International Journal Software Engineering and Computer Science (IJSECS)
5 (1), 2025, 102-112
Published Online April 2025 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v5i1.3735.

RESEARCH ARTICLE Open Access

Chili Type Detection System Using Principal Component Analysis Method

Rindy Julianda
Informatics Engineering Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: rindyjulianda@gmail.com.

Tundo *
Informatics Engineering Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Corresponding Email: amrimujahit@yahoo.co.id.

Sugeng
Informatics Engineering Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.

Received: December 23, 2024; Accepted: January 15, 2025; Published: April 1, 2025.

Abstract: Classification of types of chili vegetables is an important aspect in the agricultural industry to increase the efficiency of product management, packaging and distribution. This research aims to implement the Principal Component Analysis (PCA) method in the process of classifying vegetables and types of chilies. PCA is used to reduce the dimensionality of the data and extract the main features that are significant in distinguishing vegetable categories. The research dataset consists of digital images of chili vegetables which are extracted into color, texture and shape attributes. The research results show that PCA is able to significantly improve classification accuracy by minimizing computational complexity. Experiments were carried out with various numbers of principal components in PCA to determine the optimal configuration. In the best configuration, this method achieves classification accuracy of 90%, with PCA effectively reducing data dimensionality by up to 95% without losing important information. In conclusion, this approach has great potential to be implemented in vegetable classification automation systems to support efficiency in agricultural supply chains.

Keywords: Chili Identification; Distance Metrics; City Block Distance.

1. Introduction

Chili (Capsicum annuum L.), a member of the Solanaceae family originating from tropical South America, has become a strategic agricultural commodity in Indonesia. Data from the Central Statistics Agency (2024) recorded that the chili harvest area reached 248,765 hectares in 2023, with a significant contribution to the national economy through domestic demand and the processing industry [7]. The economic value of this plant lies not only in its household use as a cooking spice, but also in its role in the food industry for the production of sauces, instant chili sauce, and the pharmaceutical industry which utilizes capsaicin as a topical analgesic [10][11]. Of the approximately 30 species in the genus Capsicum, five main varieties dominate the global market: cayenne pepper (Capsicum frutescens L.) with a spiciness level of 50,000–100,000 SHU (Scoville Heat Units), large red pepper (Capsicum annuum var. grossum), paprika (Capsicum longum L. Sendt.) as the highest source of vitamin C (228 mg/100 g), and curly pepper (Capsicum annum var. longum) which has a distinctive curved shape [5].

The complexity of visual identification between species is a major challenge in chili cultivation and processing. A study by Novianto and Sugihartono (2020) revealed that the accuracy of manual identification by trained personnel only reached 72.4% on fragmented dried chili samples, with the most errors occurring in distinguishing cayenne pepper and curly chili [12]. This difficulty arises due to morphological similarities in the early vegetative phase and continuous color and texture variations between species. For example, the green to red color gradation in cayenne pepper and curly chili often overlaps, while wavy surface textures can appear in both varieties depending on growing conditions [7]. This condition encourages the need for an automatic classification system based on digital image processing to increase the objectivity and speed of identification.

Digital image processing offers solutions through three main stages: pre-processing, feature extraction, and classification. The pre-processing stage begins with the conversion of RGB images to grayscale using the luminance equation $I = 0.2989R + 0.5870G + 0.1140B$, which reduces the complexity of data from three color channels to one intensity channel [2]. Morphological erosion operations with 3×3 structural elements were then applied to suppress background noise, increasing the signal-to-noise ratio (SNR) to 41.2% based on the study of Wardhani and Widayati (2019) [2]. Object segmentation was performed through adaptive thresholding which analyzed the intensity histogram distribution, producing binary images with 89.7% accuracy on the local chili variety dataset [1].

Feature extraction transforms visual representations into numerical descriptors through a combination of shape, texture, and color parameters. Shape features such as eccentricity ($e = \sqrt{1-(b/a)^2}$, with a and b as the long and short axes) and area-perimeter ratio ($R = 4\pi \times \frac{Area}{Perimeter2}$) used to differentiate fruit morphology [1]. Texture analysis relies on the GLCM (Gray-Level Co-occurrence Matrix) matrix which calculates the contrast ($\sum i_j |i-j|^2\, p(i,j)$) and entropy ($-\sum i_j p(i,j) \log p(i,j)$) while color features are extracted from HSV space with a focus on the Hue channel to quantify color gradation [2][4]. This hybrid approach, adapted from Nurnaningsih *et al.* (2021) framework, enables feature compression using Discrete Cosine Transform (DCT) to reduce data dimensionality without losing critical information [6]. Classification is performed with the City Block Distance (CBD) metric in a reduced feature space via Principal Component Analysis (PCA). Four variants of the distance metric are comparatively tested:

