Classifying Chilli Plants Using Digital Images And Multiple Linear Regression

Esa Mahendra
Department of Information System, Universitas Muhammadiyah Sumatera Utara, Indonesia

ABSTRACT

The present study focuses on the application of classification methods using digital images and multiple linear regression to identify types of chili plants based on texture and shape features extracted from leaf images. In the process, digital images of chili plants undergo a pre-processing stage to enhance image quality, followed by feature extraction using methods such as the Gray-Level Co-occurrence Matrix (GLCM). The present study utilised 100 datasets of chili plant images obtained from the BRIN website, which were then divided into training data and test data to train a multiple linear regression model. However, the findings of the study indicated that the multiple linear regression model was not adept at encapsulating the intricacies of the data, as evidenced by the negative R-squared value and substantial prediction errors. Consequently, it is recommended that dimensionality reduction and cross-validation techniques be applied to enhance model performance and increase accuracy in classifying chili plant types in future.

Keyword : Linear regression, Digital imagery, Chilli

Corresponding Author:
Esa Mahendra,
Department of Information System,
Universitas Muhammadiyah Sumatera Utara,
Jalan Kapten Muktar Basri No 3 Medan 20238, Indonesia.
Email: esamahendra@gmail.com

1. INTRODUCTION

Chili plants (Capsicum spp.) are a high-value horticultural commodity with substantial economic importance in many countries, including Indonesia. Widely used in culinary applications and known for various health benefits, chili has consistently high market demand, making productivity and quality crucial aspects for farmers and stakeholders. In response to these needs, digital image processing and data analysis technologies have emerged as vital tools in precision agriculture. The application of digital imagery enables more accurate and efficient observations of plant characteristics such as leaf shape, fruit color, and surface texture. These advancements support farmers in rapidly and accurately identifying chili plant varieties, ultimately improving crop management and yield outcomes (Putra, 2022).

However, distinguishing between different types of chili plants remains a complex challenge due to variations in shape, size, and color that are often difficult to discern through basic visual inspection. Misidentification can negatively affect agricultural practices, such as inappropriate cultivation techniques or incorrect pesticide usage, ultimately lowering both the quality and quantity of harvests. Digital image processing offers a solution by enabling detailed and automated observation of chili plant features. This method reduces reliance on human observation and increases classification precision.

To enhance this technological approach, multiple linear regression (MLR) can be employed to analyze the relationship between multiple digital image features—such as color intensity, shape, and texture—and the chili plant type. This model not only facilitates accurate classification but also helps determine which features are most influential. As demonstrated by previous research (Saputra & Ardani, 2020), MLR provides interpretability and accuracy, allowing it to serve as both a prediction tool and a guide for farmers. Furthermore, similar approaches have shown success in agricultural contexts, such as the estimation of soil moisture content using vegetation indices from satellite imagery (Mutmainna, Achmad, & Suhardi, 2017).

Based on this background, this study aims to develop a classification model for chili plant types using digital images and multiple linear regression analysis, providing both practical and scientific contributions to agricultural technology.

2. RESEARCH METHOD

A. Research Data

The data used in this research pertains to the classification of chili plant types. The classification data was obtained from the website brin.go.id. In this system, registered farmers input data related to the types of chili plants they cultivate, including their physical characteristics and growth conditions. This data is then used to identify and classify the types of chili plants being grown. A total of 100 data entries were successfully collected from the brin.go.id website for this study. The dataset includes information such as images of the cultivated chili plants. The data is presented as follows:

Table 1. Research Data

No.	Plant Picture
1	
2	
3	
4	
...	

97			
98			
99			
100			

B. General Architecture

This study employs a general architecture consisting of several stages to classify chili plant types using digital image processing and multiple linear regression. The main stages of the architecture used are as follows:

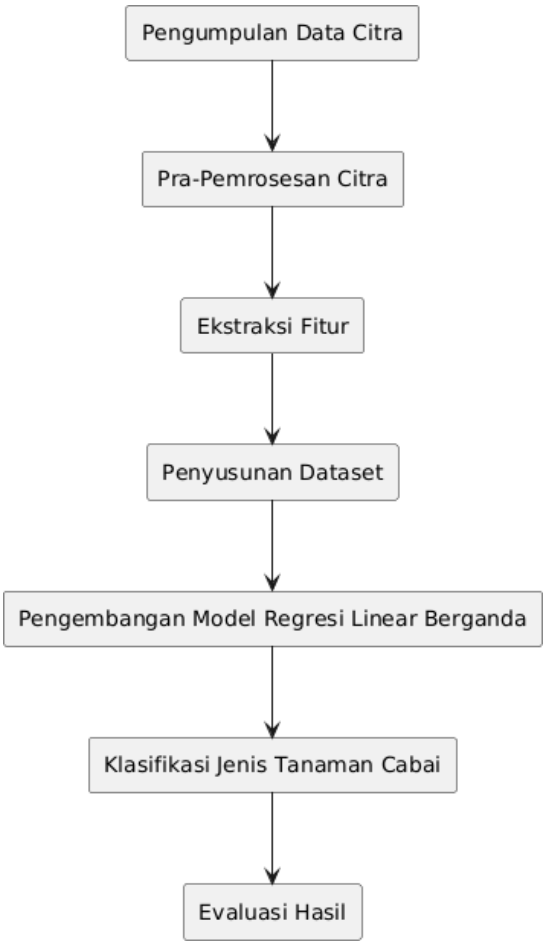

Figure 1. Research Architecture

C. Data Preparation

1. Data Download

The classification data of chili plant types is downloaded from the website brin.go.id. The data consists of images in JPG format depicting the shape of chili leaves and other characteristics.

