Gmail

Indra Laksmana <indra.puskom@gmail.com>

Proofreading result

2 pesan

icgefast icgefast <icgefast@gmail.com> 5 Februari 2022 pukul 07.07
Kepada: indra.puskom@gmail.com

Dear Presenters
We have finished proofreading the manuscript, please recheck the suggestions from proofreading, if you agree, please agree to the comments in the manuscript.
We've also attached how to approve comments. thank you

2 lampiran

Indra Laksmana.docx
71K

proofread.pdf
173K

Indra Laksmana <indra.puskom@gmail.com> 5 Februari 2022 pukul 08.26
Kepada: icgefast icgefast <icgefast@gmail.com>

ok, Done
thank you
[Kutipan teks disembunyikan]

Indra Laksmana.docx
71K

MACHINE LEARNING AS SEED IMAGE IDENTIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Indra Laksmana[1], Hendra[1], Jamaluddin[1], Trinovita Zuhara J[1], M Riza Nurtam[1], Amrizal[1], Rosda Syelly[2]

[1]Computer Engineering Study Program, Politeknik Pertanian Negeri Payakumbuh, Jalan Raya Negara Km 7, Tanjung~~P~~pati, ~~West~~ Sumat~~e~~ra Barat, Indonesia 26271
[2][1]Computer Engineering Study Program~~Program Studi Teknik Komputer~~, Sekolah Tinggi Teknologi Payakumbuh, Jalan Khatib Sulaiman Sawah Padang Payakumbuh, ~~West~~ Sumat~~e~~ra Barat, Indonesia 26227

*Corresponden Email~~ ~~: indra.puskom@gmail.com

Abstract. ~~We can identify p~~Plants can be identified using several variables, ~~including~~ such as seeds' shapes, colors, and sizes. ~~But~~ However, ~~there are~~ several types of plants hav~~e~~ing close similarities ~~in~~ to seed shapes. Therefore, ~~so one will need~~ additional characteristics are necessary to support the identification process. This study applie~~d~~s machine learning with the PCA method to identify plant species from seed shapes. The PCA simplifies the observed variables by reducing data dimensions and storing 75% of the information. The procedure ~~will~~ did not eliminate too much important information while reducing data size and processing time. We collected 100 images of plant seeds ~~that are~~ similar to one another, such as sapodilla seeds, soursop, cucumber, star fruit, grape, melon, apple, lime, watermelon, and chili. A measurement system was designed using the K-Fold Cross Validation, and ~~with~~ 10 tables of experimental results ~~yields~~ discovered a good level of accuracy of 83%. The Omission error occurred in the seeds of soursop, starfruit, grape, apple, lime, and watermelon~~,~~ while the most commission errors occurred in apple seeds (8 times).
Keywords: Seed Image Identification; Machine Learning; Principal Component Analysis

1. Introduction

Fruit plants are one of the plantation crops that are widely planted by farmers in Indonesia. These ~~production of this~~ plants ~~is~~ are usually produced ~~used both for~~ to meet self-sufficiency and ~~to meet~~ local market needs [1]. In general, fruit plants are planted around the house or in pots as garden decorations. [2]. ~~The~~ Farmers commonly propagate ~~ion of~~ fruit plants ~~that is still commonly carried out by farmers is~~ by planting seeds. ~~Some of the advantages of p~~Planting with seeds ~~include~~ provides a strong root system, ~~easier to do~~a simple technique, and a long period of fruiting~~,~~; however, ~~although~~ the nature of the offspring is not the same as that of the ~~parent~~mother plant [3].

The shapes and structures of the seeds vary~~,~~ depending on the types of species and environmental conditions in which the plant lives [4]. ~~So that~~Therefore, the seeds can be used as a marker ~~in~~ to differentiate~~ing~~ plant species [5]. ~~But~~ However, the wide diversity of forms ~~is so much~~ sometimes ~~also difficult~~ complicates ~~to~~ the identification~~y~~ of a particular plant. In general, there are four ways to identify plants ~~.~~[6]. However, these ways ~~This will~~ complicate the manual classification ~~manually, this process~~ because they require~~s~~ a long time and special understanding. ~~So~~ Therefore, we need a system ~~with a technique to~~ that can simplify the data ~~in~~ when identifying plants. One of the techniques ~~used~~ to simplify the data ~~used~~ as an image classifier is ~~to use~~ the ~~P~~principal ~~C~~component ~~A~~analysis (PCA).

The PCA is used to reduce the dimensions of data without significantly reducing the characteristics of the data [7]~~.~~. According to Santosa [8], ~~Principal Component Analysis~~the PCA is a reliable technique for extracting the structure of a data set with ~~quite a lot of~~ mass dimensions. The PCA projects ~~the~~ images into ~~its~~ eigenspace planes by finding the eigenvectors of each image; then ~~and~~ project~~sing~~ them into the obtained eigenspace. Research ~~conducted~~ by [9] ~~in his journal~~entitled "Plant classification based on leaf recognition" ~~that uses~~ revealed that the PCA method ~~can~~ could classify plants with an accuracy rate of more than 90%. Moreover, [10] ~~have also identified~~ discovered medicinal plant leaf images with an accuracy of more than 90%. [11] used the PCA to identify typical compounds from various Eucalyptus plant species and ~~using PCA~~ obtained that the compound groupings ~~that~~ could clearly describe~~d~~ the specifications of each species.

Commented [Prsmntc1]: Apa yang mempersulit? Kita butuh objek pada kalimat ini.
Sementara saya tambahkan keempat cara yg disebutkan pada kalimat sebelumnya

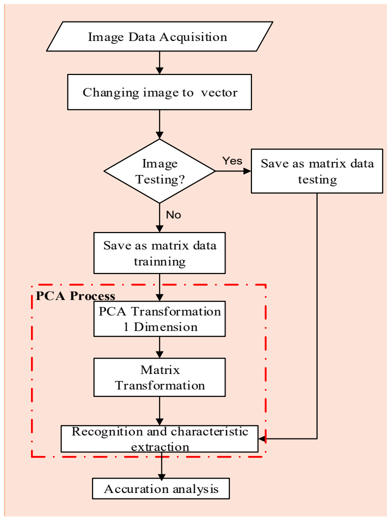

Figure 1. Research M~~m~~ethod

Commented [Prsmntc2]: Karena tdk bisa saya edit mohon diganti menjadi (dari atas sampai bawah):
Huruf diseragamkan saja. mau kapital each word atau tidak. sementara saya buat capilazied each word (kecuali preposition)
1. Acquiring Image Data
2. Changing Images to Vectors
3.Image Testing
4.Saved as Matrix Data Testing
5. Saved as Matrix Data Training
6.PCA Transformation of 1 Dimension
7.Matrix Transformation
8.Recognizing and Characterizing Extraction
9.Analyzing Accuracy

This study applie~~ds~~ the Principal Component Analysis (PCA) to identify plant species based on seed images. ~~With Principal Component Analysis,~~ The PCA ~~the~~ simplifies images of ~~the~~ seeds ~~can be simplified~~ by reducing the dimensions and not losing a lot of ~~important~~ crucial information ~~contained in it~~of the images.

2. Methodology

The stages of this research ~~are starting from~~were collecting or acquiring image data ~~collection or data acquisition~~, ~~then~~ converting the image data into vectors, sharing the data (training and testing), conducting the PCA~~Principal Component Analysis process~~, and ~~finally~~ analyzing the accuracy. These stages are shown in Figure1.

Image Data

~~In t~~This study~~,~~ employed 100 images ~~were used~~, consisting of 10 types of plants, each of which used 10 images. Seed images were taken under the same lighting conditions and distances; ~~so as not to significantly affect~~ the quality of the data and test results were not significantly affected [12] [13]. Ten types of plants ~~used~~ in this research had~~ve~~ morphologically similar seeds, such as sapodilla (*Manilkara zapota*), soursop (*Annona muricata*), cucumber (*Cucumis sativus*), star fruit (*Averrhoa carambola*), grape (*Vitis vinifera*), melon (*Cucumis melo*), apple (*Malus domestica*), lime (*Citrus aurantifolia*), watermelon (*Manilkara zapota*) and chili (*Capsicum annum L.*). All images awere saved in JPG formats and RGB modes with image dimensions of 100x75 pixels.

Image rReading Image and data sSplitting Data

From the RGB mode, the image data is were converted to greyscales. The matrix of an image is was saved as a 1x7500 row vector. This process is was repeated for the entire image. Then, and the image data is were combined into an *X* matrix *X* and an *O* matrix *O*. Matrix *X* is referred to a collection of row vectors for training images, while matrix *O* isreferred to a collection of row vectors for testing images.

Formatted: Font: Not Italic

The Iimage-sharing using applied the k-fold cross-validation with a value of k=10, also known as 10-fold cross-validation. The data is were divided into two parts:, namely, a training set of 9 images and a testing set of 1 image. So Therefore, there awere 9 x 10 = 90 images for the training data. Meanwhile, as athe testing image consisted of as much as 1x10 = 10 images. The first experiment is to used the first image as a testing image, and so forth. Thus, and so on until the tenth experiment is to used the tenth image as a testing image.