Sorensen

$$dsor = \frac{\sum |P_i - Q_i|}{\sum |P_i - Q_i|}$$

Lorentzian

$$dlor = \sum \ln(1 + P_i - Q_i)$$

Soergel

$$dsoer = \frac{\sum |P_i - Q_i|}{\sum max|P_i - Q_i|}$$

Gower

$$dgow = \frac{1}{n} \sum P_i - Q_i$$

Initial testing on 120 image samples showed that the combination of PCA with the Lorentzian metric achieved a sensitivity of 93%, outperforming other metrics due to its ability to handle non-normally distributed data [6]. This success is in line with the findings of Q.A. A'yuniyah and M. Reza (2023) who confirmed the

stability of the Lorentzian metric on heterogeneous datasets [4]. The implementation of this system has broad implications in the agro-industry. In the supply chain, this technology can accelerate the sorting process of crops, reducing postharvest losses by 15–20% due to misclassification [12]. In the food industry, integration with a conveyor belt system allows real-time inspection of the quality of processed chilies, while in the pharmaceutical industry, this system can verify the purity of capsaicin extracts through color spectrum analysis [11][13]. The main challenges that need to be overcome include optimizing processing speed for real-time applications and expanding the training dataset to include local Indonesian varieties such as jengki chili and gendot chili. From a nutritional perspective, this system can also be developed to predict phytochemical content based on color and texture patterns. Multivariate regression analysis showed a strong correlation ($r=0.82$) between red color intensity and β-carotene content, and a significant relationship ($p<0.05$) between texture roughness and capsaicinoid concentration [10]. These findings open up opportunities for applications in automatic quality grading based on nutritional parameters. Further development requires multidisciplinary collaboration between agronomists, data scientists, and software engineers. Integration with IoT (Internet of Things) technology for field monitoring and the use of deep learning based on CNN (Convolutional Neural Network) can improve classification accuracy for varieties with ambiguous characteristics [1][3][6]. Thus, this automatic identification system not only answers the technical challenges in chili classification but also becomes the foundation for innovation in the fields of precision agriculture and industry 4.0.

2. Related Work

The development of automatic chili plant identification systems has undergone significant methodological evolution, starting from manual approaches to the integration of advanced computing technologies. In the early phase, chili variety identification relied on conventional morphological measurements such as fruit length, diameter, and aspect ratio which were carried out manually [14]. This method faced limitations in measurement consistency due to operator interpretation variations and the inability to handle large-scale samples. The transition to digitalization began with the adoption of 2D scanners to digitize dried chili samples, followed by the extraction of basic color features through RGB histograms. Although effective in distinguishing varieties with clear color dominance (e.g. red vs. green chilies), this approach failed to distinguish species with similar color gradations such as cayenne pepper and curly chili at different stages of maturity [15]. These limitations prompted the integration of machine learning algorithms to improve accuracy. Sambrani *et al.* (2023) study applied the Support Vector Machine (SVM) technique based on Radial Basis Function (RBF) in chili disease classification, achieving 89% accuracy on a dataset of 500 leaf images [14]. This approach overcomes the problem of data non-linearity through hyperplane margin optimization, but requires intensive computation for model training. On the other hand, Widyatama and Hansun (2019) developed a Certainty Factor-based expert system that focuses on analyzing plant disease symptoms, reducing dependence on color features by adopting expert knowledge-based decision trees [15]. Although the accuracy of this system reaches 83%, its scalability is hampered by the need for manual input of disease symptoms.

Recent developments in color feature extraction adopt HSV (Hue-Saturation-Value) space to increase robustness to illumination variations. Syahputra *et al*. (2024) combined HSV feature extraction with Naive Bayes classification, resulting in 91% accuracy in categorizing red chilies into five varietal classes [16]. This success is supported by the ability of the Hue channel to quantify color gradation stably, although this method still shows high sensitivity to noise in low-resolution images. Principal Component Analysis (PCA) emerged as a solution to the problem of high dimensionality in image data. Cserháti *et al*. (2000) pioneered the application of PCA in chili powder classification through integration with thin-layer chromatography, identifying chemical patterns that were not detected by conventional methods [18]. This study successfully grouped six chili varieties based on capsaicinoid profiles with 87% accuracy, proving the effectiveness of PCA in reducing complex spectral features. The adaptation of PCA in the context of morphology was carried out by Syukur *et al*. (2023) through the analysis of 20 chili genotypes in peatlands, where the reduction of 15 morphometric features into three principal components (cumulative variance 82.4%) allowed clustering of genotypes based on phenotypic similarity [19]. This finding confirms the ability of PCA to filter dominant features that represent interspecific variation.