2. Format Conversion

The downloaded JPG images are converted into CSV table format to store metadata and extracted feature results. This step allows the data to be more easily processed and analyzed using data analysis tools such as Pandas in Python.

3. Data Cleaning

Identifying and removing duplicate images to ensure each entry is unique. Missing or corrupted images are either filled in or removed from the dataset. For instance, incomplete or blurry images may be discarded if they are not too numerous.

4. Image Preprocessing

Standardizing the size of all images to uniform dimensions to ensure consistency in analysis.

Table 2. Image Processing

Before	After
	

5. Texture Analysis

The identification of texture patterns on leaf surfaces is facilitated by the implementation of methodologies such as the GLCM (Gray-Level Co-occurrence Matrix).

Table 3. Gray-Level Co-occurrence Matrix

Before	After
	

	Image	Contrast	Dissimilarity	\
0	5f3ec52b-bd7a-41b7-aec7-ce388b5cd9d5.jpeg	2295.04154	31.886687	

	Homogeneity	Energy	Correlation	ASM
0	0.052813	0.009391	0.708124	0.000088

Figure 2. Extraction Results

As illustrated in Figure 4.1, the table results generated from the Gray-Level Co-occurrence Matrix (GLCM) feature extraction for chili leaf images are presented. The following table provides a comprehensive list of the image names and the GLCM feature values that have been extracted from the images.

6. Dataset Arrangement

The feature data that has been extracted is then arranged into a dataset, which is then ready for further analysis. Each row in the dataset corresponds to an image of a chili leaf, incorporating the extracted features and the designated chili type label.

3. RESULTS AND DISCUSSION

The present study proposes a system that utilises the Multiple Linear Regression method for the identification of different types of chili plants. Concurrently, a digital image processing method was employed to detect and process images of chili plants, utilising the OpenCV library and techniques such as image segmentation and GLCM (Gray-Level Co-occurrence Matrix) feature extraction. The database contains a total of 100 images of various types of chilli, obtained from the BRIN database. The database is divided into two parts: training data and test data. Each variety of chili is characterised by a multitude of images, representing diverse variations in shape, colour, and texture.

Figure 3. Dataset of Chilli Classification

3.1 Imagery Detection

In this study, a digital image processing method was utilised for the purpose of detecting the image of chili plants. Digital image processing is the process of separating the main object (foreground), namely the chili plants, from the background. The proposed methodology involves the separation of the RGB values into intensity values and colour characteristics. The original colour in the image of chili plants often still contains lighting effects that can change visual characteristics such as colour, shape, and texture. Therefore, it is necessary to convert it into a feature representation that is more stable against changes in lighting. In order to mitigate the impact of lighting conditions, the utilisation of a model such as the GLCM (Gray-Level Co-occurrence Matrix) is recommended. This model facilitates the extraction of texture features, including contrast, homogeneity, and correlation, which exhibit enhanced robustness to variations in lighting.

Figure 4. Imagery Component YCbCr

The image presented here illustrates the outcome of the conversion process of an image of a chili plant from RGB colour space to YCbCr colour space. It also demonstrates the separation of the Y, Cb, and Cr components that is an inevitable consequence of this process. The RGB image is the original representation of the leaves of the chili plant in the Red, Green, Blue colour format, which is a common format for storing and displaying digital images. This RGB format employs a combination of three primary colours to display full colour on each pixel, which closely matches the way in which the human eye responds to light. Subsequent to the conversion of the RGB image to the YCbCr colour space, the

image is divided into three components: Y (luminance), Cb (chrominance-blue), and Cr (chrominance-red). The Y (luminance) component is responsible for storing information regarding the light intensity or brightness of each pixel, independent of its colour. The image is displayed in grayscale, representing a quantitative representation of lightness or darkness in each area of the image. This component is of particular significance, given that the human eye exhibits a greater sensitivity to changes in brightness than to changes in colour. In the following section, the chrominance-blue (Cb) and chrominance-red (Cr) components are analysed as representations of the difference between blue and red in an RGB image, with a focus on their respective luminance values. The Cb component is indicative of the 'blue' hue of a pixel in relation to its luminance, whilst the Cr component is indicative of the 'red' hue of the pixel. This information is useful in a variety of image processing applications, especially since colour information can be compressed more than luminance information without significantly reducing perceived visual quality.