In one experiment, the process of reading the training image was repeated 9 nine times for 10 ten seed images, to obtain a combined matrix measuring 90x7500. The testing image in an experiment is was an exception or other than the data from the *X* matrix *X* because there is one image from each class (seed) which will would servebe as the training image or which iswould not be processed when reading the *X* matrix *X*., then there will be Meanwhile, 10ten images that will would become row vectors. The next stage was dividinge the matrix *O*; thus, so that the matrix *O* is served as a composite matrix of testing images with a size of 10x7500.

Formatted: Font: Not Italic

The Process of Principal Component Analysis (PCA) Process

The PCA procedure aims to simplify the observed variables by reducing their dimensions [14]. A digital image consists of the smallest elements which are usually called pixels. The pixels store information in the form of the color intensity of the images at these coordinates. The images can be translated as a matrix while pixels can be translated as elements of the matrix [15].

In Artificial Iintelligence, uses many use images as inputs which are processed in clustering and classification. The image or image is defined as a two-dimensional function *f(x,y)* where *x* and y are the spatial coordinates; and the amplitude f, at either x or y, is called the intensity or degree of gray. If x, y, and f are finite, the image ~~can be said to be~~ is considered a digital image [16].

In this study, seed image was detected~~ion is carried out~~ in several stages:~~, namely,~~ reading the image~~s reading process~~, determining the covariance matrix, determining the feature vector and ~~feature~~ roots that correspond to the covariance matrix, selecting the largest feature root ~~so that it~~to contribute~~s~~ to the desired information, ~~and~~ determining the matrix transformation ~~matrix~~ using the characteristic vectors according to the root of the feature, determining~~e~~ the resulted~~ing~~ matrix from the transformation of the seed image, determining~~e~~ the average vector of the transformation results for each class (plant species), determining~~e~~ the Euclidean distance between the testing image and each class average vector, ~~then~~ and recognizing the images ~~recognition~~ based on the minimum Euclidean distance. ~~If~~ When the minimum distance of the image ~~is~~ was obtained from the same class average vector, ~~then~~ the image ~~is~~ was recognized.

After ~~the *X*~~ matrix *X* ~~is~~ had been obtained, the covariance was calculated~~ion process is carried out,~~ to obtain the *sigma* matrix. The calculation of this covariance used~~s~~ the equation of $covariance(y_i) = \sigma_{y_i}^2 = a_i{}^T \Sigma a_i$. To ~~perform~~ calculate the covariance ~~calculations~~, the previous unit type matrix data must have been changed to the double type; thus, ~~so that~~ the value range of the matrix elements changed~~s~~ from 0-255 to 0.0-1.

~~After obtaining the sigma covariance matrix, t~~The next step after obtaining the sigma covariance matrix ~~is to~~ was determining~~e~~ the feature vectors and ~~the feature~~ roots of the covariance matrix. The resulted~~ing~~ output ~~i~~was a matrix v and d; each showed~~s~~ a feature vector and ~~a feature~~ root. ~~It~~ This step also determined~~s~~ the diagonal of the matrix d. The output of the diagonal matrix d ~~is~~ was lambda, ~~which is~~ referring to the characteristic root value of the *sigma* covariance matrix.

Furthermore, after obtaining the values of the feature roots and ~~feature~~ vectors, the ~~feature or~~ features were extracted~~ion process is carried out~~ from the image. Several characteristic roots ~~a~~were taken, and they ~~which a~~were large enough to represent the desired information. ~~In t~~This study, took ~~the information taken is~~ 75% of the initial information.

3. Results and Discussion

Commented [Prsmntc3]: Apa yang determined the diagonal?

Akat aIt blm jelas jadi kita ganti saja menjadi kata bendanya. Sementara saya ganti This step

~~Looking at t~~The ~~results of the~~ computer program ~~that has been~~ developed to run the PCA algorithm and the 10-fold cross-validation revealed~~,~~ ~~we get~~ a matrix from ~~the introduction of~~ the ~~1st~~ first ~~seed~~ to the ~~10th~~ tenth seed of the training ~~sample~~ and detection samples. The results ~~can be seen~~are summarized in Table 1.~~ below~~

Formatted: Font color: Red

Table 1. Results of Image Recognition ~~Results~~ for 10 Experiments

Experiment	Original ~~i~~Images									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Detected as									
1	1	2	3	4	7	6	5	8	7	10
2	1	2	3	4	4	6	7	8	7	10
3	1	2	3	4	5	6	7	8	7	10
4	1	2	3	4	5	6	4	8	7	10
5	1	2	3	8	9	6	7	8	9	10
6	1	5	3	4	7	6	7	6	7	10
7	1	2	3	4	7	6	7	8	9	10
8	1	2	3	9	5	6	7	8	9	10
9	1	2	3	4	5	6	9	8	9	10
10	1	2	3	4	5	6	9	8	9	10

~~In~~ Table 1~~,~~ shows that the images of seeds 1 (sapodilla), 3 (cucumber), 6 (melon), and 10 (chili) in each experiment ~~showed~~ did not have errors ~~in~~ when detecting the image. ~~Meanwhile~~However, image 5 (grape) was detected as an image of soursop seeds in the ~~6th~~ sixth experiment and ~~was detected~~ as an image of apple seeds in the first experiment. In the ~~5~~fifth experiment, the image of lime seeds was detected as an image of star fruit seeds. The image of star fruit seeds was detected as an image of grape seeds in the ~~2~~second experiment and ~~was detected~~ as an image of apples in the ~~4~~fourth experiment. The image of watermelon seeds was detected as an image of star fruit seeds and apples. ~~W~~Meanwhile, the image of apple seeds was detected as the images of grape seeds and watermelons. The shape of the seed images and the patterns of image data ~~that are~~ taken from the same angle influence~~d~~s this ~~detection~~ error detection.

We iterated the experiment ten times to get information on how many times ~~will~~ the system would accurately detect the plant seeds from ~~its~~ their images. The sum of each element from the experimental results ~~table can be used to~~ shows the accuracy of the image detection system. Table 2 ~~is~~ summarizes the sum of all experiments. The main diagonal shows the number of correct image detections ~~correctly~~ in 10 trials.

Table 2. ~~The r~~Result of the ~~s~~Sum of ~~a~~All ~~t~~Trials

Label	Detected as									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	10	0	0	0	0	0	0	0	0	0
2	0	9	0	0	1	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0	0
4	0	0	0	8	0	0	0	1	1	0
5	0	0	0	0	6	0	3	0	1	0
6	0	0	0	0	0	10	0	0	0	0
7	0	0	0	1	1	0	6	0	2	0
8	0	0	0	0	0	1	0	9	0	0
9	0	0	0	0	0	0	5	0	5	0
10	0	0	0	0	0	0	0	0	0	10

Accuracy ~~a~~Analysis

The ~~calculation of~~ accuracy ~~for~~ of the seed image detection system can be calculated ~~found~~ by counting the number of diagonals in the overall result table, ~~and~~ dividing them~~ed~~ by a lot of data, and multiplyin~~ged~~g them by 100%. The built system ~~that has been built~~ has an accuracy of 83%.

Table 3. Omission ~~e~~Errors

Plants	Seed ~~n~~Number	Omission Errors
Sapodilla (*Manilkara zapota*)	1	0
~~s~~Soursop (*Annona muricata*)	2	1
~~c~~Cucumber (*Cucumis sativus*)	3	0
~~s~~Starfruit (*Averrhoa carambola*)	4	2
~~g~~Grape (*Vitis vinifera*)	5	4
Melon (*Cucumis melo*)	6	0
~~a~~Apple (*Malus domestica*)	7	4
~~l~~Lime (*Citrus aurantifolia*)	8	1
~~w~~Watermelon (*Citrullus lanatus*)	9	5
~~c~~Chili (*Capsicum annum L.*)	10	0

Omission and Commission Error

~~Omission~~ The omission ~~is~~ constitutes the number of testing images that do not match the ~~class~~ classification category. Meanwhile, ~~C~~the commission is the number of detected testing images that ~~are not in accordance~~disagree with the actual situation/classification in the class. The omission error can be calculated by adding up every row of the resulted matrix, except the elements ~~o~~in the main diagonal. The output of the omission error is ~~as~~ shown in Table 3. ~~below~~

~~The m~~Most omission errors occurred in watermelon seeds~~,~~ (see Table 3). ~~From 10 experiments conducted, w~~Watermelon seeds were detected ~~5 times~~ as seeds of other plants in five of ten experiments. ~~W~~Meanwhile, the seeds of ~~S~~sapodilla, cucumber, ~~M~~melon, and chili did not occur omission.~~,~~ ~~t~~This phenomenon indicated ~~means~~ that the system had~~s~~ successfully and correctly detected the seeds~~it correctly~~. The commission errors can be calculated by adding up every column of the resulted matrix, except the elements on the main diagonal. The output of the commission errors is shown in Table 4~~ below~~.

~~Meanwhile in~~ Table 4~~,~~ shows that ~~out of 10 experiments conducted by the system, the system carried out the~~ most commission errors of ten experiments ~~on~~ occurred in apple seeds~~,~~. This phenomenon denoted ~~meaning~~ that the system had detected other seeds as themselves or ~~in other words, got~~ detected~~ion~~ errors from other plant seeds ~~8~~eight times. Meanwhile, sapodilla, soursop, cucumber seeds, and chili seeds never received errors ~~of~~ when detected~~ion~~ from other seeds.

