro Dan Vokasional), 5(1.1), 166. https://doi.org/10.24036/jtev.v5i1.1.107560

[14] Zikra, F., Usman, K., & Patmasari, R. (2021). Deteksi Penyakit Cabai Berdasarkan Citra Daun Menggunakan Metode Gray Level Co-Occurence Matrix Dan Support Vector Machine. Seminar Nasional Hasil Penelitian Dan Pengabdian Masyaraka, 1, 105–113. https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2920

[15] Syarif, R. S. (2023). Implementasi Color Moments, Grey Level Co-occurrence Matrix (GLCM), dan Random Forest untuk Mendeteksi Penyakit Daun Tomat. Skripsi, Universitas Islam Negeri Alauddin Makassar. http://repositori.uin-alauddin.ac.id/24535/

Acknowledgements

The author expresses his gratitude to Allah SWT for all His grace and blessings, enabling this research to be successfully completed. He expresses his deepest gratitude to Universitas Multi Data Palembang for the facilities and academic support provided. He also extends his deepest appreciation and gratitude to Mr. Derry Alamsyah, his supervisor, for his valuable guidance, constructive discussions, and technical guidance throughout the research and writing process. He also thanked Nirob et al. (2025) for their contribution in providing the public dataset "Chili Plant Leaf Disease and Growth Stage Dataset from Bangladesh," which served as the foundation for this study. Finally, he thanks his parents and colleagues for their continued moral support and motivation.