ASTEEC Conference Proceeding: Computer Science
3rd International Conference on Information Science and Technology Innovation (ICoSTEC)
July 27, 2024, Yogyakarta, Indonesia

Comparative Analysis of Linear Discriminant Algorithms and Support Vector Machine in Palm Fruit Image Disease Classification

Finis Hermanto Laia[1] , Hartono[2] , Zakarias Situmorang[3]

[1,2]*Magister of Computer Science, Potensi Utama University*
JL. KL. Yos Sudarso Km. 6.5 No. 3-A, Medan, Indonesia
[3]*Departement of Computer Science, Katolik Santo Thomas University*
JL. Setia Budi, kp. Tengah, Kec. Medan Tuntungan, Medan, Indonesia
[1]finishermanto@gmail.com
[2]hartonoibbi@gmail.com
[3]zakarias65@yahoo.com

Abstract — **Artificial Intelligence (AI) is a branch of computer vision used for object detection and image classification using algorithms. Approaches to comparing object characteristics in image processing can be divided into High Dimensional Feature approaches and Low Dimensional Feature approaches. Support Vector Machine (SVM) is an accurate High Dimensional Feature method, while Linear Discriminant Analysis (LDA) is a powerful Low Dimensional Feature Method. Some studies combine SVM with LDA to reduce complexity and improve performance. In the classification of palm fruit disease images, the comparison between SVM (High Dimensional Feature) and LDA (Low Dimensional Feature) can be done with variations in dataset size and the percentage of training and testing data in the research is 50:50, 60:40 and 70:30. Performance measurement is based on accuracy, precision, recall, and f-1 score. The algorithm for testing predictions for the validity of accuracy results is k=5 Cross Validation. The average test results in Linear Discriminant Analysis had the highest prediction, namely 86.00%, obtained in the 1st iteration, a percentage of variation of 50% of the image data. Meanwhile, the lowest average value was obtained in the 5th iteration, namely 66.00%, a percentage of variation in 30% of the image data. Then the average prediction value for system testing is 79.67%. Meanwhile, the support vector machine calculation test results have the highest prediction average, namely 96.00%, obtained in the 1st iteration, a percentage of variation of 30% of the image data. The lowest average accuracy value was obtained in the 1st iteration, namely 92.55% with a variation percentage of 40% of the image data. Then the prediction value for the test data system was obtained at 93.98% of the average results for each iteration. This study aims to compare the performance of LDA and SVM algorithms in classifying healthy or diseased oil palm fruit with variations in data set size and percentage of training and testing data. From the results of research testing carried out, SVM has an accuracy of 93.33% at the 1st percentage variation. Meanwhile, LDA is 86.66% with the same percentage variation. SVM showed to be more effective in classifying images of sick palm fruit or healthy fruit compared to LDA.**

Keywords — **Image Classification, Palm Fruit, Linear Discriminant Analysis, Support Vector Machine, Cross Validation**

I. INTRODUCTION

The field of artificial intelligence has become very popular and continues to be widely researched. One branch of AI is computer vision. Image Processing methods, specifically related to the comparison of characteristics between objects (visual comparison), can generally be divided into 2 (two) approaches, namely: the High Dimensional Feature approach[1] and the Low Dimensional Feature[2]. One of the High Dimensional Feature methods which is known to be very good and provides accurate results is Support Vector Machine (SVM) [3]. The Low Dimensional Feature method which is quite powerful is the Linear Discriminant Analysis (LDA) method [4]. Support Vector Machine (SVM) is a classification algorithm that has supervised learning characteristics which works by finding the optimal hyperplane (decision boundary) that separates the distance between classes, generalization and stable classification accuracy [5]. Linear Discriminant Analysis (LDA) is a dimensional reduction technique used for the pattern classification stage and machine learning applications. LDA is used to obtain image features and provides a larger distance between classes, while the distance between training data within a class is smaller [6].