Table 4. Commission ~~e~~Errors

Plants	Seed number	Commission Errors
Sapodilla (*Manilkara zapota*)	1	0
~~s~~Soursop (*Annona muricata*)	2	0
~~c~~Cucumber (*Cucumis sativus*)	3	0
~~s~~Starfruit (*Averrhoa carambola*)	4	1
~~g~~Grape (*Vitis vinifera*)	5	2
Melon (*Cucumis melo*)	6	1
~~a~~Apple (*Malus domestica*)	7	8
~~l~~Lime (*Citrus aurantifolia*)	8	1
~~w~~Watermelon (*Citrullus lanatus*)	9	4
~~c~~Chili (*Capsicum annum L.*)	10	0

Formatted: Font: Not Bold

4. Conclusion

The results of this study concluded that the PCA technique ~~can~~ could be used ~~for~~ to reduce dimensions~~ reduction~~. The information taken ~~is~~ was 75% of the 100 processed plant seed images~~,~~. This study obtained~~ing~~ good results~~,~~ ~~namely~~ because ~~successfully recognizing~~ the images were successfully ~~-~~and correctly recognized with a percentage rate of 83%. Image retrieval techniques are urgently needed so that the images taken as data have similar lighting, distances, and dimensions. Moreover, ~~identifying, assessing, and interpreting~~ all available research evidence is crucially identified, assessed, and interpreted~~ in order~~ to ~~provide~~ answers ~~to~~ specific research questions.

References

[1] N. L. M. Pradnyawathi, I. K. Sardiana, and N. N. Darmiati, “Pengembangan Bibit Buah Lokal Unggul Wani Bali Tanpa Biji,” *Bul. Udayana Mengabdi*, vol. 19, no. 1, Feb. 2020, doi: 10.24843/BUM.2020.v19.i01.p06.

[2] M. R. Suhartanto and E. Gunawan, *Untung besar dari bisnis bibit tanaman buah*. Jakarta: AgroMedia Pustaka, 2012.

[3] A. Barus and Syukri, *Agroteknologi Tanaman Buah-buahan*. Medan: USU Press, 2008.

[4] A. R. Muhammad, “Hubungan Bentuk Biji Dengan Karakteristik Morfologi Bibit Tanaman Pala (Myristica fragrans H.),” Andalas University, 2016.

[5] I. Laksmana, Y. Herdiyeni, and E. A. M. Zuhud, “Genetic Programming for Medicinal Plant Family Identification System,” *J. ICT Res. Appl.*, vol. 7, no. 3, pp. 217–234, Dec. 2013, doi: 10.5614/itbj.ict.res.appl.2013.7.3.4.

[6] I. P. B. Laboratorium Ekologi Hutan, Fakultas Kehutanan, *Diktat Dendrologi*. Bogor: Laboratorium Ekologi Hutan, Fakultas Kehutanan, Insitut Pertanian Bogor, 2004.

[7] L. I. Smith, *A tutorial on Principal Components Analysis*. John Wiley & Sons Inc, 2002.

[8] S. B, *Data Mining (Teori dan Aplikasi)*. Yogyakarta: Graha Ilmu, 2007.

[9] A. Ehsanirad, “Plant Classification Based on Leaf Recognition,” 2010.

[10] I. Laksmana *et al.*, “Identification of Tropical Plants Leaves Image Base on Principal Component Analysis,” *J. Appl. Agric. Sci. Technol.*, vol. 4, no. 1, pp. 64–75, Feb. 2020, doi: 10.32530/jaast.v4i1.156.

[11] P. J. Dunlop, C. M. Bignell, J. F. Jackson, and D. B. Hibbert, “Chemometric analysis of gas chromatographic data of oils from Eucalyptus species,” *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 59–67, Nov. 1995, doi: 10.1016/0169-7439(95)00036-4.

[12] K. Delac, M. Grgic, and T. Kos, “Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions.”

[13] V. V. Starovoitov, D. I. Samal, and D. V. Briliuk, “Image Enhancement for Face Recognition,” 2003.

[14] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[15] A. Rahmad, “Perbandingan Metode Ekstraksi Ciri FFT, PCA, dan FPE dalam Pengenalan Karakter Tulisan Tangan,” INSTITUT PERTANIAN BOGOR, 2012.

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

MACHINE LEARNING AS SEED IMAGE IDENTIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Indra Laksmana[1], Hendra[1], Jamaluddin[1], Trinovita Zuhara J[1], M Riza Nurtam[1], Amrizal[1], Rosda Syelly[2]

[1]Computer Engineering Study Program, Politeknik Pertanian Negeri Payakumbuh, Jalan Raya Negara Km 7, Tanjungpati, West Sumatera, Indonesia 26271
[2]Sekolah Tinggi Teknologi Payakumbuh, Khatib Sulaiman Sawah Padang street Payakumbuh, West Sumatera, Indonesia 26227

*Corresponden Email: indra.puskom@gmail.com

Abstract. Plants can be identified using several variables, such as seeds' shapes, colors, and sizes. However, several types of plants have close similarities to seed shapes. Therefore, additional characteristics are necessary to support the identification process. This study applied machine learning with the PCA method to identify plant species from seed shapes. The PCA simplifies the observed variables by reducing data dimensions and storing 75% of the information. The procedure did not eliminate too much important information while reducing data size and processing time. We collected 100 images of plant seeds similar to one another, such as sapodilla seeds, soursop, cucumber, star fruit, grape, melon, apple, lime, watermelon, and chili. A measurement system was designed using the K-Fold Cross Validation, and 10 tables of experimental results discovered a good level of accuracy of 83%. The Omission error occurred in the seeds of soursop, starfruit, grape, apple, lime, and watermelon while the most commission errors occurred in apple seeds (8 times).
Keywords: Seed Image Identification; Machine Learning; Principal Component Analysis

1. Introduction

Fruit plants are one of the plantation crops that are widely planted by farmers in Indonesia. These plants are usually produced to meet self-sufficiency and local market needs [1]. In general, fruit plants are planted around the house or in pots as garden decorations. [2]. Farmers commonly propagate fruit plants by planting seeds. Planting with seeds provides a strong root system, a simple technique, and a long period of fruiting; however, the nature of the offspring is not the same as that of the mother plant [3].

The shapes and structures of the seeds vary depending on the types of species and environmental conditions in which the plant lives [4]. Therefore, the seeds can be used as a marker to differentiate plant species [5]. However, the wide diversity of forms sometimes complicates the identification of a particular plant. In general, there are four ways to identify plants [6]. However, these ways complicate the manual classification because they require a long time and special understanding. Therefore, we need a system that can simplify the data when identifying plants. One of the techniques to simplify the data as an image classifier is the principal component analysis (PCA).

The PCA is used to reduce the dimensions of data without significantly reducing the characteristics of the data [7]. According to Santosa [8], the PCA is a reliable technique for extracting the structure of a data set with mass dimensions. The PCA projects images into eigenspace planes by finding the eigenvectors of each image; then projects them into the obtained eigenspace. Research by [9] entitled "Plant classification based on leaf recognition" revealed that the PCA method could classify plants with an accuracy rate of more than 90%. Moreover, [10] discovered medicinal plant leaf images with an accuracy of more than 90%. [11] used the PCA to identify typical compounds from various Eucalyptus plant species and obtained that the compound groupings could clearly describe the specifications of each species.

This study applied the Principal Component Analysis (PCA) to identify plant species based on seed images. The PCA simplifies images of seeds by reducing the dimensions and not losing a lot of crucial information of the images.

2. Methodology

The stages of this research were collecting or acquiring image data, converting the image data into vectors, sharing the data (training and testing), conducting the PCA, and analyzing the accuracy. These stages are shown in Figure1.

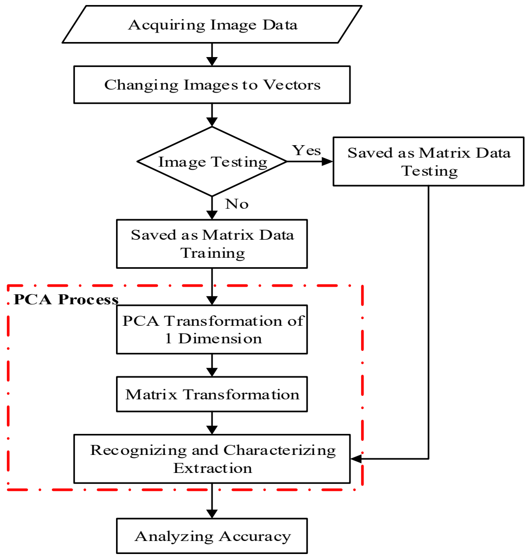

Figure 1. Research Method

Image Data

This study employed 100 images, consisting of 10 types of plants, each of which used 10 images. Seed images were taken under the same lighting conditions and distances; the quality of the data and test results were not significantly affected [12] [13]. Ten types of plants in this research had morphologically similar seeds, such as sapodilla (*Manilkara zapota*), soursop (*Annona muricata*), cucumber (*Cucumis sativus*), star fruit (*Averrhoa carambola*), grape (*Vitis vinifera*), melon (*Cucumis melo*), apple (*Malus domestica*), lime (*Citrus aurantifolia*), watermelon (*Manilkara zapota*) and chili (*Capsicum annum L.*). All images were saved in JPG formats and RGB modes with image dimensions of 100x75 pixels.