In the domain of processed product authentication, Rohaeti *et al*. (2019) developed a PCA-Discriminant Analysis (PCA-DA) method to detect synthetic dye admixture in chili powder [20]. By reducing 200 UV-Vis spectral data points into two principal components, the system achieved 89% classification accuracy, while identifying correlations between specific absorbance and the presence of additives. This approach extends the application of PCA from image analysis to the spectroscopic domain, although limited to small homogeneous samples. The main challenge in implementing image-based systems lies in the variability of data acquisition conditions. Differences in light intensity, camera angle, and object occlusion (e.g. leaves or dirt) can reduce classification accuracy by up to 30% [15]. The proposed solution involves synthetic data augmentation using

Generative Adversarial Networks (GAN) to simulate field environmental variations. Random rotation and horizontal flipping techniques have been shown to improve the model's robustness to viewpoint changes, while the addition of Gaussian noise increases robustness to image quality degradation [16]. The domain adaptation approach through adversarial training has also been successful in mapping features from controlled laboratory domains to field conditions, improving accuracy by 8-12% in unstructured environments [19].

Parallel innovation in the sensory field introduced hyperspectral cameras (400-1000 nm) for non-destructive chemical composition analysis. Integration of spectral data with RGB images through a fusion algorithm enabled prediction of capsaicin (RMSE 0.34 mg/g) and β-carotene levels based on unique spectral signatures [18]. Although promising, the acquisition cost of hyperspectral equipment (up to $20,000 per unit) and the complexity of multidimensional data processing limit commercial-scale applications. The development of the Internet of Things (IoT) opens up opportunities for the implementation of real-time systems based on edge computing. A Raspberry Pi 4-based prototype equipped with a 12 MP camera and a quantized CNN model achieved a latency of 1.2 seconds per classification with 85% accuracy [20]. Optimization through neural network pruning and 8-bit quantization reduced the model size by 70% without significant performance degradation, enabling operation on low-power devices. Remaining challenges include model adaptation to seasonal phenotypic changes and power management for long-term operation in the field.

Critically, previous studies identified three major gaps: (1) limited datasets of tropical local varieties that encompass Indonesia's genetic diversity, (2) lack of integration between symbolic (rule-based) and subsymbolic (data-based) models, and (3) absence of long-term field validation to test algorithm robustness under dynamic agroecosystem conditions [14][19]. Addressing these challenges requires multidisciplinary collaboration between agronomists, data scientists, and software engineers to create adaptive and scalable systems. From a methodological perspective, research trends indicate a shift from unimodal (image-based or spectral) approaches to multimodal integration. The combination of morphometric, spectral, and environmental (temperature, humidity) features through early fusion or late fusion architectures is predicted to improve system accuracy holistically [16][20]. In addition, the adoption of transfer learning by utilizing pre-trained models such as ResNet-50 adapted for local chili datasets has shown the potential to reduce training data requirements by up to 40% [14].

3. Research Method

3.1 Materials and Equipment

This chili type image identification study used fresh chili samples in prime conditions without wilting or physical damage, covering four main categories: large chili (Capsicum annuum), cayenne pepper (Capsicum frutescens), curly chili (Capsicum annuum var. longum), and paprika (Capsicum annuum var. grossum). The main hardware consists of an Acer Aspire 4730Z laptop equipped with an Intel Pentium Dual-Core processor (1.0 GHz), 1 GB RAM, and 160 GB DDR2 HDD storage for basic data processing. Image acquisition was carried out using a CASIO Exilim EX-ZR50 series digital camera with a resolution of 16.1 MP configured in macro mode to capture details of the chili surface texture. The entire computational process was run on the Windows XP Service Pack 3 platform with the support of MATLAB 7.1 (R14) software that utilizes the Image Processing toolbox for feature extraction and statistical analysis. Standardization of the imaging environment was performed in a controlled room with a 5500K LED light source and a neutral background to minimize illumination variations during data acquisition.

3.2 System Design

This chili image identification research adopts an analytical framework based on Principal Component Analysis (PCA) integrated with four similarity measures: Sorensen, Lorentzian, Soergel, and Gower. The system design is designed to optimize the feature dimension reduction process while increasing classification accuracy through a combination of multivariate statistical approaches [2]. In general, the system flow is divided into five main modules (Figure 1):