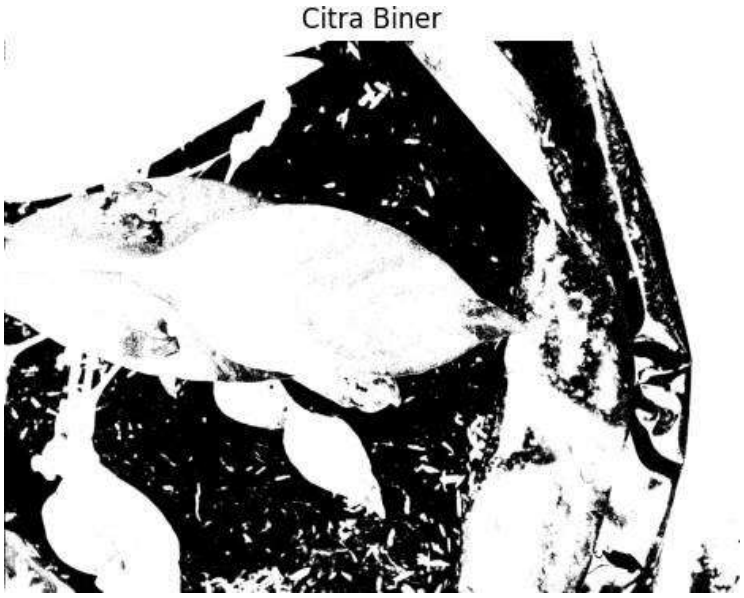

Figure 5. Binner Image

The image presented here illustrates the outcome of the conversion process of chili plant images into binary images. This binary image is a representation of the original image that has been converted into a form with two intensity values, namely black and white. The process is initiated by converting the image to grayscale, whereby the image is reduced to a series of shades of grey, devoid of any colour information. Subsequently, a thresholding technique is employed to categorise pixels into two distinct groups: those exceeding a specified threshold value are rendered as white, while those falling below the threshold are displayed as black. In this binary image, the regions of the chili plant that exhibit lighter pigmentation, such as the leaves, are represented by white areas, while darker pigmentation, such as the background or shadows, is represented by black areas. The utilisation of this binary image facilitates subsequent analysis processes, including edge detection, object segmentation, and the calculation of specific image areas. It is evident that the utilisation of binary images, characterised by a mere two levels of intensity, engenders a pronounced representation of contrasting properties. This property of binary images has the potential to reduce the complexity of visual data processing, thereby facilitating the identification of salient features within the image, such as the shape and contour of chili leaves.

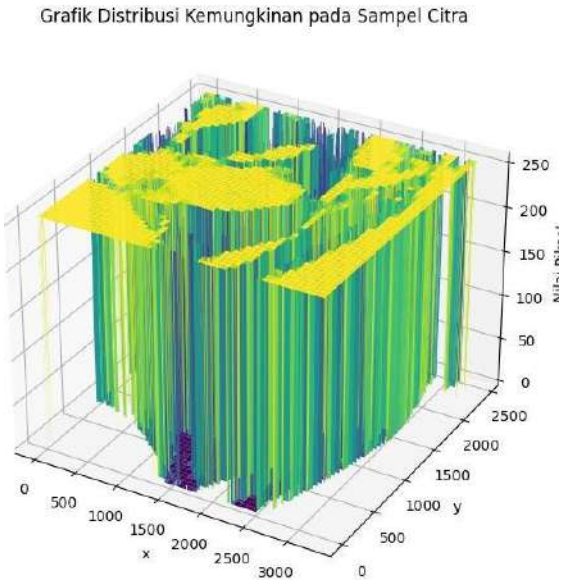

Figure 6. Probability Distribution Graph On Image Samples

The image presented here illustrates a three-dimensional distribution graph of pixel values in a sample image. The graph illustrates the relationship between the position of a pixel in the image (represented by the x and y axes) and the intensity value of that pixel (represented by the z axis) in the image. Each bar in the graph is representative of the intensity value of a pixel at a specific point in the image. The colour and height of the bar are indicative of the difference in intensity or value of the pixel. Visually, this graph provides a clear representation of the distribution of brightness or darkness across the image. Regions within the image exhibiting high intensity values are indicative of the presence of a bright or luminous object. These regions are represented by tall, brightly coloured bars. Conversely, darker areas in the image are represented by lower, darker bars. The analysis of the graph enables the comprehension of the distribution of light intensity across the image, facilitating the identification of significant patterns or features, including edges, textures, and shapes, that are present within the image.

Figure 7. Erosion Image and Dilation Image

The image presented here illustrates the outcomes of two prevalent morphological operations in image processing, namely erosion and dilation, applied to an image of a chili plant. The initial image illustrates the outcome of the erosion process. Erosion is a technique that removes pixels at the edges of objects in an image, thereby reducing the size of the bright areas, with the potential result that fine details or small objects may be lost. In this image, the edges of the leaves and other minor elements have been reduced or even eliminated, thereby rendering the primary objects appear more slender and concentrated. Conversely, the second image illustrates the outcome of the dilation process, which results

in the expansion of the bright areas of the image. In the process of dilation, the objects in the image appear to increase in size and are distributed over a greater area, thereby filling in any gaps or small holes that may have appeared subsequent to the process of erosion or as a result of noise in the original image. In this image, the white regions, which represent leaf shapes, are more substantial and voluminous, with thicker edges and more pronounced forms when compared to the result of the erosion process.

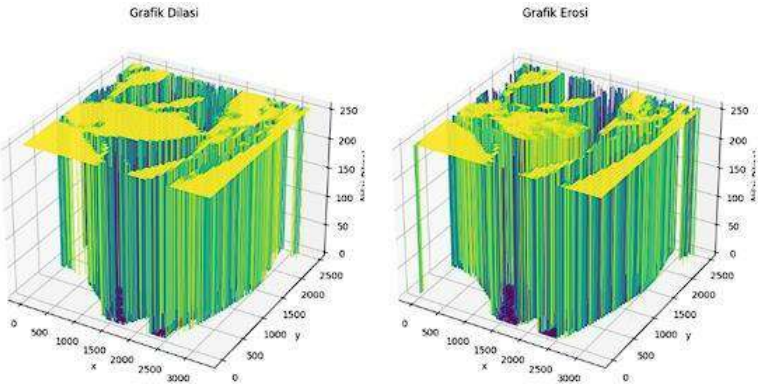

Figure 8. Dilation Graph and Erosion Graph

The image displays two three-dimensional graphs, illustrating the outcomes of the morphological process on the image, specifically dilation and erosion. The graph on the left illustrates the outcomes of the dilation process, while the graph on the right demonstrates the results of the erosion process. As demonstrated by both graphs, the distribution of pixel values is illustrated following the application of the morphological operation to the binary image. The dilation graph provides a visual representation of the objects in the image, demonstrating an increase in size and a decrease in density. This is due to the fact that the operation of dilation introduces pixels to the object's edges, thereby expanding the bright areas and filling in minor gaps in the image. This effect of image dilation serves to accentuate the primary structures within the image, such as the contour of a leaf, by increasing both the surface area and the pixel values in various regions. Conversely, the erosion graph provides a divergent set of results. Erosion is a process that removes pixels at the object's edges, thereby causing the bright areas in the image to become smaller. This phenomenon can be observed by examining the erosion graph, where the object structure appears diminished in size and the regions exhibiting high pixel values are reduced in number when compared to the dilation graph. This phenomenon of image simplification, characterised by the elimination of minor details and the reduction of primary objects, is known as the erosion effect.