Some researchers generally combine the application of SVM with LDA. Where LDA is used as a Feature Reduction method to reduce complexity in SVM [7]. However, several researchers such as [4] have used LDA as a fairly good image processing method with good computing time for image recognition. Research conducted by [8] also shows that the LDA method provides good results as a supervised learning method. [9] conducted research on which method is better between High Dimensional Features and Low Dimensional Features and obtained the results that which method is better depends on the size of the existing dataset and also the size of the training data and testing data used. This is also the same as research regarding the use of the Linear Discriminant Analysis method for face recognition by comparing the amount of training data and research by [10] regarding the influence of the size of the

training dataset on the performance of Support Vector Machines and Discision Trees. In research conducted by [11] with the research title "Building a Medium Scale Dataset for Nondestructive Disease Classification in Mango Fruits Using Machine Learning and Deep Learning Models". The problem in this research is that the quality of fruit is often assessed based on sensory attributes such as taste and aroma, shape, size, color. By using computer image processing combined with learning algorithms. The results of the study showed that SVM achieved a significant disease classification accuracy of mango fruit of 95%, while CNN was 91.52%. Research by [12] with the research title "Machine learning for cultivar classification of apricots (Prunus armeniaca L.) based on shape features". This research aims to classify apricot varieties using machine learning based on shape features. Identification of aprico varieties uses six machine learning methods, namely decision tree, K-nearest neighbor, naive Bayes, linear discriminant analysis, support vector machine, and back propagation neural network. Based on the research results, it shows that apricot varieties have significant differences in shape features with the support vector machine classification algorithm providing the best results with an accuracy rate of 90.7% in the test dataset.

It is interesting to observe the comparison between SVM as High Dimensional Features and LDA as Low Dimensional Features in its application in recognizing diseases in oil palm fruit, especially with variations in the amount of data and percentage of training data and testing data. Recognizing diseases in oil palm plants is very important considering that palm oil is a mainstay source of income, an export commodity and supports domestic industry. Therefore, researchers in this study are interested in comparing the performance of the LDA and SVM algorithms in disease classification in oil palm fruit plants, especially in relation to variations in dataset size and also the percentage of training data and testing data.

II. Research Methods

In this study, a comparison of the linear discriminant analysis (LDA) classification algorithm with support vector machine (SVM) was carried out on healthy and diseased palm fruit. This research will produce the highest accuracy results, when using the LDA and SVM algorithms. Fig. 1. shows the general flow diagram of research methodology.

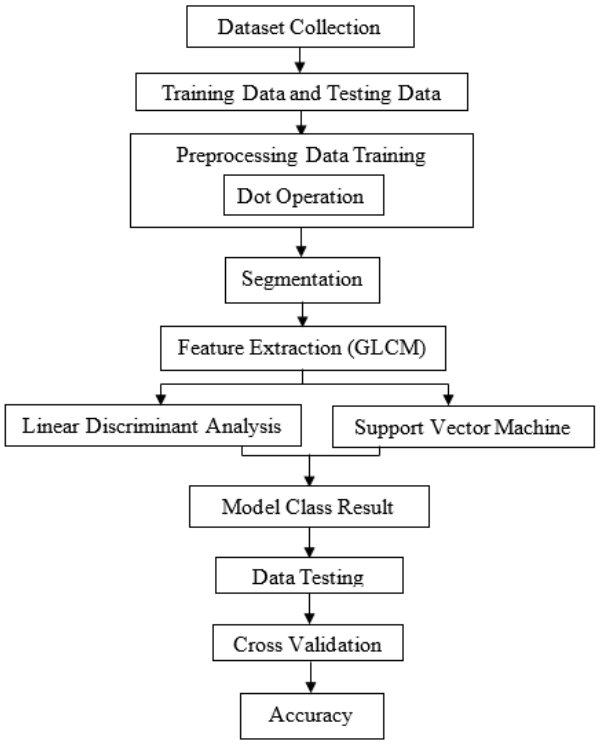

Fig. 1 Stages of research methodology

A. Data Collection

The data used in this research is secondary data sourced from the largest collection of open source computer vision datasets which can be found at the website address universe.robof Palm Fruit Image Dataset, the author collected with training data totaling 120 images and testing data totaling 120 of them, class sick fruit and healthy fruit. The following image shows an example of an image dataset of palm fruit:

(a) (b)