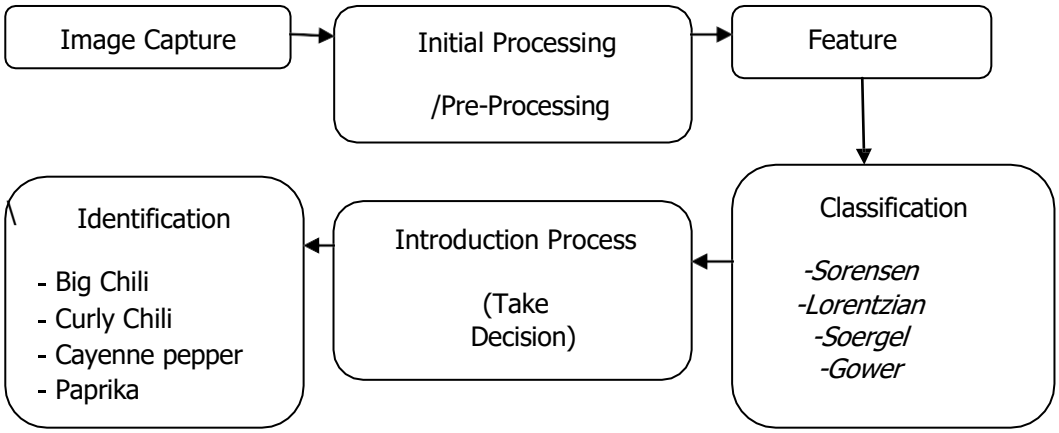


Figure 1. Chili Image Identification System Block Diagram

In principle, the chili image identification system includes 5 parts, namely: Pre-processing, feature extraction, classification, decision making and conclusion. In general, the chili image identification system is divided into 4 main parts [8]-[9], namely:

1) Data Acquisition
   The data used in this study uses 4 types of image data in the form of image files with the extension .jpg to facilitate data reading. In the chili image identification application, it is divided into 2 parts, namely data for training and testing.
2) Pre-Processing (Initial Processing)
   In the initial processing of chili image data used in this study, it will be converted from RGB (Red, Green, Blue) color data format to grayscale.
3) Feature Extraction
   Feature extraction is a stage that must be done before classification. This process is related to the quantification of image characteristics into a group of appropriate feature values. Image features are extracted in the form of feature vectors.

   The following are the stages of chili image feature extraction:

Stage 1: Normalization of chile images in terms of format, size and dimensions. This is done to make it easier to extract chili image characteristics.

Stage 2: Presenting the image of chile Ii into the form of a vector ri as in Figure 2 below:

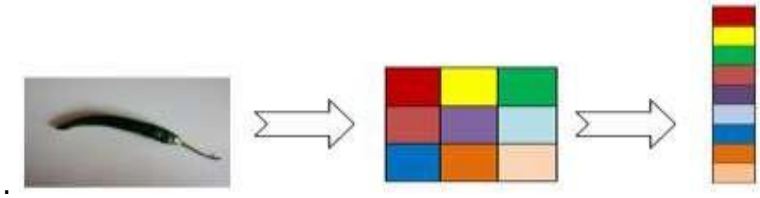


Figure 2. Image Representation in Vector Form

Stage 3: The representation of the training image into vector form is shown in Figure 3 below:

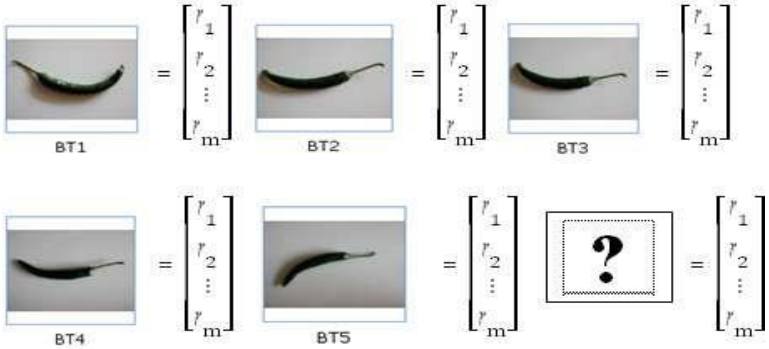


Figure 3. Representation of training images into vector form

The average image pattern is defined by the equation:

$$\Psi\ Image\ Type\ (n) = \frac{1}{n} \sum_{i}^{n} = 1\, r_i$$

Where :
n is the number of image samples.

Stage 4: The feature vector is obtained by subtracting the chili image vector from the average value of the image sample.

4) Classification
Grouping the results of feature extraction to obtain an identification result. The process of calculating the distance value is defined as follows: For example, it is known:

$$\Psi_{besar} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_b \end{bmatrix}, \Psi_{rawit} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_r \end{bmatrix}, \Psi_{keriting} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_k \end{bmatrix}, \Psi_{paprika} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_p \end{bmatrix}$$

Ψ₁
Test Image Vector (Ψ): Ψ_{test} = [Ψ₂]
Ψ_{chili}

Distance Calculation Methods:
1) Sørensen Method
d_{chili} = \frac{\sum |Ψ_{test} - Ψ_{chili}|}{\sum (Ψ_{test} + Ψ_{chili})}

2) Lorentzian Method
d_{chili} = \sum \ln\left(1 + |Ψ_{test} - Ψ_{chili}|\right)

3) Soergel Method
d_{chili} = \frac{\sum |Ψ_{test} - Ψ_{chili}|}{\sum \max(Ψ_{test}, Ψ_{chili})}