Reading Image and Splitting Data

From the RGB mode, the image data were converted to greyscales. The matrix of an image was saved as a 1x7500 row vector. This process was repeated for the entire image. Then, the image data were combined into matrix X and matrix O. Matrix X referred to a collection of row vectors for training images, while matrix O referred to a collection of row vectors for testing images.

The image-sharing applied the k-fold cross-validation with a value of k=10, also known as 10-fold cross-validation. The data were divided into two parts: a training set of 9 images and a testing set of 1 image. Therefore, there were 9 x 10 = 90 images for the training data. Meanwhile, the testing image consisted of 1x10 = 10 images. The first experiment used the first image as a testing image, and so forth. Thus, the tenth experiment used the tenth image as a testing image.

In one experiment, the process of reading the training image was repeated nine times for ten seed images to obtain a combined matrix measuring 90x7500. The testing image in an experiment was an exception or other than the data from the matrix *X* because one image from each class (seed) would serve as the training image or would not be processed when reading the matrix *X*. Meanwhile, ten images would become row vectors. The next stage was dividing the matrix *O;* thus, the matrix *O* served as a composite matrix of testing images with a size of 10x7500.

The Process of Principal Component Analysis (PCA)

The PCA procedure aims to simplify the observed variables by reducing their dimensions [14]. A digital image consists of the smallest elements usually called pixels. The pixels store information in the form of the color intensity of images at these coordinates. The images can be translated as a matrix while pixels can be translated as elements of the matrix [15].

Artificial intelligence uses many images as inputs which are processed in clustering and classification. The image is defined as a two-dimensional function *f(x,y)* where *x* and *y* are the spatial coordinates; and the amplitude *f*, at either *x* or *y*, is called the intensity or degree of gray. If *x*, *y*, and *f* are finite, the image is considered a digital image [16].

In this study, seed image was detected in several stages: reading the images, determining the covariance matrix, determining the feature vector and roots that correspond to the covariance matrix, selecting the largest feature root to contribute to the desired information, determining the matrix transformation using the characteristic vectors according to the root of the feature, determining the resulted matrix from the transformation of the seed image, determining the average vector of the transformation results for each class (plant species), determining the Euclidean distance between the testing image and each class average vector, and recognizing the images based on the minimum Euclidean distance. When the minimum distance of the image was obtained from the same class average vector, the image was recognized.

After matrix *X* had been obtained, the covariance was calculated to obtain the *sigma* matrix. The calculation of this covariance used the equation of $covariance(y_i) = \sigma_{y_i}^2 = {a_i}^T \Sigma a_i$. To calculate the covariance, the previous unit type matrix data must have been changed to the double type; thus, the value range of the matrix elements changed from 0-255 to 0.0-1.

The next step after obtaining the sigma covariance matrix was determining the feature vectors and roots of the covariance matrix. The resulted output was a matrix *v* and *d*; each showed a feature vector and root. This step also determined the diagonal of the matrix *d*. The output of the diagonal matrix *d* was lambda, referring to the characteristic root value of the *sigma* covariance matrix.

Furthermore, after obtaining the values of the feature roots and vectors, the features were extracted from the image. Several characteristic roots were taken, and they were large enough to represent the desired information. This study took 75% of the initial information.

3. Results and Discussion

The computer program developed to run the PCA algorithm and the 10-fold cross-validation revealed a matrix from the first to the tenth seed of the training and detection samples. The results are summarized in Table 1.

Table 1. Results of Image Recognition for 10 Experiments

Experiment	**Original Images**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Detected as									
1	1	2	3	4	7	6	5	8	7	10
2	1	2	3	4	4	6	7	8	7	10
3	1	2	3	4	5	6	7	8	7	10
4	1	2	3	4	5	6	4	8	7	10
5	1	2	3	8	9	6	7	8	9	10
6	1	5	3	4	7	6	7	6	7	10
7	1	2	3	4	7	6	7	8	9	10
8	1	2	3	9	5	6	7	8	9	10
9	1	2	3	4	5	6	9	8	9	10
10	1	2	3	4	5	6	9	8	9	10

Table 1 shows that the images of seeds 1 (sapodilla), 3 (cucumber), 6 (melon), and 10 (chili) in each experiment did not have errors when detecting the image. However, image 5 (grape) was detected as an image of soursop seeds in the sixth experiment and as an image of apple seeds in the first experiment. In the fifth experiment, the image of lime seeds was detected as an image of star fruit seeds. The image of star fruit seeds was detected as an image of grape seeds in the second experiment and as an image of apples in the fourth experiment. The image of watermelon seeds was detected as an image of star fruit seeds and apples. Meanwhile, the image of apple seeds was detected as the images of grape seeds and watermelons. The shape of the seed images and the patterns of image data taken from the same angle influenced this error detection.

We iterated the experiment ten times to get information on how many times the system would accurately detect the plant seeds from their images. The sum of each element from the experimental results shows the accuracy of the image detection system. Table 2 summarizes the sum of all experiments. The main diagonal shows the number of correct image detections in 10 trials.

Table 2. Result of the Sum of All Trials

Label	**Detected as**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	10	0	0	0	0	0	0	0	0	0
2	0	9	0	0	1	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0	0
4	0	0	0	8	0	0	0	1	1	0
5	0	0	0	0	6	0	3	0	1	0
6	0	0	0	0	0	10	0	0	0	0
7	0	0	0	1	1	0	6	0	2	0
8	0	0	0	0	0	1	0	9	0	0
9	0	0	0	0	0	0	5	0	5	0
10	0	0	0	0	0	0	0	0	0	10

Accuracy Analysis

The accuracy of the seed image detection system can be calculated by counting the number of diagonals in the overall result table, dividing them by a lot of data, and multiplying them by 100%. The built system has an accuracy of 83%.

Table 3. Omission Errors

Plants	**Seed Number**	**Omission Errors**
Sapodilla (*Manilkara zapota*)	1	0
Soursop (*Annona muricata*)	2	1
Cucumber (*Cucumis sativus*)	3	0
Starfruit (*Averrhoa carambola*)	4	2
Grape (*Vitis vinifera*)	5	4
Melon (*Cucumis melo*)	6	0
Apple (*Malus domestica*)	7	4
Lime (*Citrus aurantifolia*)	8	1
Watermelon (*Citrullus lanatus*)	9	5
Chili (*Capsicum annum L.*)	10	0

Omission and Commission Error

The omission constitutes the number of testing images that do not match the classification category. Meanwhile, the commission is the number of detected testing images that disagree with the actual situation/classification in the class. The omission error can be calculated by adding up every row of the resulted matrix, except the elements in the main diagonal. The output of the omission error is shown in Table 3.

Most omission errors occurred in watermelon seeds (see Table 3). Watermelon seeds were detected as seeds of other plants in five of ten experiments. Meanwhile, the seeds of sapodilla, cucumber, melon, and chili did not occur omission. This phenomenon indicated that the system had successfully and correctly detected the seeds. The commission errors can be calculated by adding up every column of the resulted matrix, except the elements on the main diagonal. The output of the commission errors is shown in Table 4.

Table 4 shows that most commission errors of ten experiments occurred in apple seeds. This phenomenon denoted that the system had detected other seeds as themselves or detected errors from other plant seeds eight times. Meanwhile, sapodilla, soursop, cucumber seeds, and chili seeds never received errors when detected from other seeds.

Table 4. Commission Errors

Plants	Seed number	Commission Errors
Sapodilla (*Manilkara zapota*)	1	0
Soursop (*Annona muricata*)	2	0
Cucumber (*Cucumis sativus*)	3	0
Starfruit (*Averrhoa carambola*)	4	1
Grape (*Vitis vinifera*)	5	2
Melon (*Cucumis melo*)	6	1
Apple (*Malus domestica*)	7	8
Lime (*Citrus aurantifolia*)	8	1
Watermelon (*Citrullus lanatus*)	9	4
Chili (*Capsicum annum L.*)	10	0

4. Conclusion

The results of this study concluded that the PCA technique could be used to reduce dimensions. The information taken was 75% of the 100 processed plant seed images. This study obtained good results because the images were successfully and correctly recognized with a percentage rate of 83%. Image retrieval techniques are urgently needed so that the images taken as data have similar lighting, distances, and dimensions. Moreover, all available research evidence is crucially identified, assessed, and interpreted to answer specific research questions.

References

[1] N. L. M. Pradnyawathi, I. K. Sardiana, and N. N. Darmiati, “Pengembangan Bibit Buah Lokal Unggul Wani Bali Tanpa Biji,” *Bul. Udayana Mengabdi*, vol. 19, no. 1, Feb. 2020, doi: 10.24843/BUM.2020.v19.i01.p06.

[2] M. R. Suhartanto and E. Gunawan, *Untung besar dari bisnis bibit tanaman buah*. Jakarta: AgroMedia Pustaka, 2012.

[3] A. Barus and Syukri, *Agroteknologi Tanaman Buah-buahan*. Medan: USU Press, 2008.

[4] A. R. Muhammad, “Hubungan Bentuk Biji Dengan Karakteristik Morfologi Bibit Tanaman Pala (Myristica fragrans H.),” Andalas University, 2016.

