ILKOM Jurnal Ilmiah Vol. 14, No. 2, August 2022, pp. 134-141
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Comparative analysis to determine the best accuracy of classification methods

Warnia Nengsih [a,1,*]; Yuli Fitrisia [a,2]; Mardhiah Fadhli [a,3]
[a] *Computer of Departement Politeknik Caltex, Jl. Umban Sari No.13, Pekanbaru and 28265, Indonesia*
[1] *warnia@pcr.ac.id;* [2] *uli@pcr.ac.id;* [3] *mardhiah@pcr.ac.id*
** Corresponding author*

Article history: Received April 06, 2022; Revised April 07, 2022; Accepted August 02, 2022; Available online August 30, 2022

Abstract

The classification method is one of the methods of supervised learning and predictive learning. This method can be used to detect an object in the image presented, whether it is in accordance with the existing object in the training phase. There are several classification methods used, including Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) and Decision Tree. To determine the accuracy in detecting these objects, it is necessary to measure the accuracy of each used classification method. The object that became simulation in this research was the object image of Guava and Pear fruit. Testing using confusion matrix. The results showed that the Support Vector Machine (SVM) method was able to detect with an accuracy of 98.09%. Then the K-Nearest Neighbors (K-NN) method with an accuracy of 98.06%, then the Decision Tree method with an accuracy of 97.57%. From the results of the accuracy test, it can be concluded that basically these three classification methods have high accuracy with a difference of 0.49% and the overall average accuracy of the classification of the three methods is 97.89%.

Keywords: SVM; K-NN; Decision Tree; Classification

Introduction

Classification is a data grouping where the data used has a label or target class. So that algorithms for solving classification problems are categorized into supervised learning. The purpose of supervised learning is that the label or target data plays a role as a 'supervisor' that oversees the learning process in achieving a certain level of accuracy or precision. To detect an image, accuracy in detecting an object is very important. Some of the methods used in the classification and are used in this study include Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) and Decision Tree. Support Vector Machine (SVM). The method used by SVM works with the principle of Structural Risk Minimization which aims to find the hyperplane 2 class in the input space. The K-Nearest Neighbors algorithm is a supervised learning algorithm where the results of new instances are classified based on most of the K-Nearest Neighbors category. The purpose of this algorithm is to classify new objects based on attributes and samples of training data. The K-Nearest Neighbors algorithm uses the Neighborhood Classification as a predictive value of the new instance value 1. The purpose of this algorithm is to classify new objects based on attributes and samples of training data. The K-Nearest Neighbor algorithm uses the Neighborhood Classification as a predictive value of the new instance value. The purpose of this algorithm is to classify new objects based on attributes and samples of training data. Trials of different k values, to produce different levels of accuracy in the classification process carried out. Decision tree is a prediction model using a tree structure or hierarchical structure. The concept of a decision tree is to convert data into a decision tree and decision rules. The main benefit of using a decision tree is its ability to break down complex decision-making processes into simpler ones, so that decision makers will better interpret solutions to problems. The Decision Tree method is used to calculate predictive accuracy. The research uses the normalized Confusion Matrix plot to calculate the metric score.

In previous studies, only one test method was used to measure the accuracy of each used classification method. However, the fundamental difference in this study is the comparison of several evaluation methods used to measure the accuracy and error rate of each of these methods.

http://dx.doi.org/10.33096/ilkom.v14i2.1128.134-141

Method

The dataset used was an image data set for guava and pear fruits. The dataset used was a public dataset. Here is a link from the https://www.kaggle.com/datasets/omkarmanohardalvi/guava-disease-dataset-4-types dataset. The content of variable fruits with Guava and Pear was the data to be processed. Contains training data and testing data that were used to retrieve images and labels for classification The getYourFruits method functions to retrieve data from the dataset and start with the declaration of empty images and labels then looping so that it displayed patterns and data and it can perform classification.