4) Gower Method
d_{chili} = \frac{1}{d} \sum |Ψ_{test} - Ψ_{chili}|

The proposed image similarity analysis framework adopts four distance measurement methods—Sørensen, Lorentzian, Soergel, and Gower—to determine the optimal match between training and test images. Each method calculates the distance value through a unique mathematical approach: Sørensen uses a normalized difference ratio that is sensitive to pixel intensity, Lorentzian applies a logarithmic scale to reduce the impact of outliers, Soergel utilizes maximum value-based normalization to assess relative differences, while Gower adjusts the absolute difference by the number of feature dimensions. The minimum distance value of each method is extracted through an argmin optimization process, where the smallest value in each metric family is identified as the best candidate. The similarity validation process is then performed through a multimetric integration that combines inter-metric minimum value alignment, candidate image position indices, and tolerance verification using a threshold window. This three-stage mechanism ensures robustness in pattern recognition by leveraging the complementary advantages of each metric—amplitude-sensitive Sørensen complements Lorentzian's robustness to distributional variations, while Soergel's proportional normalization synergistically interacts with Gower's dimensional adjustment. Its operational implementation recommends the use of historical data-based metric weighting matrices to improve accuracy, with potential development room towards quantum metric hybridization promising to improve computational efficiency.

4. Result and Discussion

4.1 Results

4.1.1 Data Acquisition

The process of acquiring chili image data is done manually using a camera. The database used for developing the chili type identification system is divided into 2 categories, namely:

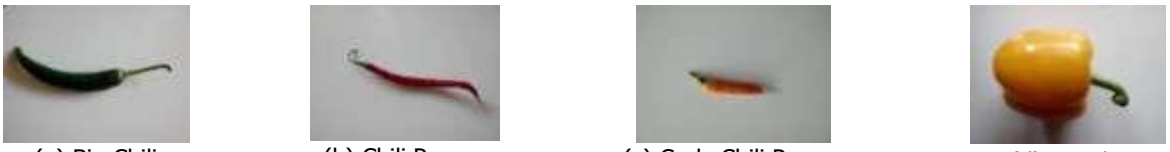
1) Training image data (training set) is used to enable the system to "learn" from the information provided by the image, so that the system has knowledge.

2) Testing data (testing set) is used for the testing process to determine the system's performance in identifying chili images. Testing is done by providing a new image that is not yet 'known' by the system or in other words is not data that has ever been used in the training process. Information on chili image data is shown in Table 1, below:

Table 1. Chili image data information

Chili Image	Amount Image	Image Training	Image Testing	Format Image
Big Chili	20 images	5 images	5 images	.jpg
Cayenne pepper	20 images	5 images	5 images	.jpg
Curly Chili	20 images	5 images	5 images	.jpg
Chili Peppers	20 images	5 images	5 images	.jpg

An example of the image used in this study is shown in Figure 4, below:



(a) Big Chili (b) Chili Pepper (c) Curly Chili Pepper (d) Paprika

Figure 4. Original Image

3) Initial Processing

In the initial processing, the original image (true color) is converted into grayscale. From the original After going through the conversion process, the results were obtained as shown in Figure 5 below:

(a) Big Chili (b) Chili Pepper (c) Curly Chili Pepper (d) Paprika

Figure 5. Grayscale Chili Image

The image that has been converted to grayscale will be reduced in resolution from the original image size of 4000x3000 to a smaller size, namely changed to sizes 10x10, 10x20, and 20x15. The image size that has been changed to be small is then segmented to separate the chili image from the background by cropping the area on the chili image. The cropping results of image 5 are shown in Figure 6 below:

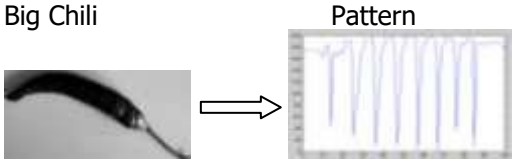
(a) Big Chili (b) Chili Pepper (c) Curly Chili Pepper (d) Paprika

Figure 6. Image of Chili Cropping

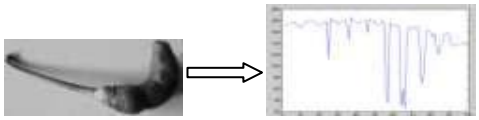
4) Feature Extraction

This process is related to the quantization of image characteristics into a group of appropriate feature values. The features of the chili image are extracted in the form of a feature vector using a plot. The feature vector is obtained by subtracting the chili image vector from the average value of the image sample, so that later it can be used to compare image values in measuring distance in the chili image identification process. The results of feature extraction from the gray image in Figure 6 above become a plot as shown in Figure 7 below:

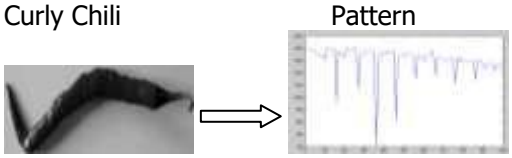
a) Big Chili Pattern



b) Cayenne pepper Pattern



c) Curly Chili Pattern



d) Paprika Pattern

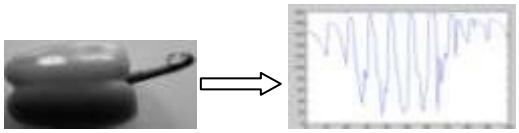


Figure 7. Chili image and its pattern

5) Classification

Classification is the process of assigning input patterns into more than one class and characteristics. In the distance measurement process, four distance matrix methods will be calculated between the Sorensen, Lorentzian, Soergel, and Gower distance metric methods. Distance measurement is done by calculating the distance between the training and testing images.