Fig. 2 Categories of fruit (a) sick fruit and (b) healthy fruit

The scenarios carried out include the first percentage of 50% training data : 50% testing data, second 60% training data : 40% testing data and third 70% training data : 30% testing data. Complete data can be seen in table 1. Variations in creating training and testing data are as follows:

TABLE 1
VARIATIONS IN PALM FRUIT IMAGE DATASET

No	Dataset Variations	Training Data	Data Testing	Sum
1.	50% : 50%	90 images	90 images	180 images
2.	60% : 40%	108 images	72 images	
3.	70% : 30%	126 images	54 images	

B. *Preprocessing Training Data*

The image preprocessing stage in processing is image quality improvement (Image Enhancement) which consists of dot operations and special operations and the resizing process is carried out to homogenize the size of the same data set. The image quality was improved using one dot operation using the increase contrast technique. The following is a grayscale image that has been quality improved into an Intensity Adjustment Image image shown in Fig. 3. Dataset preprocessing stage:

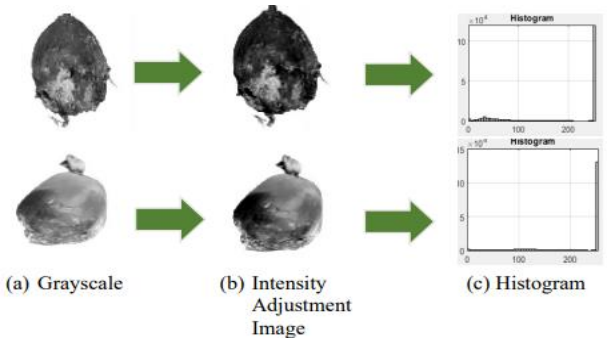

Fig. 3 (a) Grayscale (b) Adjustment Image (c) Histogram

In Fig. 3. improving image quality from grayscale to Intensity Adjustment Image (Histogram) aims to increase the difference between pixel intensity values in the image and help make it easier to distinguish objects in the image, such as reducing noise, increasing feature visibility, for the classification process. At this stage, the pixel intensity value in the image is changed to a higher intensity value so that the objects in the image are more visible and the segmentation process becomes easier.

C. *Segmentation*

At the segmentation stage, separating objects or features in the image, the pixel intensity value in the image is converted into a threshold measuring threshold value (128) with the aim of dividing the categorized image into objects and background before being converted from a thresholding image to a binary image. Binary images are produced from a thresholding process to simplify the segmentation process. Binary images have two contrasting colors, so that objects in the image are easier to separate from the background. The following are the results of images of sick palm fruit (a) and healthy fruit (b) from converting thresholding image values to binary, which can be seen in the following image:

Fig. 4 Image thresholding conversion to binary

D. *Gray Level Co-occurrence Matrix (GLCM) Feature Extraction*

The process of feature extraction or taking characteristics from an image object, the results are used as input values for the next stage of the classification process. The results of feature extraction used in this research are using texture feature extraction (GLCM) based on contrast, correlation and energy values, to differentiate objects with certain textures you can use the contrast values used to differentiate objects with smooth and rough textures and the correlation value differentiates textures. which has regular and irregular directions in the pattern, while the energy value differentiates objects with repeated and non-repeated textures from objects in an area.

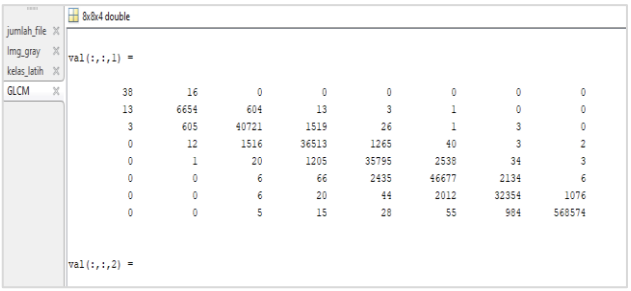

Fig. 5 Texture Feature Matrix Values

The following are the results of GLCM texture feature extraction. The percentage of training data and testing data for 120 images can be seen in the table below, an example of the results of feature values in GLCM:

TABLE 2
GLCM TEXTURE FEATURE EXTRACTION

Training Data				
No.	File Name	Contrast	Correlation	Energy
1.	'healthy fruit1.jpg'	0.18132	0.95948	0.68838
2.	'sick fruit2.jpg'	0.20412	0.97701	0.65626
Data Testing				
1.	'healthy fruit1.jpg'	0.12192	0.98158	0.64240
2.	'sick fruit2.jpg'	0.097177	0.97901	0.71385

In table 2, the results of texture feature extraction using the Gray Level Co-occurrence Matrix (GLCM) on two sets of data, training data and testing data. glcm is used to analyze texture in images by examining the spatial relationship between pixel intensities. In the testing data, a similar pattern can be seen. The 'healthy fruit1.jpg' image has a lower contrast value and a higher correlation value compared to the 'sick fruit2.jpg' image. The results obtained were used to train a classification model to differentiate between healthy and diseased fruit images based on GLCM texture feature extraction.

III. Results and Discussions

A series of trials and performance evaluations of research conducted using the Matlab R2021a programming language to create classifications for diseases of healthy and diseased palm fruit. The application of the model using LDA and SVM in this research carried out three experiments on predetermined scenarios with percentage variations in the training data and test data datasets.

A. Linear Discriminant Analysis (LDA)

In the process of classifying palm fruit using linear discriminant analysis, the following is a visualization of the results of the fitcdiscr modeling of the training data and test data in the following image:

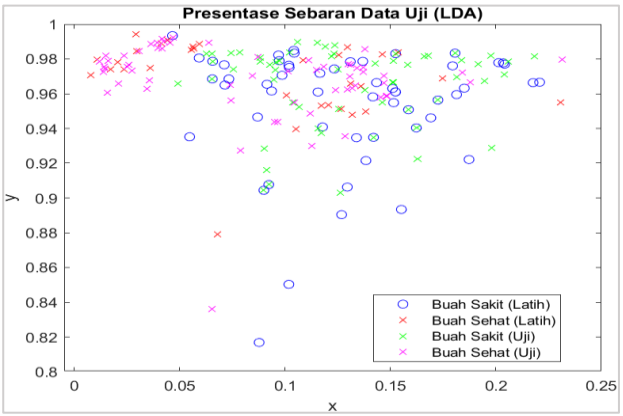

Fig. 6 Visualization of testing data results (LDA)

In the test results, the diseased fruit object (test) is described with a green x-mark plot and the healthy fruit (test) is described with a magenta x-mark plot where the points represent the sick and healthy fruit classes.

B. Support Vector Machine (SVM)

Visualization of the results of modeling the distribution of support vector machine kernel *linear test data* .

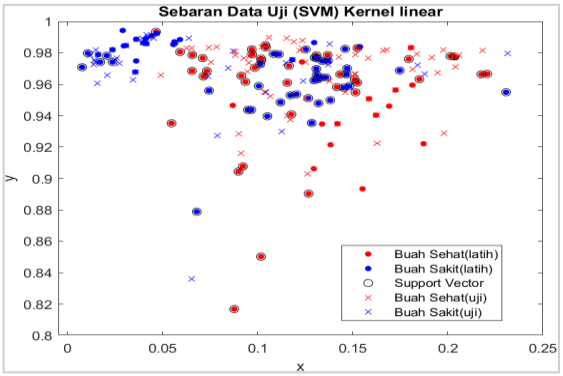

Fig. 7 Visualization of SVM test data results

In the visualization process, the test plot displays the test data and training data separated using the SVM model. The borders created by the SVM model divide two classes of sick fruit and healthy fruit. The test data classes are shown with red and blue crosses and the support vector machine points are marked with black dots.

To test the models for each of the two algorithms that have been built, the first step is to input the image to be tested from the testing data, then the image data is subjected to a process of improving image quality at the preprocessing stage for segmentation, then the image texture features are extracted by GLCM characteristics based on contrast values, correlation, energy. The next step is to classify the image using the Linear Discriminant Analysis and Support Vector Machine methods. The following displays the test results from one of the palm fruit image data.