Figure 1 are the mechanisms and stages of the research carried out

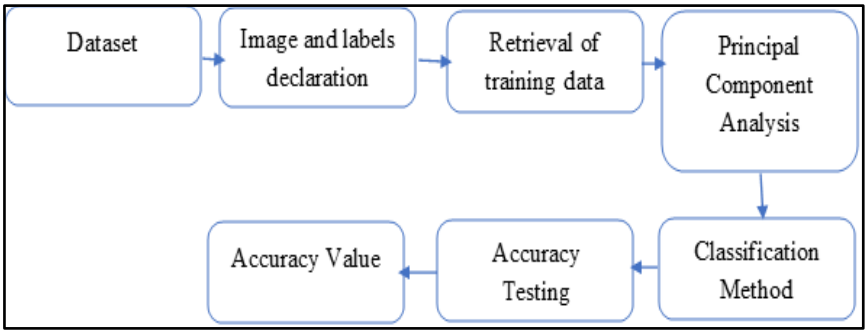

Figure 1. mechanisms and stages of the research

The dataset was dataset of guava and pear fruit images, then determining and declaring the labels. Before determining the training data, implementing PCA and applying the classification algorithm were used. In this study, the classification algorithms used were SVM, K-NN and Decision Tree see in **Figure 2**.

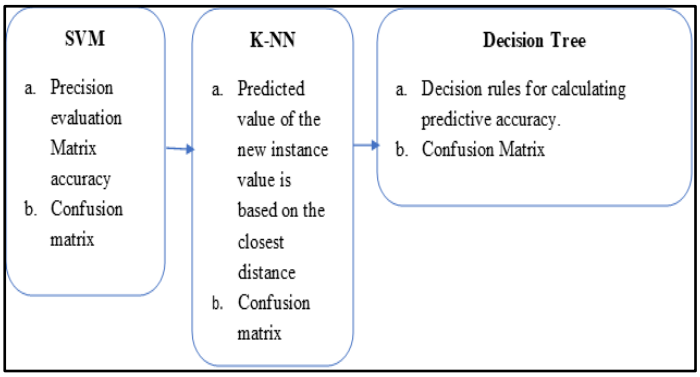

Figure 2. SVM, K-NN and Decision Tree

For the SVM algorithm, there were related concepts including Precision evaluation matrix accuracy and Confusion matrix, K-NN algorithm related concepts, including Predicted value of the new instance value was based on the closest distance, while for Decision Tree, related concepts included Decision rules for calculating predictive, accuracy and Confusion Matrix

Discussion and Results

This research used the plot_image_grid method using the parameters images, nb_rows, nb_cols, figrize. The plot_image_grid method functioned to display images in the form of a grid table with a certain size for rows x columns. The results can be seen that there were 490 training data in the form of Guava images, 492 training data in the form of Pear images, 104 testing data in the form of Guava images. 102 testing data in the form of Pear images. To display Guava fruits that were at index 0 to 100, the plot_image_grid method was used. The image was then displayed as 10 columns by 10 rows.

Guava results were taken from the dataset see in **Figure 3**.

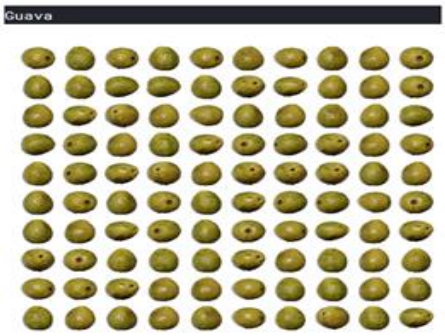

Figure 3. The 10x10 Guava dataset, 102 testing data in the form of Pear images

Displaying pears that were at index 490 to 590, we use the plot_image_grid method. The image was displayed as 10 columns by 10 rows. **Figure 4** are Pear results were taken from the dataset.