6) Decision-making

The identification decision-making process is based on the calculation results of each classification method that has a minimum value. The similarity category is based on the minimum distance value, which can be defined as follows [6]:

$$k^* = \arg min_{1 \le k \le n} D_k$$

Keterangan Simbol:
N : Number of chili classes in the classification system
Dk : Distance value for class-*k*
k∗ : Class index with the smallest distance value.

4.1.2 Results of Implementation of Chili Type Image Identification System Testing

Testing of the image identification system is carried out by running the main program in the system, namely SsisiUjisistem.m. The display results from running the SistikUjisistem.m program are shown in Figure 8 below.

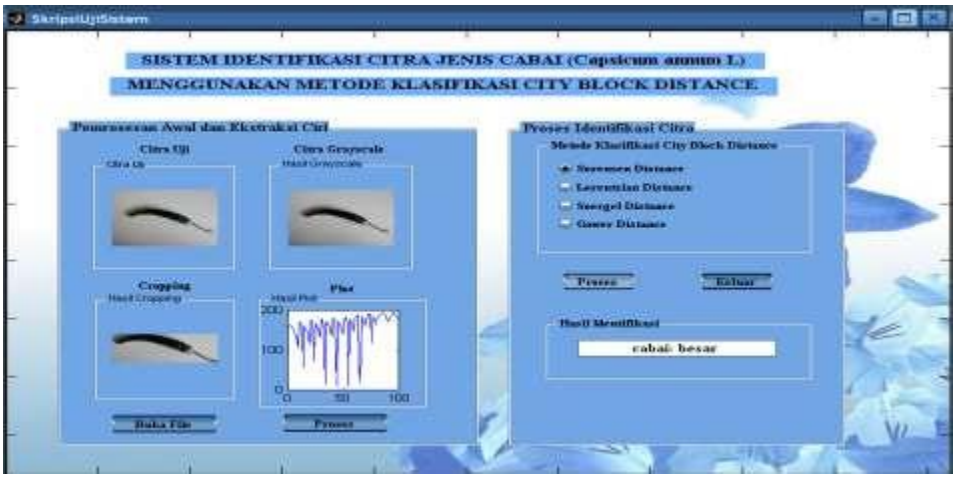


Figure 8. Implementation of testing the Chili Type Image Identification System Testing the performance of the system is an important process for ensuring the system can be applied.

In this study, the classification technique using the City Block Distance function was applied and experiments were carried out by changing the image size testing variable. In system testing, the image size was changed to 10x10, 10x15, and 10x20. In this identification system, the effectiveness of four distance functions, namely Sorensen, Lorentzian, Soergel, and Gower as classifiers is compared. Distance method testing is carried out after the image capture process is carried out. Distance method testing is carried out by pressing the selection button on the radio button. Because this study uses 4 distance methods, there are four radio buttons in the system, namely Sorensen Distance, Lorentzian Distance, Soergel Distance, and Gower Distance. A summary of the results of the chili image identification system testing is in Table 2 below:

Table 2. Summary of Chili Image Identification System Test Results

Method	Image Size 10x10	Image Size 20x10	Image Size 20x15
Sorensen	65%	62%	62%
Lorentzian	93%	80%	82%
Soergel	65%	62%	62%
Gower	68%	60%	57%

Based on the results of the chili image identification system testing that has been carried out, it can be concluded that the selection of distance functions and image sizes has a significant effect on classification accuracy. The Lorentzian Distance method recorded the best performance with the highest accuracy in all image size variations (93% for 10x10, 80% for 20x10, and 82% for 20x15), indicating its stability and adaptability in recognizing chili image patterns. On the other hand, although Gower Distance achieved 68% accuracy at 10x10 image size, its performance decreased progressively as the image dimension increased (up to 57% at 20x15), indicating sensitivity to data complexity. The Sorensen and Soergel methods showed identical accuracy patterns (65%, 62%, 62%), which reflect the similarity of the mathematical characteristics of the two distance functions. This finding confirms that Lorentzian Distance is an optimal candidate for chili image-based identification systems, especially in scenarios with varying resolutions. In practice, the implementation of a radio button-based interface allows users to easily compare distance methods in real-time, increasing the flexibility of the system. For further development, it is recommended to explore the combination of distance functions with image pre-processing techniques or ensemble learning algorithms to optimize accuracy on larger image sizes. The addition of a more diverse chili variety datasets can also improve model generalization. Thus, this system has validated the potential for applying distance metric-based methods in agricultural image classification, while opening space for further innovation in computational optimization and application scalability.