3.2 Pre-processing Technique

Pre-processing constitutes a pivotal preliminary stage within the broader framework of image processing, wherein raw data is transformed into a format that is more amenable to subsequent analysis. The purpose of this step is to enhance image quality and reduce complexity by means of the removal of noise, the alignment of sizes, and the normalisation of pixel intensities. The implementation of pre-processing techniques facilitates the preparation of image data, thereby rendering significant features more readily extractable and analyzable in subsequent stages. This, in turn, ensures more accurate and efficient results in a range of image processing applications.

1. RGB to grayscale conversion

The RGB to grayscale conversion is a common procedure in digital image processing systems, as it has the capacity to streamline the analysis process by reducing colour information to a single brightness level. This is of particular significance in applications such as the classification of chili plants, where shape and texture information are more pertinent than colour information. This process also reduces memory requirements and speeds up calculations, making it easier and faster for subsequent analysis steps. Figure 4.6 presents several illustrations of the outcomes derived from the conversion of RGB images to grayscale within the domain of chili plant image processing.

Figure 9. Grayscale Image Results

The image presented here illustrates two visualisation outcomes of chili plant images. The first outcome is the original image in RGB format, and the second is the image that has been converted to grayscale. In the original RGB image, the leaves of the chili plant are displayed with all the colour information recorded by the camera, which includes the red, green, and blue colour spectrum. This colour information is imperative for natural visualisation, yet it frequently lacks pertinence in the context of shape and texture analysis within digital image processing. Conversely, the grayscale image displayed on the right has been obtained through conversion of an RGB image. In a grayscale image, each pixel possesses a single intensity value that denotes brightness, ranging from black (the lowest value) to white (the highest value). The elimination of colour information from the image results in the creation of a greyscale representation, which simplifies the visual data and facilitates the analysis of significant features such as texture, edges, and the shape of the chili leaves. This conversion process is particularly advantageous in the context of image processing applications, such as plant classification, where colour is not invariably the primary factor and the emphasis is instead placed on the structural and patterned elements present in the image.

	Contrast	Correlation	Energy	Homogeneity
0	0.000285	184.041622	0.971970	3284.990604
1	0.000199	181.724724	0.978169	4161.784851
2	0.000264	122.080939	0.979250	2945.219606
3	0.000248	173.212799	0.977758	3894.892279
4	0.000264	122.080939	0.979250	2945.219606

Figure 10. Extraction Results

The GLCM (Gray-Level Co-occurrence Matrix) feature extraction results from the displayed image show several important texture characteristics. The GLCM is utilised for the purpose of quantifying the frequency with which a specific pair of pixels, characterised by a given intensity value, manifests within a particular spatial configuration within the image. The resultant extraction process has revealed several salient features, namely Contrast, Correlation, Energy, and Homogeneity. Contrast is defined as the intensity of contrast or variation between a pixel and its immediate neighbours within the image. In this particular result, the contrast value is found to be very low (approximately 0.0002), indicating that the image exhibits low contrast or minimal variation in intensity between adjacent pixels. The magnitude of the correlation coefficient, which is a measure of the degree to which a pixel is correlated with its neighbours, is found to be exceedingly high (greater than 100), thereby signifying a robust linear correlation between the pixel intensities across the image. Energy, defined as a measure of uniformity in an image, exhibits a value close to 1 (approximately 0.97 to 0.98), thereby indicating that the image is highly homogeneous with a uniform intensity distribution. Homogeneity, a measure of the proximity of the distribution of elements in GLCM to the diagonal, also exhibits a high value (thousands), signifying that the distribution of pixel intensities in the image is highly uniform, indicative of a high level of consistency.

2. Data Labeling

The data labeling process constitutes a pivotal step in this study, with the objective of categorizing each sample of chili plant images according to its specific type. The application of labels to each image facilitates the organisation of data according to the type of chili, whether this is red chili,

green chili, or cayenne pepper. This label will serve as the foundation for training a multiple linear regression model, which will subsequently be employed to classify the type of chili plant based on the texture features that have been extracted from the digital image. This labelling process is instrumental in ensuring that the model is endowed with well-structured data, thereby facilitating more accurate predictions in identifying the type of chili from the image.

```
Dataset awal:
   Contrast  Correlation    Energy  Homogeneity  Label
0  0.000285   184.041622  0.971970  3284.990604      0
1  0.000199   181.724724  0.978169  4161.784851      0
2  0.000264   122.080939  0.979250  2945.219606      0
3  0.000248   173.212799  0.977758  3894.892279      0
4  0.000264   122.080939  0.979250  2945.219606      0

Jumlah baris dalam dataset: 108

Dataset setelah penambahan label:
   Contrast  Correlation    Energy  Homogeneity  Label
0  0.000285   184.041622  0.971970  3284.990604      0
1  0.000199   181.724724  0.978169  4161.784851      0
2  0.000264   122.080939  0.979250  2945.219606      0
3  0.000248   173.212799  0.977758  3894.892279      0
4  0.000264   122.080939  0.979250  2945.219606      0
```