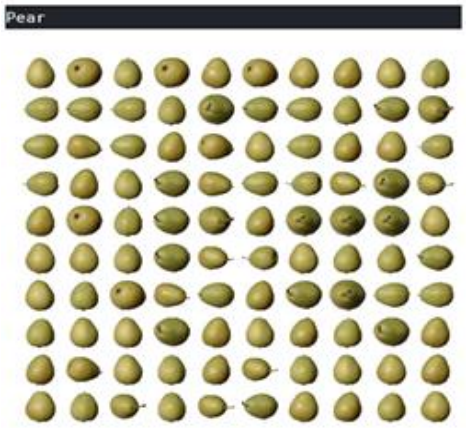

Figure 4. The 10x10 Guava dataset, displaying pears that are at index 490 to 590,

Displaying pears that were at index 490 to 590, the plot_image_grid method was employed. The image would then be displayed as 10 columns by 10 rows. **Figure 5** are Pear results were taken from the dataset.

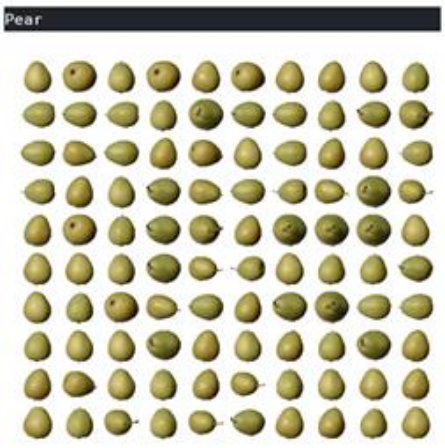

Figure 5. 10x10 Pear dataset, displaying that are at index 490 to 590, we use the plot_image_grid method

Principal Component Analysis (PCA) is a technique used to simplify data by transforming data linearly so that a new coordinate system with maximum variance is formed. GetClassNumber method which functions to retrieve numbering in the previous class.

The PCA algorithm method was used to display images in 2 dimensions. Initialize the PCA class by passing several components to the constructor. The transformation method returns the specified number of principal components see in **Figure 6**.

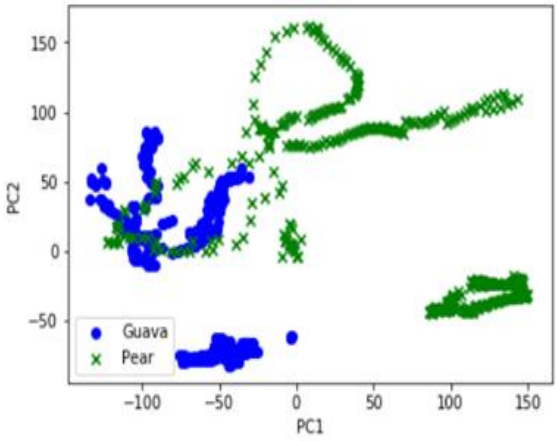

Figure 6. PCA transformation, display images in 2 dimensions

Principal Component Analysis (PCA) algorithm displays images in 3-dimensional form see in **Figure 7**.

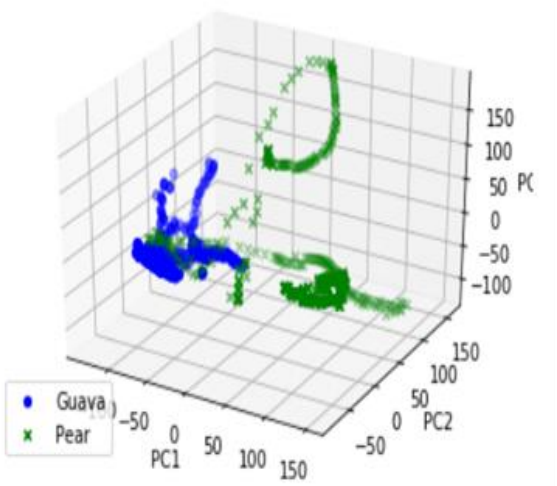

Figure 7. dimensional PCA transformation, guava=blue, pear=green

Based on **Figure 7** the showPCA method declaration was utalised to display part of the image in the Principal Component Analysis. The images shown were original images, setting 50 PCs, 10 PCs, and 2PCs. While the computePCA method was used to calculate PCA results. The PCA used were 2.10, and 50 for the training process and re-shaping (change in image size). Displays the original guava image and the results after the fruit image was changed.