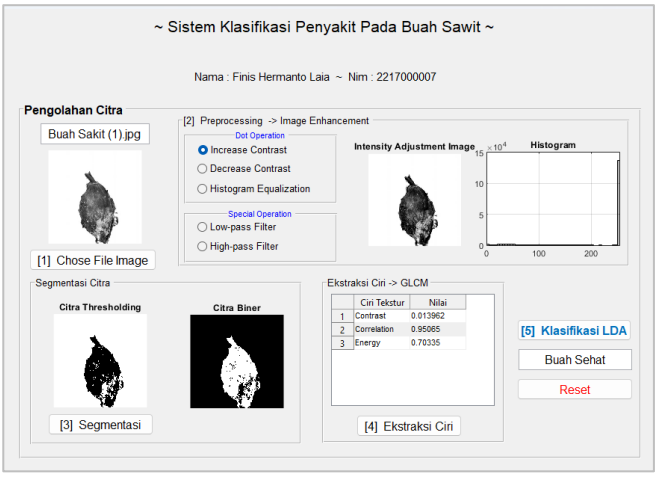

Fig. 8 Display of test results

Testing uses test data of 120 images, second 96 images and third 72 images. There are two classes, namely sick fruit and healthy fruit, all test images will be matched with the classification results.

C. Cross Validation Results

The algorithm used to test the prediction of the validity of the accuracy results is Cross validation. To test the validity of prediction accuracy results with Cross validation, K=5 and K=10 iterations were used, where iterations were carried out 5 times, 1 to 5 epochs and 1 to 10 epochs. For testing dataset selection, it is adjusted to the order of iteration with fold. The iteration process is carried out five times based on the number of variations in the percentage of image data, the first percentage is 50% training data of 90 images, 50% of test data is 90 images, the second percentage variation is 60% of training data from 180, 108 images, 40% test data from 180 , as many as 72 images and the third percentage variation is 70% training data from 180, as many as 126 images, 30% test data from 180, as many as 54 images. From the total training data and data after each training, testing is immediately carried out to find the prediction value, then the level of accuracy is calculated on average:

TABLE 3
ACCURACY OF TRAINING RESULTS USING THE K-FOLD CROSS VALIDATION METHOD

Types of ML Models	Iteration To	Training n-1%	Training n-2%	Training n-3%	Average
Linear Discriminant Analysis (LDA)	1	86,11	86,96	85,04	86,04
	2	85,55	85,32	88,86	86,58
	3	87,22	84,37	87,29	86,29
	4	86,66	85,19	88,12	86,66
	5	85	86,96	88,89	86,95
	Average	86,11	85,76	87,64	86,50
	Iteration To	Training n-1%	Training n-2%	Training n-3%	Average
Support Vector Machine (SVM)	1	91,11	91,60	93,66	92,68
	2	92,77	91,55	94,46	92,93
	3	92,77	90,69	94,46	92,64
	4	91,11	92,55	95,23	92,96
	5	92,22	91,60	94,49	92,77
	Average	92,33	91,60	94,46	92,80

Based on table 3 above, the results of cross validation calculations for training data in linear dicriminant analysis classification have the highest average accuracy, namely 87.29% in the 3rd iteration with a training data percentage of 70%. Meanwhile, the lowest average accuracy value was obtained in the 2nd iteration of 84.37% with a training data percentage of 60%. The system accuracy value of 86.50% is obtained from the average results of each iteration. Meanwhile, in calculating the cross validation results of support vector machine training data, the highest average accuracy, namely 94.46%, was obtained in the 2nd iteration with a training data percentage of 70%. The lowest average accuracy value was obtained in the 3rd iteration, namely 90.69%. Meanwhile, the system's accuracy value was 92.80%, obtained from the average results of each iteration.