[5] I. Laksmana, Y. Herdiyeni, and E. A. M. Zuhud, “Genetic Programming for Medicinal Plant Family Identification System,” *J. ICT Res. Appl.*, vol. 7, no. 3, pp. 217–234, Dec. 2013, doi: 10.5614/itbj.ict.res.appl.2013.7.3.4.

[6] I. P. B. Laboratorium Ekologi Hutan, Fakultas Kehutanan, *Diktat Dendrologi*. Bogor: Laboratorium Ekologi Hutan, Fakultas Kehutanan, Insitut Pertanian Bogor, 2004.

[7] L. I. Smith, *A tutorial on Principal Components Analysis*. John Wiley & Sons Inc, 2002.

[8] S. B, *Data Mining (Teori dan Aplikasi)*. Yogyakarta: Graha Ilmu, 2007.

[9] A. Ehsanirad, “Plant Classification Based on Leaf Recognition,” 2010.

[10] I. Laksmana *et al.*, “Identification of Tropical Plants Leaves Image Base on Principal Component Analysis,” *J. Appl. Agric. Sci. Technol.*, vol. 4, no. 1, pp. 64–75, Feb. 2020, doi: 10.32530/jaast.v4i1.156.

[11] P. J. Dunlop, C. M. Bignell, J. F. Jackson, and D. B. Hibbert, “Chemometric analysis of gas chromatographic data of oils from Eucalyptus species,” *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 59–67, Nov. 1995, doi: 10.1016/0169-7439(95)00036-4.

[12] K. Delac, M. Grgic, and T. Kos, “Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions.”

[13] V. V. Starovoitov, D. I. Samal, and D. V. Briliuk, “Image Enhancement for Face Recognition,” 2003.

[14] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[15] A. Rahmad, “Perbandingan Metode Ekstraksi Ciri FFT, PCA, dan FPE dalam Pengenalan Karakter Tulisan Tangan,” INSTITUT PERTANIAN BOGOR, 2012.

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

Gmail **Indra Laksmana <indra.puskom@gmail.com>**

Pre-review naskah seminar internasional ICGEFAST

2 pesan

---

icgefast icgefast <icgefast@gmail.com> 15 Desember 2021 pukul 14.16
Kepada: indra.puskom@gmail.com

Yth Bpk/Ibu Presenter ICGEFAST

Dalam rangka peningkatan kualitas naskah untuk mengurangi tingkat reject dari IOP maka kami lampirkan saran perbaikan dari tim ICGEFAST.

Mohon diperbaiki segera agar bisa dilanjutkan ke proses selanjutnya, terima kasih

---

artikelindraIOP (1).docx
71K

Indra Laksmana <indra.puskom@gmail.com> 19 Desember 2021 pukul 10.31
Kepada: icgefast icgefast <icgefast@gmail.com>

Terimakasih atas masukannya, jika masih ada kelihatan kurang pas. mohon jangan sungkan koreksiannya kembali..
[Kutipan teks disembunyikan]

---

artikel indral Laksmana .docx
68K

MACHINE LEARNING AS SEED IMAGE IDENTIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Indra Laksmana[1], Hendra[1], Jamaluddin[1], Trinovita Zuhara J[1], M Riza Nurtam[1], Amrizal[1], Rosda Syelly[2]

[1]Computer Engineering Study Program, Politeknik Pertanian Negeri Payakumbuh, Jalan Raya Negara Km 7, TanjungPati, West Sumatra, Indonesia 26271

[2]Program Studi Teknik Komputer, Sekolah Tinggi Teknologi Payakumbuh, Jalan Khatib Sulaiman Sawah Padang Payakumbuh, West Sumatra, Indonesia 26227

*Corresponden Email : indra.puskom@gmail.com

Abstract. We can identify plants using several variables, including seeds' shape, color, and size. But there are several types of plants having close similarities in seed shape, so one will need additional characteristics to support the identification process. This study applies machine learning with the principal component analysis (PCA) method to identify plant species from seed shapes. PCA simplifies the observed variables by reducing data dimensions and storing 75% of the information. The procedure will not eliminate too much important information while reducing data size and processing time. We collected 100 images of plant seeds that are similar to one another, such as sapodilla seeds, soursop, cucumber, star fruit, grape, melon, apple, lime, watermelon, and chili. A measurement system designed using K-Fold Cross Validation with 10 tables of experimental results yields a good level of accuracy of 83%. The Omission error occurred in the seeds of soursop, starfruit, grape, apple, lime, and watermelon, while the most commission errors occurred in apple seeds (8 times).

Keywords: *Seed Image Identification;Machine Learning; Principal Component Analysis*

Commented [G1]: Sesuaikan dengan penulisan templete

1. Introduction

Fruit plants are one of the plantation crops that are widely planted by farmers in Indonesia. The production of this plant is usually used both for self-sufficiency and to meet local market needs [1]. In general, fruit plants are planted around the house or in pots as garden decorations. [2]. The propagation of fruit plants that is still commonly carried out by farmers is by planting seeds. Some of the advantages of planting with seeds include a strong root system, easier to do, and a long period of fruiting, although the nature of the offspring is not the same as the parent [3].
The shape and structure of the seeds vary, depending on the type of species and environmental conditions in which the plant lives [4]. So that the seeds can be used as a marker in differentiating plant species [5]. But the diversity of forms is so much sometimes also difficult to identify a particular plant. In general, there are four ways to identify plants [6]. This will complicate the classification manually, this process requires a long time and special understanding. So we need a system with a technique to simplify the data in identifying plants. One of the techniques used to simplify the data used as an image classifier is to use Principal Component Analysis (PCA).

PCA is used to reduce the dimensions of data without significantly reducing the characteristics of the data [7] . According to Santosa [8], Principal Component Analysis is a reliable technique for extracting the structure of a data set with quite a lot of dimensions. PCA projects the image into its eigenspace plane by finding the eigenvectors of each image and projecting them into the obtained eigenspace. Research conducted by [9] in his journal "Plant classification based on leaf recognition" that uses the PCA method can classify plants with an accuracy rate of more than 90%. [10] have also identified medicinal plant leaf images with an accuracy of more than 90%. Dunlop et al [11] used PCA to identify typical compounds from various Eucalyptus plant species and using PCA obtained compound groupings that clearly described the specifications of each species.

This study applies Principal Component Analysis to identify plant species based on seed images. With Principal Component Analysis, the image of the seed can be simplified by reducing the dimensions and not losing a lot of important information contained in it.

2. Methodology

The stages of this research are starting from image data collection or data acquisition, then converting the image data into vectors, sharing data (training and testing), Principal Component Analysis process, and finally analyzing accuracy. These stages are shown in Figure 1

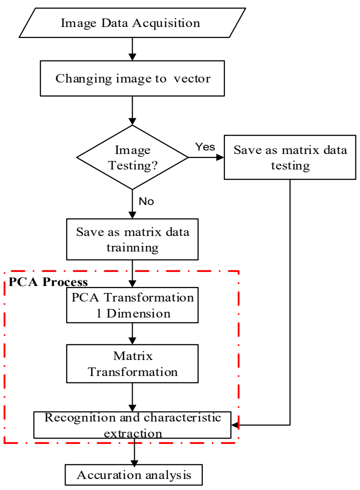

Figure 1. Research method

Image Data

In this study, 100 images were used, consisting of 10 types of plants, each of which used 10 images. Seed images were taken under the same lighting conditions and distances so as not to significantly affect the quality of the data and test results [12] [13]. Ten types of plants used have morphologically similar seeds, such as sapodilla (*Manilkara zapota*), soursop (*Annona muricata*), cucumber (*Cucumis sativus*), star fruit (*Averrhoa carambola*), grape (*Vitis vinifera*), melon (*Cucumis melo*), apple (*Malus domestica*), lime (*Citrus aurantifolia*), watermelon (*Manilkara zapota*) and chili (*Capsicum annum L.*). All images are saved in JPG format and RGB mode with image dimensions of 100x75 pixels.

Image reading and data split

From RGB mode, image data is converted to greyscale. The matrix of an image is saved as a 1x7500 row vector. This process is repeated for the entire image and the image data is combined into an *X* matrix and an *O* matrix. Matrix *X* is a collection of row vectors for training images, while matrix *O* is a collection of row vectors for testing images.

Image sharing using k-fold cross-validation with a value of k=10, also known as 10-fold cross-validation. The data is divided into two parts, namely, a training set of 9 images and a testing set of 1 image. So there are 9 x 10 = 90 images for training data. Meanwhile, as a testing image as much as 1x10 = 10 images. The first experiment is to use the first image as a testing image and so on until the tenth experiment is to use the tenth image as a testing image.

In one experiment, the process of reading the training image was repeated 9 times for 10 seed images, to obtain a combined matrix measuring 90x7500. The testing image in an experiment is an exception or other than the data from the *X* matrix because there is one image from each class (seed) which will be the training image or which is not processed when reading the *X* matrix, then there will be 10 images that will become row vectors. divide the matrix *O*, so that the matrix *O* is a composite matrix of testing images with a size of 10x7500.