Figure 11. Labelling Process

The initial dataset comprises four primary columns: The texture features that have been extracted from the images are represented by contrast, correlation, energy, and homogeneity. The dataset under consideration consists of 108 rows, with each row representing a single image sample. Prior to the application of labels, the dataset comprises solely the values of these features, devoid of any information pertaining to the category or type of the image under analysis. Following the addition of the Label column, each row in the dataset has now been enriched with information pertaining to the category of the image to which it corresponds. In this particular instance, all the labels in the rows displayed are set to 0, which may be indicative of the fact that these rows originate from the same category, for instance, "Red Chili". The significance of these labels lies in their utilisation within a classification model, which aims to predict the type of chili based on the extracted features. The incorporation of labels facilitates the utilisation of the dataset in machine learning algorithms, wherein the model is trained to discern patterns in the texture features that are indicative of a specific chilli variety. This labelling process constitutes a pivotal step in constructing a classification model, given that these labels are the projections upon which the model will subsequently base its predictions. In instances where multiple types of chili are present in the dataset, the Label column will display the respective variations, with distinct numerical values assigned to each type of chili. This process guarantees that the model is endowed with well-structured data for the purposes of training and testing, thereby ensuring its effective classification of chili types based on their digital images.

3. Datasets and Their Division

In this study, the subsequent step is to divide the dataset into two primary components: the training set and the test set. This division is implemented to ensure that the model to be trained can be evaluated accurately on data that has not been utilised during the training process. The training set is employed for the purpose of training the multiple linear regression model, while the test set is utilised for the evaluation of the model's performance. The division of the dataset is typically accomplished through a ratio of 80:20, wherein the majority of the data is allocated for model training, with the residual portion utilised for testing purposes. It is imperative to ensure that the model functions effectively on the training data and is also capable of generalising effectively on new data that has not been previously encountered. In order to achieve the random division of the dataset and ensure a similar data distribution between the training set and test set, it is necessary to use the train_test_split function from the scikit-learn library. This methodological approach assists in minimising bias and provides a

more accurate evaluation of the model's performance in classifying chili plant types based on digital images.

3.2 Application of Multiple Linear Regression Method

The objective of this study is to utilise the Multiple Linear Regression method for the prediction of the variety of chili plants based on texture features extracted from digital images. The model has been trained using a labelled dataset, with the objective of achieving accurate classification of chili types based on the linear relationship between existing features.

```
Mean Squared Error (MSE): 0.6583044725730125
R-squared (R²): -0.04465365483717365

Perbandingan Nilai Asli dan Prediksi:
   Actual  Predicted
0       1   0.813997
1       0   0.796617
2       0   0.914845
3       2   0.898951
4       1   0.492673
```

Figure 12. The Results of Method Implementation

The results obtained allow the author to observe multiple performance indicators of the multiple linear regression model that has been applied to the dataset. Mean Squared Error (MSE) is a metric used to measure the average of the squared errors between the actual and predicted values. A smaller MSE value indicates that the model exhibits reduced error in prediction. In this instance, the MSE obtained is approximately 0.6583, which indicates a substantial discrepancy between the model's prediction and the actual value. Furthermore, R-squared (R^2) is a metric that demonstrates the extent to which the model explains the variability of the data. The R^2 value ranges from 0 to 1, where a value of 1 indicates an excellent model for predicting the data. However, in this case, the R^2 value obtained is negative (-0.0447), which indicates that the model is unable to adequately explain the variability of the data. This negative value indicates that the model performs worse than the baseline model, which only predicts the average of the target values. The comparison between the actual and predicted values further demonstrates that the model frequently exhibits inaccuracies in its predictions, particularly in relation to categories. To illustrate this point, consider the first and fourth data points, which have actual values of 0 and 1, respectively. However, the model predicts values that are significantly different from the target. This further indicates that the multiple linear regression model may not be suitable for this classification task, or alternatively there is an issue with the data or features that have been utilised.