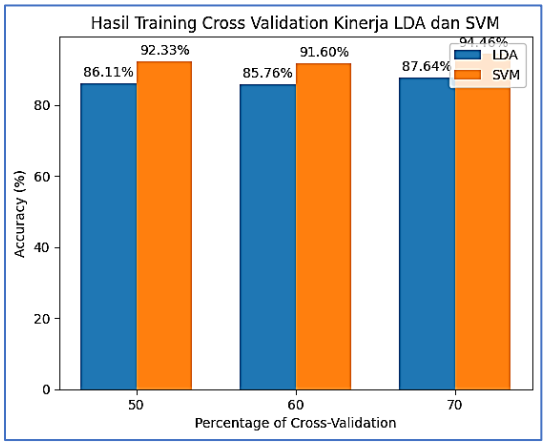

Fig. 9 LDA and SVM Cross validation training graphs

Cross validation data testing is carried out after training in each literacy, following are the results of test data calculations from each ML classification model below:

TABLE 4
ACCURACY OF TESTING RESULTS USING THE K-FOLD CROSS VALIDATION METHOD

Types of ML Models	Iteration To	Testing n-1%	Testing n-2%	Testing n-3%	Average
Linear Discriminant Analysis (LDA)	1	86,0	83,0	76,0	81,67
	2	86,0	79,0	69,0	78,00
	3	86,0	79,0	73,0	79,33
	4	86,0	88,0	73,0	82,33
	5	86,0	79,	66,0%	77,00
	Average	86,00	81,60	71,40	79,67
	Iteration To	Testing n-1%	Testing n-2%	Testing n-3%	Average
Support Vector Machine (SVM)	1	93,33	92,55	96,00	93,96
	2	93,33	92,59	96,03	93,98
	3	93,33	92,64	96,00	93,99
	4	93,33	92,59	96,09	94,00
	5	93,33	92,59	96,00	93,97
	Average	93,33	92,59	96,02	93,98

From table 4. above the average testing results in linear dicriminant analysis, the highest prediction percentage of test data was 30%, namely 73.33%, obtained in the 4th iteration. Meanwhile, the lowest average value was obtained in the 5th iteration, namely 44.0%, in the test data the percentage was 50%. Then the average predicted value for system testing is 79.67%.

Meanwhile, the calculation of the support vector machine testing results has the highest average prediction, namely 96.09%, obtained in the 4th iteration at a test data percentage of 30%. The lowest average accuracy value was obtained in the 1st iteration, namely 92.55%. Then the prediction value for the data testing system was 93.98%, obtained from the average results of each iteration.

The following results from k=5 cross validation are depicted in the following graphical form:

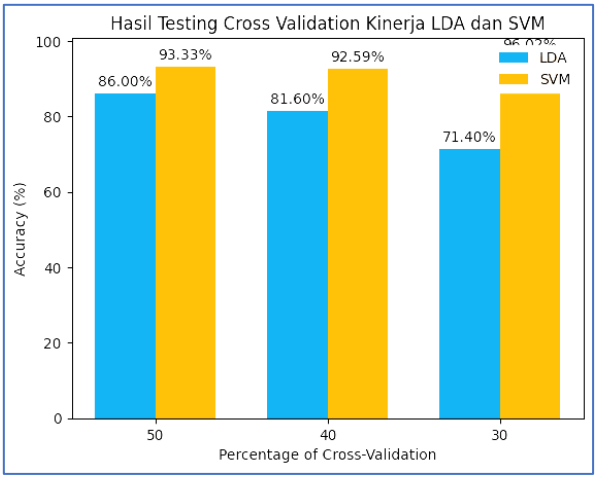

Fig. 10 LDA and SVM Cross validation testing graphs

D. Accuracy Results

Results evaluation was carried out using accuracy, precision, recall and f1-score metrics on the training and test systems to measure the effectiveness and performance of both LDA and SVM algorithms in classifying images of healthy or sick palm fruit. The following are the results of the accuracy of the percentage of training data from the total image data, the first variation is 50% with 90 training images, the second percentage is 60% from 180 images with 108 images, and the third is 70% with 180 images with 126 images of palm fruit.

TABLE 5
LDA AND SVM TRAINING DATA ACCURACY RESULTS

Types of ML Models	Percentage Variation	Amount of Training Data	Accuracy	Precision	Recall	F1-Score
Linear Discriminant Analysis (LDA)	Ke-1	90	86,66 %	98,0 %	74,0 %	84,0 %
	Ke-2	108	88,88 %	100%	77,0 %	87,0 %
	Ke-3	126	89,68 %	100%	80,0 %	88,0 %
Support Vector Machine (SVM)	Ke-1	90	93,33 %	100%	87,0 %	92,0 %
	Ke-2	108	92,59 %	97,0 %	87,0 %	92,0 %
	Ke-3	126	96,03 %	98,0 %	93,0 %	96,0 %

Table. 5. shows the evaluation and effectiveness and performance of two algorithms from training data in classifying palm fruit images. The SVM algorithm shows better accuracy results than LDA. The following is a graph of the accuracy of the training data.