Principal Component Analysis (PCA) Process

The PCA procedure aims to simplify the observed variables by reducing their dimensions [14]. A digital image consists of the smallest elements which are usually called pixels. The pixels store information in the form of the color intensity of the image at these coordinates. The image can be translated as a matrix while pixels can be translated as elements of the matrix [15].

In Artificial Intelligence, many use images as input which are processed in clustering and classification. The image or image is defined as a two-dimensional function $f(x,y)$ where x and y are the spatial coordinates and the amplitude f, at either x or y, is called the intensity or degree of gray. If x, y, and f are finite, the image can be said to be a digital image [16].

In this study, seed image detection is carried out in several stages, namely, the image reading process, determining the covariance matrix, determining the feature vector and feature roots that correspond to the covariance matrix, selecting the largest feature root so that it contributes the desired information and determining the transformation matrix using the characteristic vector according to the root of the feature, determine the resulting matrix from the transformation of the seed image, determine the average vector of the transformation results for each class (plant species), determine the Euclidean distance between the testing image and each class average vector, then image recognition based on the minimum Euclidean distance. If the minimum distance of the image is obtained from the same class average vector, then the image is recognized.

After the *X* matrix is obtained, the covariance calculation process is carried out, to obtain *sigma* matrix. The calculation of this covariance uses the equation $covariance(y_i) = \sigma_{y_i}^2 = a_i{}^T \Sigma a_i$. To perform covariance calculations, the previous unit type matrix data must be changed to double type so that the value range of the matrix elements changes from 0-255 to 0.0-1.

After obtaining the sigma covariance matrix, the next step is to determine the feature vector and the feature root of the covariance matrix. The resulting output is a matrix *v* and *d* each shows a feature vector and a feature root. It also determines the diagonal of the matrix *d*. The output of the diagonal matrix *d* is lambda, which is the characteristic root value of the *sigma* covariance matrix

Furthermore, after obtaining the values of the feature roots and feature vectors, the feature or feature extraction process is carried out from the image. Several characteristic roots are taken which are large enough to represent the desired information. In this study, the information taken is 75% of the initial information.

3. Result and Discussion

Looking at the results of the computer program that has been developed to run the PCA algorithm and 10-fold cross-validation, we get a matrix from the introduction of the 1st seed to the 10th seed of the training sample and detection sample. The results can be seen in Table 1 below

Table 1. Image Recognition Results for 10 Experiments

Experiment	Original image									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Detected as									
1	1	2	3	4	7	6	5	8	7	10
2	1	2	3	4	4	6	7	8	7	10
3	1	2	3	4	5	6	7	8	7	10
4	1	2	3	4	5	6	4	8	7	10
5	1	2	3	8	9	6	7	8	9	10
6	1	5	3	4	7	6	7	6	7	10
7	1	2	3	4	7	6	7	8	9	10
8	1	2	3	9	5	6	7	8	9	10
9	1	2	3	4	5	6	9	8	9	10
10	1	2	3	4	5	6	9	8	9	10

Commented [G2]: Sesuaikan dengan templete

In Table 1, the images of seeds 1 (sapodilla), 3 (cucumber), 6 (melon), and 10 (chili) in each experiment showed no errors in detecting the image. Meanwhile, image 5 (grape) was detected as an image of soursop seeds in the 6th experiment and was detected as an image of apple seeds in the first experiment. In the 5th experiment, the image of lime seeds was detected as an image of star fruit seeds. The image of star fruit seeds was detected as an image of grape seeds in the 2nd experiment and was detected as an apple in the 4th experiment. The image of watermelon seeds was detected as an image of star fruit seeds and apples. While the image of apple seeds was detected as the image of grape seeds and watermelon. The shape of the seed image and the pattern of image data that are taken from the same angle influences this detection error.

We iterated the experiment ten times to get information on how many times will the system accurately detect the plant seeds from its image. The sum of each element from the experimental results table can be used to show the accuracy of the image detection system. Table 2 is the sum of all experiments. The main diagonal shows the number of image detections correctly in 10 trials.

Table 2. The result of the sum of all trials

Label	**Detected as**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	10	0	0	0	0	0	0	0	0	0
2	0	9	0	0	1	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0	0
4	0	0	0	8	0	0	0	1	1	0
5	0	0	0	0	6	0	3	0	1	0
6	0	0	0	0	0	10	0	0	0	0
7	0	0	0	1	1	0	6	0	2	0
8	0	0	0	0	0	1	0	9	0	0
9	0	0	0	0	0	0	5	0	5	0
10	0	0	0	0	0	0	0	0	0	10

Commented [G3]: Sesuaikan dengan templete

Accuracy analysis

The calculation of accuracy for the seed image detection system can be found by counting the number of diagonals in the overall result table and divided by a lot of data and multiplied by 100%. The system that has been built has an accuracy of 83%.

Omission and Commission Error

Omission is the number of testing images that do not match the class classification category. Meanwhile, Commission is the number of detected testing images that are not in accordance with the actual situation/classification in the class. The omission error can be calculated by adding up every row of the result matrix, except the elements on the main diagonal. The output of the omission error is as shown in Table 3 below

Table 3. Omission error

Plants	Seed number	Omission Error
Sapodilla (*Manilkara zapota*)	1	0
soursop (*Annona muricata*)	2	1
cucumber (*Cucumis sativus*)	3	0
starfruit (*Averrhoa carambola*)	4	2
grape (*Vitis vinifera*)	5	4
Melon (*Cucumis melo*)	6	0
apple (*Malus domestica*)	7	4
lime (*Citrus aurantifolia*)	8	1
watermelon (*Citrullus lanatus*)	9	5
chili (*Capsicum annum L.*)	10	0

Commented [G4]: Sesuaikan Dengan Penulisa Templete

The most omission errors occurred in watermelon seeds, see Table 3. From 10 experiments conducted, watermelon seeds were detected 5 times as seeds of other plants. While the seeds of Sapodilla, cucumber, Melon, and chili did not occur omission, this means the system has successfully detected it correctly. The commission error can be calculated by adding up every column of the result matrix, except the elements on the main diagonal. The output of the commission error is shown in Table 4 below.

Meanwhile in Table 4, out of 10 experiments conducted by the system, the system carried out the most commission errors on apple seeds, meaning that the system detected other seeds as themselves or in other words, got detection errors from other plant seeds 8 times. Meanwhile, sapodilla, soursop, cucumber seeds, and chili seeds never received errors of detection from other seeds.

Table 4. Commission error

Nama Biji Tumbuhan	Biji ke	Omission Error
Sapodilla (*Manilkara zapota*)	1	0
soursop (*Annona muricata*)	2	0
cucumber (*Cucumis sativus*)	3	0
starfruit (*Averrhoa carambola*)	4	1
grape (*Vitis vinifera*)	5	2
Melon (*Cucumis melo*)	6	1
apple (*Malus domestica*)	7	8
lime (*Citrus aurantifolia*)	8	1
watermelon (*Citrullus lanatus*)	9	4
chili (*Capsicum annum L.*)	10	0

4. Conclusion

The results of this study concluded that the PCA technique can be used for dimension reduction. The information taken is 75% of the 100 processed plant seed images, obtaining good results, namely successfully recognizing the image correctly with a percentage rate of 83%. Image retrieval techniques are urgently needed so that the images taken as data have similar lighting, distances, and dimensions. identifying, assessing, and interpreting all available research evidence in order to provide answers to specific research questions.

References

[1] N. L. M. Pradnyawathi, I. K. Sardiana, and N. N. Darmiati, "Pengembangan Bibit Buah Lokal Unggul Wani Bali Tanpa Biji," *Bul. Udayana Mengabdi*, vol. 19, no. 1, Feb. 2020, doi: 10.24843/BUM.2020.v19.i01.p06.

[2] M. R. Suhartanto and E. Gunawan, *Untung besar dari bisnis bibit tanaman buah*. Jakarta: AgroMedia Pustaka, 2012.

[3] A. Barus and Syukri, *Agroteknologi Tanaman Buah-buahan*. Medan: USU Press, 2008.

[4] A. R. Muhammad, "Hubungan Bentuk Biji Dengan Karakteristik Morfologi Bibit Tanaman Pala (Myristica fragrans H.)," Andalas University, 2016.

[5] I. Laksmana, Y. Herdiyeni, and E. A. M. Zuhud, "Genetic Programming for Medicinal Plant Family Identification System," *J. ICT Res. Appl.*, vol. 7, no. 3, pp. 217–234, Dec. 2013, doi: 10.5614/itbj.ict.res.appl.2013.7.3.4.

[6] I. P. B. Laboratorium Ekologi Hutan, Fakultas Kehutanan, *Diktat Dendrologi*. Bogor: Laboratorium Ekologi Hutan, Fakultas Kehutanan, Insitut Pertanian Bogor, 2004.

[7] L. I. Smith, *A tutorial on Principal Components Analysis*. John Wiley & Sons Inc, 2002.

[8] S. B, *Data Mining (Teori dan Aplikasi)*. Yogyakarta: Graha Ilmu, 2007.

[9] A. Ehsanirad, "Plant Classification Based on Leaf Recognition," 2010.

[10] I. Laksmana *et al.*, "Identification of Tropical Plants Leaves Image Base on Principal Component Analysis," *J. Appl. Agric. Sci. Technol.*, vol. 4, no. 1, pp. 64–75, Feb. 2020, doi: 10.32530/jaast.v4i1.156.