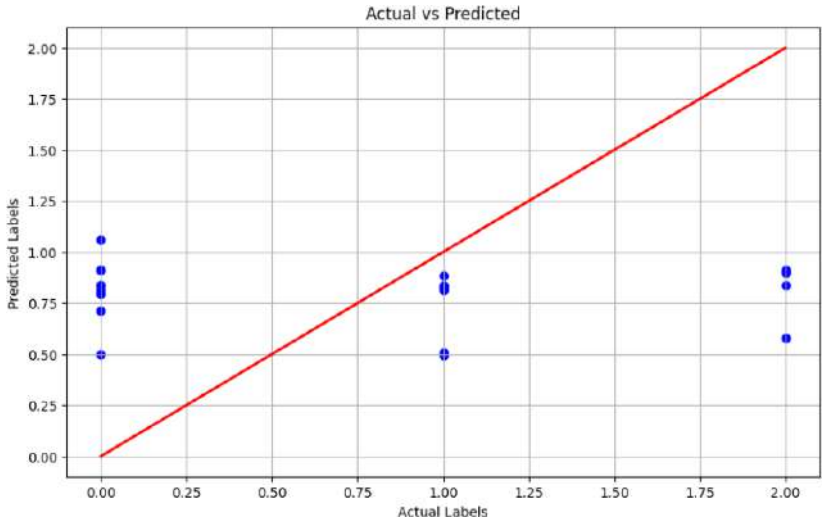

Figure 13. Scatter Plot Graph

The scatter plot graph presented here demonstrates the relationship between the actual and predicted values of the multiple linear regression model applied to the dataset pertaining to the chili plant. The red diagonal line in this graph represents the identity line, i.e. the line along which the predicted values should lie if the model were to make perfect predictions. However, the results indicate that the majority of the data points do not align with this identity line, suggesting a substantial discrepancy between the predicted and actual values. A considerable proportion of data points deviate substantially from the red line, indicating a substantial prediction error by the model. Furthermore, the clusters of values on the x-axis, which represent the actual labels 0, 1, and 2, are not proximate to the anticipated predicted values. For instance, a considerable proportion of predictions for label 0 are elevated above the anticipated values, while the predictions for other labels exhibit inconsistency with the expected values. The prediction errors observed suggest that the multiple linear regression model is not adequately capturing the patterns present in the data, consequently leading to inaccurate classification of the different types of chili plants.

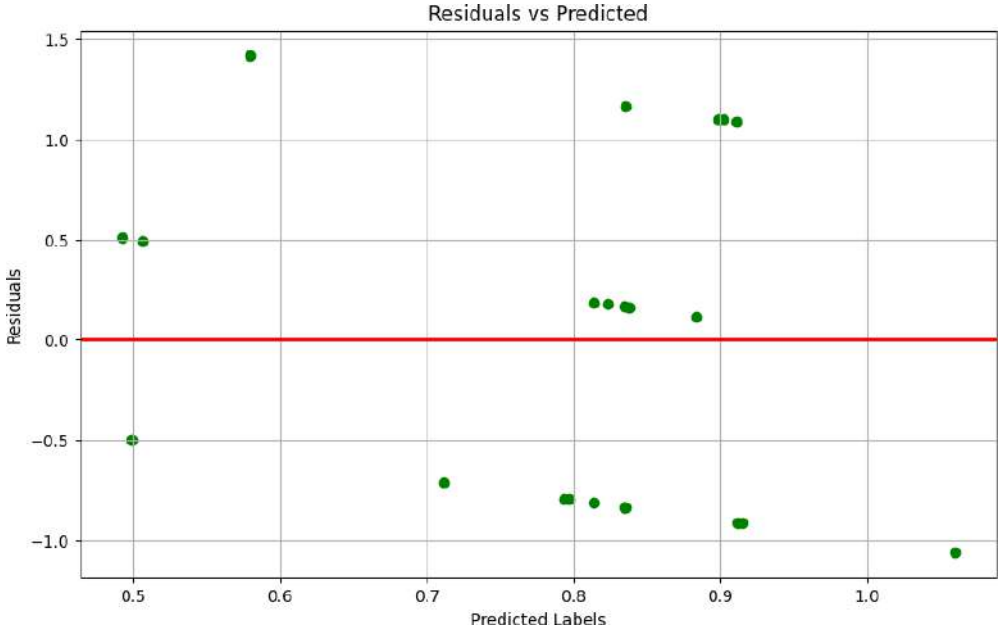

Figure 14. Residuals

As demonstrated in Figure 1, the graph depicting the residuals versus predicted values illustrates the relationship between the predicted values of a multiple linear regression model and the residuals, which are defined as the differences between the actual and predicted values. The horizontal

red line on this graph indicates the zero line, which is defined as the point at which the model predictions exactly match the actual values, resulting in zero residuals. The graph indicates that the majority of the data points do not align with the zero line, suggesting that the model generates substantial residuals or prediction errors. It is evident that certain observations exhibit elevated positive residuals, signifying that the model's predictions were considerably lower than the actual values. Conversely, there are also some points with large negative residuals, indicating that the model predicted values that were higher than the actual values. The distribution of residuals, which are not randomly distributed around the zero line, indicates that the model may not be capturing patterns in the data well. In the ideal scenario, if the model is functioning effectively, the residuals should be randomly distributed around the zero line, exhibiting no discernible pattern. This indicates that prediction errors occur consistently throughout the prediction range. However, in this case, the residuals tend to cluster at certain values, indicating a pattern of errors that could be due to the poor fit of the multiple linear regression model for these data.