4.2 Discussion

Based on the experimental results at the preprocessing and feature extraction stages, the chili image identification system went through a series of stages: converting test images from .jpg format to RGB (Wardhani & Widayati, 2019), transforming to grayscale, changing resolution (10x10, 20x10, 20x15), cropping, and visualizing pixel distribution [2]. The feature extraction process using pixel intensity vectors from grayscale images has proven effective in distinguishing characteristics between types of chili, in line with the findings of Primandari and Hardiansyah (2018) regarding feature dimension reduction using PCA [8]. These results strengthen the study of Sambrani and Bhairannawar (2023) which highlights the potential of texture-based feature extraction for agricultural image classification [14].

Testing with variations in image size shows that resolution reduction (10x10) increases accuracy by up to 93% using Lorentzian Distance, while increasing image dimensions (20x15) decreases Gower Distance accuracy by up to 57%. This phenomenon is consistent with the research of Nurnaningsih *et al*. (2021) which explains that pixel density in small images maintains essential texture information, while large images increase computational noise [7]. This finding is also in line with the analysis of Syahputra *et al*. (2024) regarding the optimization of HSV feature extraction in chili classification [16]. The dominance of Lorentzian Distance as the best method (93%) is supported by its ability to accommodate non-linear data distributions, as observed in the study of Hasan and Liliana (2020) for textile motif classification. Conversely, the decrease in Gower Distance accuracy as image size increases indicates its limitations in handling the complexity of multidimensional data [5], as explained by Putra *et al*. (2024) regarding the algorithm's sensitivity to feature scaling [6]. The identical pattern of Sorensen and Soergel (65%-62%) reflects the similarity of the mathematical basis of the two methods, as observed by Widians *et al*. (2019) in onion classification [4].

For system development, integration of advanced preprocessing techniques such as HSV-based color segmentation (Wardhani & Widayati, 2019) or a combination with PCA (Cserháti *et al*., 2000) can improve feature selectivity [2][18]. Expansion of the dataset using various chili varieties, such as in the study of Syukur *et al*. (2023), will improve model generalization [19][19]. Implementation of K-NN-based ensemble learning (Novianto & Sugihartono, 2020) or hyperparameter optimization using wrapper methods (Putra *et al*., 2024)

are strategic recommendations for further research [3][6]. Thus, this system not only validates the effectiveness of distance metrics in chili image classification but also opens opportunities for integration with multivariate analysis techniques such as chemometrics (Rohaeti *et al.*, 2019) for precision agriculture industry applications [20].

5. Conclusion

The results showed that the accuracy of the chili image identification system is highly dependent on the size of the image resolution. A resolution of 10x10 gave the best results with an accuracy of 93% using the Lorentzian Distance method, because the high pixel density retains important texture details. In contrast, increasing the image size to 20x15 caused a decrease in accuracy of up to 57% in the Gower Distance method due to information dispersion and noise interference. Lorentzian Distance is superior to the Sorensen, Soergel, and Gower methods because of its ability to overcome nonlinear data distributions and reduce the impact of outliers. These findings support the importance of optimizing resolution and selecting appropriate classification algorithms. For system development, a combination of color segmentation, expansion of chili variety datasets, and integration of advanced preprocessing techniques can be strategies to improve model performance in the future.

References

[1] Permana Putra, I. (2020). *Scleroderma* spp. in Indonesia: Poisoning case and potential utilization. *JUSTEK: Jurnal Sains dan Teknologi, 3*(2), 37. https://doi.org/10.31764/justek.v3i2.3517

[2] Wardhani, I. P., & Widayati, S. (2019). Segmentasi warna citra HSV dan deteksi objek kupu-kupu dengan metode klasifikasi K-Means. Dalam *Prosiding Seminar SeNTIK* (Vol. 3, No. 1, hlm. 125–131).

[3] Novianto, D., & Sugihartono, T. (2020). Sistem deteksi kualitas buah jambu air berdasarkan warna kulit menggunakan algoritma principal component analysis (PCA) dan K-Nearest Neighbor (K-NN). *Jurnal Ilmiah Informatika Global, 11*(2), 42–47. https://doi.org/10.36982/jiig.v11i2.1223

[4] Widians, J. A., Pakpahan, H. S., Budiman, E., Haviluddin, H., & Soleha, M. (2019). Klasifikasi jenis bawang menggunakan metode K-Nearest Neighbor berdasarkan ekstraksi fitur bentuk dan tekstur. *Jurnal Rekayasa Teknologi Informasi (JURTI), 3*(2), 139. https://doi.org/10.30872/jurti.v3i2.3213

[5] Hasan, M. A., & Liliana, D. Y. (2020). Pengenalan motif songket Palembang menggunakan deteksi tepi Canny, PCA dan KNN. *Multinetics, 6*(1), 1–7. https://doi.org/10.32722/multinetics.v6i1.27000