[11] P. J. Dunlop, C. M. Bignell, J. F. Jackson, and D. B. Hibbert, "Chemometric analysis of gas chromatographic data of oils from Eucalyptus species," *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 59–67, Nov. 1995, doi: 10.1016/0169-7439(95)00036-4.

[12] K. Delac, M. Grgic, and T. Kos, “Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions.”

[13] V. V. Starovoitov, D. I. Samal, and D. V. Briliuk, “Image Enhancement for Face Recognition,” 2003.

[14] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[15] A. Rahmad, “Perbandingan Metode Ekstraksi Ciri FFT, PCA, dan FPE dalam Pengenalan Karakter Tulisan Tangan,” INSTITUT PERTANIAN BOGOR, 2012.

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

MACHINE LEARNING AS SEED IMAGE IDENTIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Indra Laksmana[1], Hendra[1], Jamaluddin[1], Trinovita Zuhara J[1], M Riza Nurtam[1], Amrizal[1], Rosda Syelly[2]

[1]Computer Engineering Study Program, Politeknik Pertanian Negeri Payakumbuh, Jalan Raya Negara Km 7, TanjungPati, West Sumatra, Indonesia 26271

[2]Program Studi Teknik Komputer, Sekolah Tinggi Teknologi Payakumbuh, Jalan Khatib Sulaiman Sawah Padang Payakumbuh, West Sumatra, Indonesia 26227

*Corresponden Email : indra.puskom@gmail.com

Abstract. We can identify plants using several variables, including seeds' shape, color, and size. But there are several types of plants having close similarities in seed shape, so one will need additional characteristics to support the identification process. This study applies machine learning with the PCA method to identify plant species from seed shapes. PCA simplifies the observed variables by reducing data dimensions and storing 75% of the information. The procedure will not eliminate too much important information while reducing data size and processing time. We collected 100 images of plant seeds that are similar to one another, such as sapodilla seeds, soursop, cucumber, star fruit, grape, melon, apple, lime, watermelon, and chili. A measurement system designed using K-Fold Cross Validation with 10 tables of experimental results yields a good level of accuracy of 83%. The Omission error occurred in the seeds of soursop, starfruit, grape, apple, lime, and watermelon, while the most commission errors occurred in apple seeds (8 times).

Keywords: Seed Image Identification; Machine Learning; Principal Component Analysis

1. Introduction

Fruit plants are one of the plantation crops that are widely planted by farmers in Indonesia. The production of this plant is usually used both for self-sufficiency and to meet local market needs [1]. In general, fruit plants are planted around the house or in pots as garden decorations. [2]. The propagation of fruit plants that is still commonly carried out by farmers is by planting seeds. Some of the advantages of planting with seeds include a strong root system, easier to do, and a long period of fruiting, although the nature of the offspring is not the same as the parent[3].

The shape and structure of the seeds vary, depending on the type of species and environmental conditions in which the plant lives [4]. So that the seeds can be used as a marker in differentiating plant species [5]. But the diversity of forms is so much sometimes also difficult to identify a particular plant. In general, there are four ways to identify plants [6]. This will complicate the classification manually, this process requires a long time and special understanding. So we need a system with a technique to simplify the data in identifying plants. One of the techniques used to simplify the data used as an image classifier is to use Principal Component Analysis (PCA).

PCA is used to reduce the dimensions of data without significantly reducing the characteristics of the data [7] . According to Santosa [8], Principal Component Analysis is a reliable technique for extracting the structure of a data set with quite a lot of dimensions. PCA projects the image into its eigenspace plane by finding the eigenvectors of each image and projecting them into the obtained eigenspace. Research conducted by [9] in his journal "Plant classification based on leaf recognition" that uses the PCA method can classify plants with an accuracy rate of more than 90%. [10] have also identified medicinal plant leaf images with an accuracy of more than 90%. Dunlop et al [11] used PCA to identify typical compounds from various Eucalyptus plant species and using PCA obtained compound groupings that clearly described the specifications of each species.

This study applies Principal Component Analysis to identify plant species based on seed images. With Principal Component Analysis, the image of the seed can be simplified by reducing the dimensions and not losing a lot of important information contained in it.

2. Methodology

The stages of this research are starting from image data collection or data acquisition, then converting the image data into vectors, sharing data (training and testing), Principal Component Analysis process, and finally analyzing accuracy. These stages are shown in Figure1

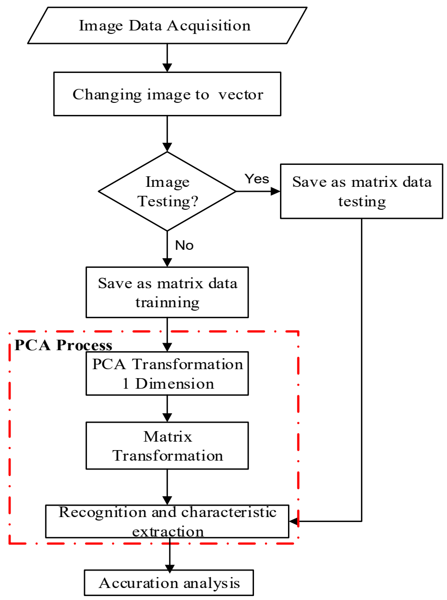

Figure 1. Research method

Image Data

In this study, 100 images were used, consisting of 10 types of plants, each of which used 10 images. Seed images were taken under the same lighting conditions and distances so as not to significantly affect the quality of the data and test results [12] [13]. Ten types of plants used have morphologically similar seeds, such as sapodilla (*Manilkara zapota*), soursop (*Annona muricata*), cucumber (*Cucumis sativus*), star fruit (*Averrhoa carambola*), grape (*Vitis vinifera*), melon (*Cucumis melo*), apple (*Malus domestica*), lime (*Citrus aurantifolia*), watermelon (*Manilkara zapota*) and chili (*Capsicum annum L.*). All images are saved in JPG format and RGB mode with image dimensions of 100x75 pixels.

Image reading and data split

From RGB mode, image data is converted to greyscale. The matrix of an image is saved as a 1x7500 row vector. This process is repeated for the entire image and the image data is combined into an *X* matrix and an *O* matrix. Matrix *X* is a collection of row vectors for training images, while matrix *O* is a collection of row vectors for testing images.

Image sharing using k-fold cross-validation with a value of k=10, also known as 10-fold cross-validation. The data is divided into two parts, namely, a training set of 9 images and a testing set of 1 image. So there are 9 x 10 = 90 images for training data. Meanwhile, as a testing image as much as 1x10 = 10 images. The first experiment is to use the first image as a testing image and so on until the tenth experiment is to use the tenth image as a testing image.

In one experiment, the process of reading the training image was repeated 9 times for 10 seed images, to obtain a combined matrix measuring 90x7500. The testing image in an experiment is an exception or other than the data from the *X* matrix because there is one image from each class (seed) which will be the training image or which is not processed when reading the *X* matrix, then there will be 10 images that will become row vectors. divide the matrix *O*, so that the matrix *O* is a composite matrix of testing images with a size of 10x7500.

Principal Component Analysis (PCA) Process

The PCA procedure aims to simplify the observed variables by reducing their dimensions [14]. A digital image consists of the smallest elements which are usually called pixels. The pixels store information in the form of the color intensity of the image at these coordinates. The image can be translated as a matrix while pixels can be translated as elements of the matrix [15]. In Artificial Intelligence, many use images as input which are processed in clustering and classification. The image or image is defined as a two-dimensional function *f(x,y)* where *x* and *y* are the spatial coordinates and the amplitude *f*, at either *x* or *y*, is called the intensity or degree of gray. If *x*, *y*, and *f* are finite, the image can be said to be a digital image [16].

In this study, seed image detection is carried out in several stages, namely, the image reading process, determining the covariance matrix, determining the feature vector and feature roots that correspond to the covariance matrix, selecting the largest feature root so that it contributes the desired information and determining the transformation matrix using the characteristic vector according to the root of the feature, determine the resulting matrix from the transformation of the seed image, determine the average vector of the transformation results for each class (plant species), determine the Euclidean distance between the testing image and each class average vector, then image recognition based on the minimum Euclidean distance. If the minimum distance of the image is obtained from the same class average vector, then the image is recognized.

After the *X* matrix is obtained, the covariance calculation process is carried out, to obtain *sigma* matrix. The calculation of this covariance uses the equation $covariance(y_i) = \sigma_{y_i}^2 = {a_i}^T \Sigma a_i$.

To perform covariance calculations, the previous unit type matrix data must be changed to double type so that the value range of the matrix elements changes from 0-255 to 0.0-1.

After obtaining the sigma covariance matrix, the next step is to determine the feature vector and the feature root of the covariance matrix. The resulting output is a matrix *v* and *d* each shows a feature vector and a feature root. It also determines the diagonal of the matrix *d*. The output of the diagonal matrix *d* is lambda, which is the characteristic root value of the *sigma* covariance matrix

Furthermore, after obtaining the values of the feature roots and feature vectors, the feature or feature extraction process is carried out from the image. Several characteristic roots are taken which are large enough to represent the desired information. In this study, the information taken is 75% of the initial information.