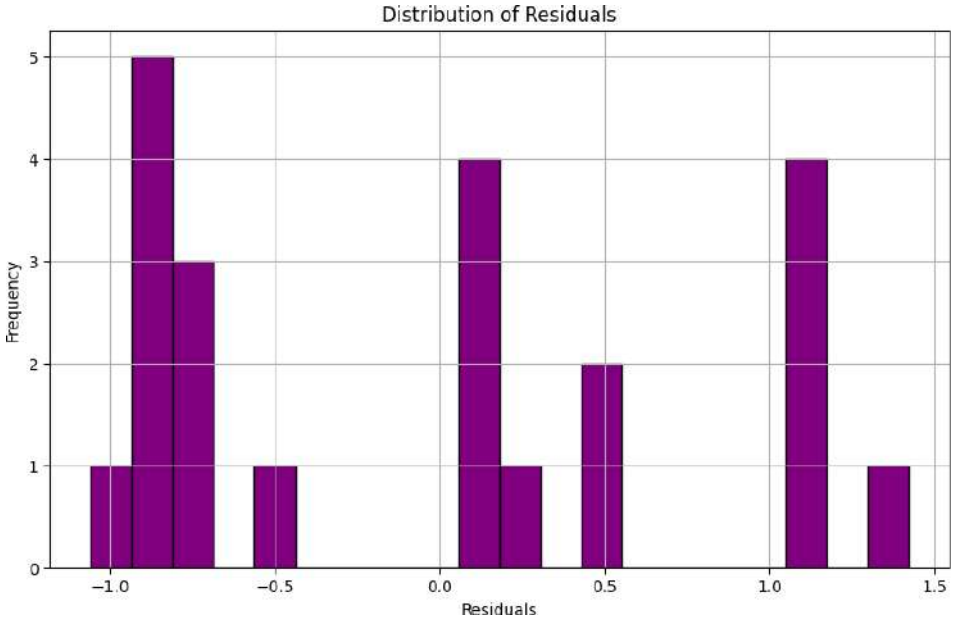

Figure 15. Histogram Graph

The histogram graph presented herein illustrates the distribution of residuals, defined as the differences between the actual and predicted values of the multiple linear regression model utilised in this study. These residuals are indicative of the prediction errors made by the model, and the distribution of the residuals can provide insight into the performance of the model. The residuals in this graph are distributed across multiple intervals, exhibiting distinct clusters. It is evident that the residuals are distributed between approximately -1.0 and 1.5. In the ideal scenario, where the model is performing well, we would expect a symmetrical bell-shaped distribution of residuals (normal distribution) centred around zero. This would indicate that the prediction errors are evenly distributed without bias. However, the graph reveals three distinct peaks in the distribution of residuals: one at -1, another at 0.5, and a third at 1.0. This distribution suggests that the model tends to make errors in several distinct clusters, rather than randomly. The presence of these large and scattered residuals suggests that the model is not adequately capturing the relationship between features and labels, resulting in less precise predictions.

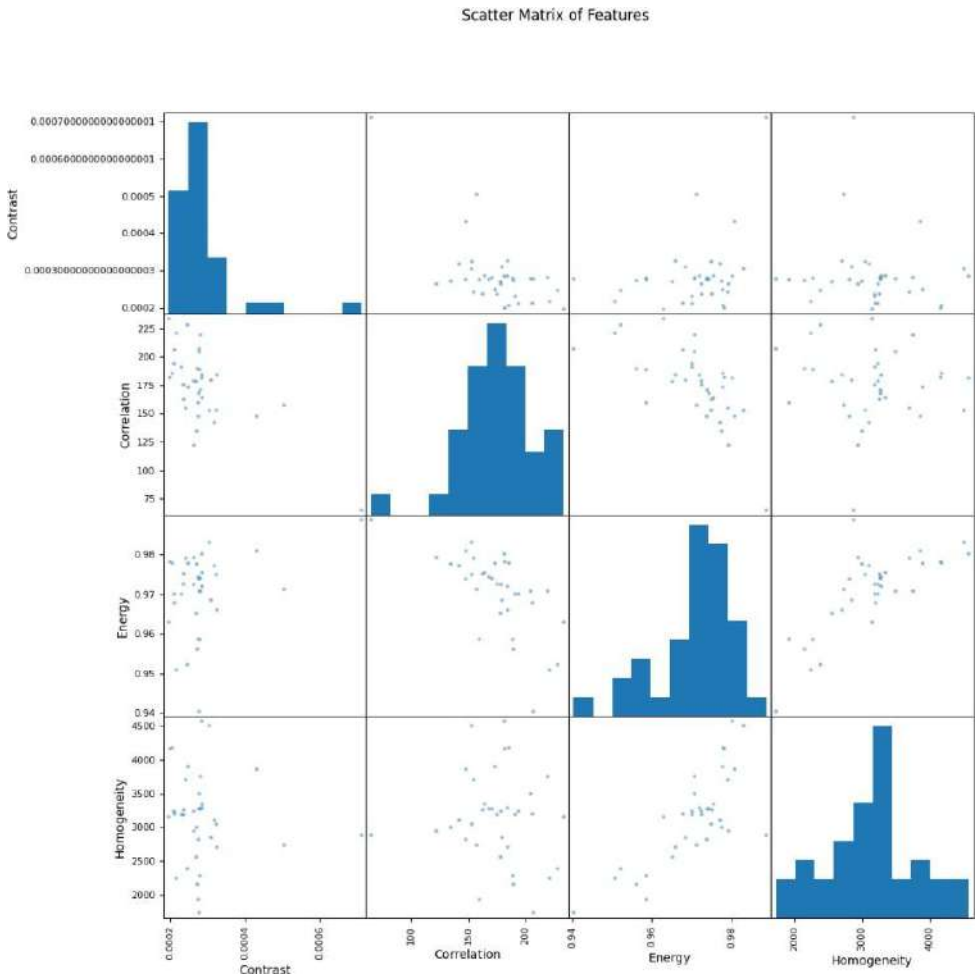

Figure 16. Scatter Matrix Graph

The scatter matrix graph presented herein provides a visualisation of the relationship between the four primary features extracted from the images in this study, namely contrast, correlation, energy, and homogeneity. Each cell in this matrix displays a scatter plot of a distinct feature pair, while the histogram on the diagonal illustrates the distribution of each feature individually. As demonstrated by the histogram for Contrast, the majority of values are concentrated within the very low range. By contrast, the histogram for Correlation exhibits a more uniform distribution, albeit with a peak that extends from 150 to 200. Furthermore, the distribution of energy and homogeneity demonstrates a tendency to be concentrated within specific value ranges. This finding suggests that the majority of samples exhibit uniform texture characteristics with regard to energy and homogeneity. The scatter plot between features in this matrix does not demonstrate any significant linear correlation. For instance, the relationship between contrast and correlation appears to be randomly distributed, exhibiting no clear linear pattern. A similar phenomenon is observed in the relationship between energy and the other features. This absence of a strong correlation suggests the possibility of complex or non-linear relationships between the features employed, which may prove challenging for multiple linear regression models to adequately capture existing patterns.