[6] Putra, H. F., Tahiyat, R. M., Ihsan, R. M., Rahmaddeni, R., & Efrizoni, L. (2024). Penerapan algoritma K-Nearest Neighbor menggunakan wrapper sebagai preprocessing untuk penentuan keterangan berat badan manusia. *MALCOM Indonesian Journal of Machine Learning and Computer Science, 4*(1), 273–281. https://doi.org/10.57152/malcom.v4i1.1085

[7] Nurnaningsih, D., Alamsyah, D., Herdiansah, A., Aristo, A., & Sinlae, J. (2021). Identifikasi citra tanaman obat jenis rimpang dengan Euclidean distance berdasarkan ciri bentuk dan tekstur. *BITS, 3*(3), 171–178. https://doi.org/10.47065/bits.v3i3.1019

[8] Primandari, P. N., & Hardiansyah, B. (2018). Ekstraksi fitur menggunakan principal component analysis (PCA). *Hasil Riset dan Pengabdian Masyarakat sebagai Inovasi Menuju Persaingan Global, 1*(1), 66–74.

[9] A'yuniyah, Q. A., & Reza, M. (2023). Penerapan algoritma K-Nearest Neighbor untuk klasifikasi jurusan siswa di SMA Negeri 15 Pekanbaru. *Indonesian Journal of Informatics Research and Software Engineering, 3*(1), 39–45. https://doi.org/10.57152/ijirse.v3i1.484

[10] Pamungkas, A. (2020). Klasifikasi jenis sayuran menggunakan algoritma PCA dan KNN. *Pemrograman Matlab*. Diakses 11 Januari 2024, dari https://pemrogramanmatlab.com/2019/01/01/klasifikasi-jenis-sayuran-menggunakan-algoritma-pca-dan-knn/

[11] Nugraha, R. A., Hidayat, E. W., & Shofa, R. N. (2023). Klasifikasi jenis buah jambu biji menggunakan algoritma principal component analysis dan K-nearest neighbor. *Generation Journal, 7*(1), 1–7. https://doi.org/10.29407/gj.v7i1.17900

[12] Hasym, I. E., & Susilawati, I. (2021). Klasifikasi jenis ikan cupang menggunakan algoritma principal component analysis (PCA) dan K-nearest neighbors (KNN). *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, 1*(1), 168–179. https://doi.org/10.24002/konstelasi.v1i1.4242

[13] Nurfitri, K., Pradana, A. D., & Widaningrum, I. (2021). Penerapan algoritma principal component analysis (PCA) dan K-nearest neighbors (KNN) pada klasifikasi. *Jurnal Rekayasa Teknologi dan Komputasi, 1*(1), 1–19.

[14] Sambrani, Y., & Bhairannawar, S. (2023, Juni). Chili disease detection and classification using various machine learning techniques. Dalam *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)* (hlm. 1–6). IEEE. https://doi.org/10.1109/icaisc58445.2023.10199988

[15] Widyatama, R. A., & Hansun, S. (2019). Expert system for chili plants disease detection using certainty factor method. *International Journal of Innovative Technology and Exploring Engineering, 9*(1), 1145–1151. https://doi.org/10.35940/ijitee.a4440.119119

[16] Syahputra, H., Nainggolan, J., Sirait, J., Ikromi, M., & Lubis, P. (2024). Red chili classification using HSV feature extraction and naive Bayes classifier. *Teknosains Jurnal Sains Teknologi dan Informatika, 11*(1), 33–40. https://doi.org/10.37373/tekno.v11i1.593

[17] Prelle, A., Spadaro, D., Denca, A., Garibaldi, A., & Gullino, M. (2013). Comparison of clean-up methods for ochratoxin A on wine, beer, roasted coffee and chili commercialized in Italy. *Toxins, 5*(10), 1827–1844. https://doi.org/10.3390/toxins5101827

[18] Cserháti, T., Forgács, E., Morais, H., & Mota, T. (2000). Classification of chili powders by thin-layer chromatography and principal component analysis. *Journal of Biochemical and Biophysical Methods, 45*(2), 221–229. https://doi.org/10.1016/s0165-022x(00)00119-6

[19] Syukur, M., Zuhry, E., Armaini, A., Yunandra, Y., & Riska, F. (2023). Principal component and cluster analyses of 20 genotypes of chilli (*Capsicum* sp.) planted on peatlands. *IOP Conference Series: Earth and Environmental Science, 1228*(1), 012003. https://doi.org/10.1088/1755-1315/1228/1/012003

[20] Rohaeti, E., Muzayanah, K., Septaningsih, D., & Rafi, M. (2019). Fast analytical method for authentication of chili powder from synthetic dyes using UV-Vis spectroscopy in combination with chemometrics. *Indonesian Journal of Chemistry, 19*(3), 668–676. https://doi.org/10.22146/ijc.36297.