3. Result and Discussion

Looking at the results of the computer program that has been developed to run the PCA algorithm and 10-fold cross-validation, we get a matrix from the introduction of the 1st seed to the 10th seed of the training sample and detection sample. The results can be seen in Table 1 below

Table 1. Image Recognition Results for 10 Experiments

Experiment	**Original image**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Detected as									
1	1	2	3	4	7	6	5	8	7	10
2	1	2	3	4	4	6	7	8	7	10
3	1	2	3	4	5	6	7	8	7	10
4	1	2	3	4	5	6	4	8	7	10
5	1	2	3	8	9	6	7	8	9	10
6	1	5	3	4	7	6	7	6	7	10
7	1	2	3	4	7	6	7	8	9	10
8	1	2	3	9	5	6	7	8	9	10
9	1	2	3	4	5	6	9	8	9	10
10	1	2	3	4	5	6	9	8	9	10

In Table 1, the images of seeds 1 (sapodilla), 3 (cucumber), 6 (melon), and 10 (chili) in each experiment showed no errors in detecting the image. Meanwhile, image 5 (grape) was detected as an image of soursop seeds in the 6th experiment and was detected as an image of apple seeds in the first experiment. In the 5th experiment, the image of lime seeds was detected as an image of star fruit seeds. The image of star fruit seeds was detected as an image of grape seeds in the 2nd experiment and was detected as an apple in the 4th experiment. The image of watermelon seeds was detected as an image of star fruit seeds and apples. While the image of apple seeds was detected as the image of grape seeds and watermelon. The shape of the seed image and the pattern of image data that are taken from the same angle influences this detection error.

We iterated the experiment ten times to get information on how many times will the system accurately detect the plant seeds from its image. The sum of each element from the experimental results table can be used to show the accuracy of the image detection system. Table 2 is the sum of all experiments. The main diagonal shows the number of image detections correctly in 10 trials.

Table 2. The result of the sum of all trials

Label	**Detected as**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	10	0	0	0	0	0	0	0	0	0
2	0	9	0	0	1	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0	0
4	0	0	0	8	0	0	0	1	1	0
5	0	0	0	0	6	0	3	0	1	0
6	0	0	0	0	0	10	0	0	0	0
7	0	0	0	1	1	0	6	0	2	0
8	0	0	0	0	0	1	0	9	0	0
9	0	0	0	0	0	0	5	0	5	0
10	0	0	0	0	0	0	0	0	0	10

Accuracy analysis

The calculation of accuracy for the seed image detection system can be found by counting the number of diagonals in the overall result table and divided by a lot of data and multiplied by 100%. The system that has been built has an accuracy of 83%.

Omission and Commission Error

Omission is the number of testing images that do not match the class classification category. Meanwhile, Commission is the number of detected testing images that are not in accordance with the actual situation/classification in the class. The omission error can be calculated by adding up every row of the result matrix, except the elements on the main diagonal. The output of the omission error is as shown in Table 3 below

Table 3. Omission error

Plants	Seed number	Omission Error
Sapodilla (*Manilkara zapota*)	1	0
soursop (*Annona muricata*)	2	1
cucumber (*Cucumis sativus*)	3	0
starfruit (*Averrhoa carambola*)	4	2
grape (*Vitis vinifera*)	5	4
Melon (*Cucumis melo*)	6	0
apple (*Malus domestica*)	7	4
lime (*Citrus aurantifolia*)	8	1
watermelon (*Citrullus lanatus*)	9	5
chili (*Capsicum annum L.*)	10	0

The most omission errors occurred in watermelon seeds, see Table 3. From 10 experiments conducted, watermelon seeds were detected 5 times as seeds of other plants. While the seeds of Sapodilla, cucumber, Melon, and chili did not occur omission, this means the system has successfully detected it correctly. The commission error can be calculated by adding up every column of the result matrix, except the elements on the main diagonal. The output of the commission error is shown in Table 4 below.

Meanwhile in Table 4, out of 10 experiments conducted by the system, the system carried out the most commission errors on apple seeds, meaning that the system detected other seeds as themselves or in other words, got detection errors from other plant seeds 8 times. Meanwhile, sapodilla, soursop, cucumber seeds, and chili seeds never received errors of detection from other seeds.

Table 4. Commission error

Plants	Seed number	Comission Error
Sapodilla (*Manilkara zapota*)	1	0
soursop (*Annona muricata*)	2	0
cucumber (*Cucumis sativus*)	3	0
starfruit (*Averrhoa carambola*)	4	1
grape (*Vitis vinifera*)	5	2
Melon (*Cucumis melo*)	6	1
apple (*Malus domestica*)	7	8
lime (*Citrus aurantifolia*)	8	1
watermelon (*Citrullus lanatus*)	9	4
chili (*Capsicum annum L.*)	10	0

4. Conclusion

The results of this study concluded that the PCA technique can be used for dimension reduction. The information taken is 75% of the 100 processed plant seed images, obtaining good results, namely successfully recognizing the image correctly with a percentage rate of 83%. Image retrieval techniques are urgently needed so that the images taken as data have similar lighting, distances, and dimensions. identifying, assessing, and interpreting all available research evidence in order to provide answers to specific research questions.

References

[1] N. L. M. Pradnyawathi, I. K. Sardiana, and N. N. Darmiati, "Pengembangan Bibit Buah Lokal Unggul Wani Bali Tanpa Biji," *Bul. Udayana Mengabdi*, vol. 19, no. 1, Feb. 2020, doi: 10.24843/BUM.2020.v19.i01.p06.

[2] M. R. Suhartanto and E. Gunawan, *Untung besar dari bisnis bibit tanaman buah*. Jakarta: AgroMedia Pustaka, 2012.

[3] A. Barus and Syukri, *Agroteknologi Tanaman Buah-buahan*. Medan: USU Press, 2008.

[4] A. R. Muhammad, "Hubungan Bentuk Biji Dengan Karakteristik Morfologi Bibit Tanaman Pala (Myristica fragrans H.)," Andalas University, 2016.

[5] I. Laksmana, Y. Herdiyeni, and E. A. M. Zuhud, "Genetic Programming for Medicinal Plant Family Identification System," *J. ICT Res. Appl.*, vol. 7, no. 3, pp. 217–234, Dec. 2013, doi: 10.5614/itbj.ict.res.appl.2013.7.3.4.

[6] I. P. B. Laboratorium Ekologi Hutan, Fakultas Kehutanan, *Diktat Dendrologi*. Bogor: Laboratorium Ekologi Hutan, Fakultas Kehutanan, Insitut Pertanian Bogor, 2004.

[7] L. I. Smith, *A tutorial on Principal Components Analysis*. John Wiley & Sons Inc, 2002.

[8] S. B, *Data Mining (Teori dan Aplikasi)*. Yogyakarta: Graha Ilmu, 2007.

[9] A. Ehsanirad, "Plant Classification Based on Leaf Recognition," 2010.

[10] I. Laksmana *et al.*, "Identification of Tropical Plants Leaves Image Base on Principal Component Analysis," *J. Appl. Agric. Sci. Technol.*, vol. 4, no. 1, pp. 64–75, Feb. 2020, doi: 10.32530/jaast.v4i1.156.

[11] P. J. Dunlop, C. M. Bignell, J. F. Jackson, and D. B. Hibbert, "Chemometric analysis of gas chromatographic data of oils from Eucalyptus species," *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 59–67, Nov. 1995, doi: 10.1016/0169-7439(95)00036-4.

[12] K. Delac, M. Grgic, and T. Kos, "Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions."

[13] V. V. Starovoitov, D. I. Samal, and D. V. Briliuk, “Image Enhancement for Face Recognition,” 2003.

[14] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[15] A. Rahmad, “Perbandingan Metode Ekstraksi Ciri FFT, PCA, dan FPE dalam Pengenalan Karakter Tulisan Tangan,” INSTITUT PERTANIAN BOGOR, 2012.

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

Gmail

Indra Laksmana <indra.puskom@gmail.com>

Proceedings of 1st Lekantara Annual Conference on Natural Science and Environment (LeNS 2021) now available online

1 pesan

IOP Conference Series team <jpcs@ioppublishing.org> 31 Oktober 2022 pukul 18.04
Balas Ke: IOP Conference Series team <jpcs@ioppublishing.org>
Kepada: indra.puskom@gmail.com

Your article Machine Learning As Seed Image Identification Using Principal Component Analysis (PCA) is online.
Visit iopscience.org/jpcs | View this email online I Unsubscribe

IOP Conference Series

Proceedings of 1st Lekantara Annual Conference on Natural Science and Environment (LeNS 2021).

Thank you for publishing your paper 'Machine Learning As Seed Image Identification Using Principal Component Analysis (PCA)' in the IOP Conference Series: Earth and Environmental Science™. Your article has now been published online.

Create an account in ScholarOne
As part of our commitment to provide the best possible publishing service to our authors, we encourage you to create an account in ScholarOne, so you can benefit from the following:

- Invitations to write and referee papers in your research area.
- Stay up-to-date on resources available to help you with getting your work published and promoted.
- Associate your existing ORCID ID with your account or create a free ORCID iD.

Create a ScholarOne Account →

Thank you, and we hope to work with you again soon.

Anete Ashton
Publisher, Conference Series
Journal of Physics: Conference Series
jpcs@ioppublishing.org

IOP Publishing