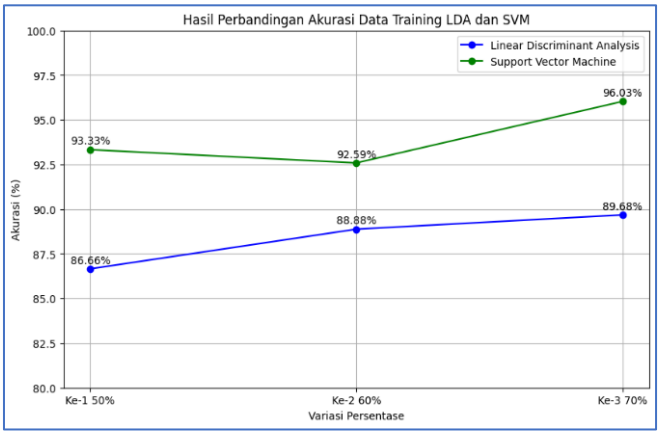

Fig. 11 Graph of LDA and SVM Training percentage variation

From the results in the graph above, SVM has an accuracy of 93.33% at the 1st percentage variation. Meanwhile, LDA is 86.66% with the same percentage variation.

TABLE 6
ACCURACY RESULTS OF LDA AND SVM TESTING DATA

Types of ML Models	Percentage Variation	Amount of Training Data	Accuracy	Precision	Recall	F1-Score
Linear Discriminant Analysis (LDA)	Ke-1	90	86,66 %	100%	74,0 %	84,0 %
	Ke-2	72	60,18 %	100%	80,0 %	89,0 %
	Ke-3	54	38,09 %	100%	77,0 %	87,5 %
Support Vector Machine (SVM)	Ke-1	90	93,33 %	100%	87,0 %	92,0 %
	Ke-2	72	62,03 %	100%	86,0 %	92,0 %
	Ke-3	54	42,85 %	100%	100 %	100 %

In the testing data, the accuracy of the SVM algorithm is higher than LDA in the 1st and 2nd percentage variations. Based on the results of precision, recall and F1-Score parameters, LDA gives results in the 1st percentage variation (50%) with an accuracy of 86.66%, precision 98.0%, recall 74.0%, and F1-Score 84.0 %. SVM gives the best results at the 1st percentage variation (50%) with an accuracy of 93.33%, precision 98.0%, recall 87.0%, and F1-Score 92.0% where SVM tends to be better. The following is a graphic visualization of the accuracy of the data testing process for the two algorithms (LDA and SVM) based on data percentage variations.

The following is a graph of the accuracy of testing data.

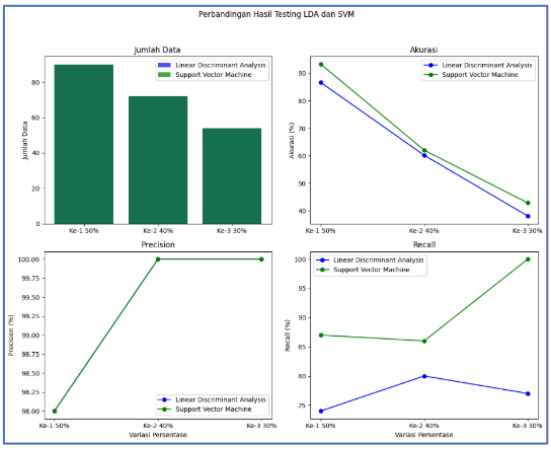

Fig. 12 Graphs of LDA and SVM Testing percentage variations

The research results show that SVM is more effective in classifying images of sick or healthy palm fruit compared to LDA. Results at the data testing process stage of 1st percentage variation (50%), 2nd percentage variation (40%), LDA performance decreased significantly while SVM remained relatively stable, 3rd percentage variation (30%) both models showed performance which is very low based on the number of test datasets used which are not the same as the total training data dataset stored in the model, but with precision, recall and F1-Score values reaching 100% for both algorithms.