4. CONCLUSION

a. The findings of the multiple linear regression model demonstrate that this model does not provide optimal prediction performance in classifying chili plant types based on texture features extracted from digital images. This can be evidenced by the negative R-squared (R^2) value and the distribution of residuals, which are asymmetric and extend significantly beyond zero.
b. The scatter matrix analysis demonstrates that there is no significant linear correlation between the features employed (Contrast, Correlation, Energy, Homogeneity). This is one of the reasons why the

multiple linear regression model is unable to make accurate predictions, because this model relies on the existence of a linear relationship between these features.

c. The distribution of residuals, as observed in the histogram, reveals the presence of numerous significant error groups, both positive and negative in nature. This distribution indicates that the model frequently exhibits substantial prediction errors, suggesting a deficiency in its capacity to adequately capture the intricacies inherent in the data.

REFERENCES

[1] Andani, Ir, M T Dr Eng Muhammad Niswar, Andini Dani Achmad, and Eng Ir Wardi. n.d. "APLIKASI IOT DAN JARINGAN SENSOR NIRKABEL UNTUK MENURUNKAN RISIKO PENULARAN COVID-19."

[2] Batubara, Nur Arkhamia, and Rolly Maulana Awangga. 2020. Tutorial Object Detection Plate Number With Convolution Neural Network (CNN). Vol. 1. Kreatif.

[3] Hartatik, Hartatik, Ghefra Rizkan Gaffara, Habibi Azka Nasution, Ardiansyah Ardiansyah, I Nyoman Alit Arsana, Urnika Mudhifatul Jannah, and S T Iwan Adhicandra. 2023. PENGENALAN PEMROGRAMAN DASAR DUNIA KODING. PT. Sonpedia Publishing Indonesia.

[4] Idris, Mohamad, Riza Ibnu Adam, Yulrio Brianorman, Rinaldi Munir, and Dimitri Mahayana. 2022. "Kebenaran Dalam Perspektif Filsafat Ilmu Pengetahuan Dan Implementasi Dalam Data Science Dan Machine Leaning." Jurnal Filsafat Indonesia 5 (2): 173–81.

[5] Kamal, Silmi Afifah. n.d. "Implementasi Hidden Markov Model Dengan Algoritma Forward-Backward Untuk Prediksi Buku Peminjamandi Perpustakaan Studi Kasus: Perpustakaan SPS UIN Jakarta." Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

[6] Mutaqin, Ghani, and S Kom. 2023. Teknik Penghapusan Kabut Pada Citra Digital. Nas Media Pustaka.

[7] Mutmainna, Nurilmi Dwi, Mahmud Achmad, and Suhardi Suhardi. 2017. "Pendugaan Lengas Tanah Inceptisol Pada Tanaman Hortikultura Menggunakan Citra Landsat 8." Jurnal Agritechno, 135–51.

[8] Nugraha, Fikri Aldi, Nisa Hanum Harani, and Roni Habibi. 2020. Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning. Kreatif.

[9] Pramanda, Heru, Wahyu Diara, and David Sarana. 2024. "Analisis Pengaruh Pengalaman Dan Karakter Sumber Daya Manusia Terhadap Kualitaskerjaan Proyek Konstruksi: Studi Kasus: Kecamatan Kuta Alam Dan Syiah Kuala." Tameh 13 (1): 55–66.

[10] PUTRA, I KOMANG WIDARMA. 2022. "UJI POTENSI ATRAKTAN DAUN CENGKEH DAN DAUN KEMANGI TERHADAP HAMA LALAT BUAH (Bactrocera Spp) PADA CABAI (Capsicum Annuum L.)." Universitas Mahasaraswati Denpasar.

[11] Putra, I Nyoman Tri Anindia, Ketut Sepdyana Kartini, Yoga Kristian Suyitno, I Made Sugiarta, and Ni Kadek Era Puspita. 2023. "Penerapan Library Tensorflow, Cvzone, Dan Numpy Pada Sistem Deteksi Bahasa Isyarat Secara Real Time." Jurnal Krisnadana 2 (3): 412–23.

[12] Rifky, Sehan, Lalu Puji Indra Kharisma, H Achmad Ruslan Afendi, Segar Napitupulu, Mustika Ulina, Wulan Sri Lestari, I Made Dendi Maysanjaya, Kelvin Kelvin, Frans Mikael Sinaga, and Mutmainnah Muchtar. 2024. Artificial Intelligence: Teori Dan Penerapan AI Di Berbagai Bidang. PT. Sonpedia Publishing Indonesia.

[13] Saputra, Gede Wisnu, and I Gusti Agung Ketut Sri Ardani. 2020. "Pengaruh Digital Marketing, Word of Mouth, Dan Kualitas Pelayanan Terhadap Keputusan Pembelian." Udayana University.

[14] Saputra, Yogi Aditya, Syafrial Fachri Pane, and Rolly Maulana Awangga. 2020. Big Data: Implementasi Hadoop Mapreduce Pada Pemetaan Sekolah Menggunakan Python. Kreatif.

[15] Setiawan, Zunan, Muhammad Fajar, Arif Mudi Priyatno, Anggi Yhurinda Perdana Putri, Mediana Aryuni, Siti Yuliyanti, Harya Widiputra, Budanis Dwi Meilani, Rohmat Nur Ibrahim, and Rezania Agramanisti Azdy. 2023. Buku Ajar Data Mining. PT. Sonpedia Publishing Indonesia.