IV. CONCLUSIONS

Based on the results of the tests carried out, the implementation of the linear discriminant analysis algorithm and support vector machine in classifying images of healthy or sick palm fruit has been successfully carried out. The results obtained from the test system which produces the highest level of accuracy from the SVM algorithm testing data are better than LDA in terms of percentage variations, the first level of accuracy is 93.33% and the second is 62.03%. Then in the second percentage LDA achieved an accuracy of 60.18% which is close to the high level of SVM. From the results of precision, recall and F1-Score parameters, LDA gives results in the 1st percentage variation (50%) with precision 98.0%, recall 74.0% and F1-Score 84.0%. SVM gives the best results at the 1st percentage variation (50%) with precision 98.0%, recall 87.0%, and F1-Score 92.0% where SVM tends to be better.

REFERENCES

[1] B. Ghaddar and J. Naoum-sawaya, “PT US CR,” Eur. J. Opera. Res., 2017, doi: 10.1016/j.ejor.2017.08.040.

[2] C. Li, Y. Shao, W. Chen, Z. Wang, and N. Deng, “Generalized two-dimensional linear discriminant analysis with regularization,” Neural Networks, vol. 142, pp. 73–91, 2021, doi: 10.1016/j.neunet.2021.04.030.

[3] SF Hussain, “A novel robust kernel for classifying high-dimensional data using Support Vector Machines,” Expert Syst. Appl., vol. 131, pp. 116–131, 2019, doi: 10.1016/j.eswa.2019.04.037.

[4] Z. He, M. Wu, X. Zhao, S. Zhang, and J. Tan, “Representative null space LDA for discriminative dimensionality reduction,” Pattern Recognit., vol. 111, p. 107664,2021, doi: 10.1016/j.patcog.2020.107664.

[5] A. Tasari, DD Tarigan, E. Nia, D. Br, and KS S, "Comparison of Support Vector Machine and KNN Algorithms in Predicting Protein Secondary Structure," vol. 9, no. 2, pp. 172–179, 2022.

[6] RD Yunita, C. Rozikin, and M. Jajuli, "Implementation of the Linear Discriminant Analysis Method for Abstract Coffee Bean Classification," vol. 8, no. 1, pp. 27–39, 2022.

[7] JVCI R, H. Shi, and L. Tao, “Visual comparison based on linear regression models and linear discriminant analysis,” J. Vis. Commun. Image Represent., vol. 57, pp. 118– 124, 2018, doi: 10.1016/j.jvcir.2018.10.026.

[8] Y. Wei, K. Gu, and L. Tan, “A positioning method for maize seed laser-cutting slices using linear discriminant analysis based on isometric distance measurement,” Inf. Process. Agric., vol. 9, no. 2, pp. 224–232, 2022, doi: 10.1016/j.inpa.2021.05.002.

[9] M. Scholz and T. Wimmer, “Jou rna,” Expert Syst. Appl., p. 114217, 2020, doi: 10.1016/j.eswa.2020.114217.

[10] T. Kavzoglu and I. Colkesen, “The effects of training set size for performance of support vector machines and decision trees The effects of training set size for performance of support vector machines and decision trees,” no. April 2015, 2012.

[11] V. Ashok, RK Bharathi, and P. Shivakumara, “Building a Medium Scale Dataset for Non-destructive Disease Classification in Mango Fruits Using Machine Learning and Deep Learning Models,” Int. J. Image, Graph. Signal Process., vol. 15, no. 4, pp. 83– 95, 2023, doi: 10.5815/ijigsp.2023.04.07.

[12] X. Yang, R. Zhang, Z. Zhai, Y. Pang, and Z. Jin, “Machine learning for cultivar classification of apricots (Prunus armeniaca L.) based on shape features,” Sci. Hortic. (Amsterdam)., vol. 256, no. May, p. 108524, 2019, doi: 10.1016/j.scienta.2019.05.051.