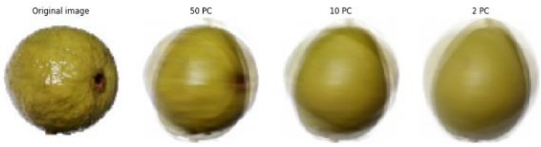

Figure 8. The training process, and re-shapping,

Support Vector Machine (SVM) has the basic principle of linear and non-linear classification by adding the kernel concept to a high-dimensional workspace. Linear SVM is the separation of data according to its linearity. The best separator can separate not only data, but also margins. The method used in SVM worked with the principle of Structural Risk Minimization which aimed to find the best hyperplane that fitted 2 classes in the input space.

To determine the image of guava or pear (classification), the Support Vector Machine Algorithm was used. The accuracy of the matrix accuracy where the Support Vector Machine accuracy was obtained from the calculated Confusion matrix plot was also examined. The results of the Confusion Matrix from the X line (fact) and the Y line (prediction) were the prediction results using the Support Vector Machine with an accuracy of 98.06% see in **Figure 9**.

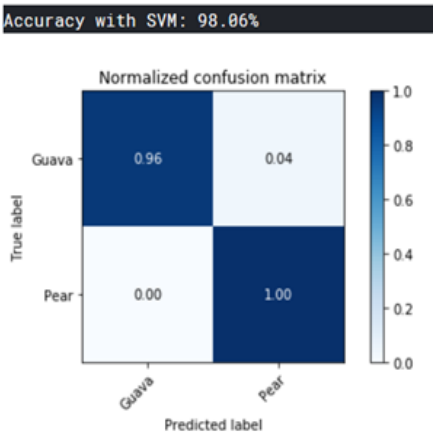

Figure 9. Accuracy using SVM

The K-Nearest Neighbors algorithm is a supervised learning algorithm where the results of new instances are classified based on most of the closest values. The purpose of this algorithm is to classify new objects based on attributes and samples of training data. The result of the accuracy of KNN was 98.06%. The results of confusion matrix, namely the X line (fact) and the Y line (prediction) were 0.96 and 1.00 respectively indicating the correct image prediction see in **Figure 10**.

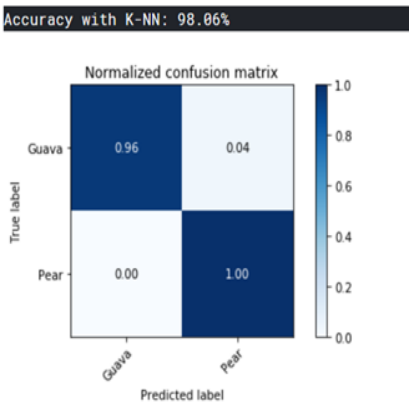

Figure 10. Accuracy using K-NN

The decision tree is one of the most popular classification methods, because it is easy to interpret. Decision tree is a prediction model using a tree structure or hierarchical structure. The concept of a decision tree is to convert data into a decision tree and decision rules. The main benefit of using a decision tree is its ability to break down complex decision-making processes into simpler ones, so that decision makers will better interpret solutions to problems. The normalized Confusion Matrix plot was used to calculate the metric score. The results of the Confusion Matrix, namely the X line (fact) and the Y line (prediction) were 0.95% and 0.97% respectively, with an accuracy of l value of 96.12% see in **Figure 9**.

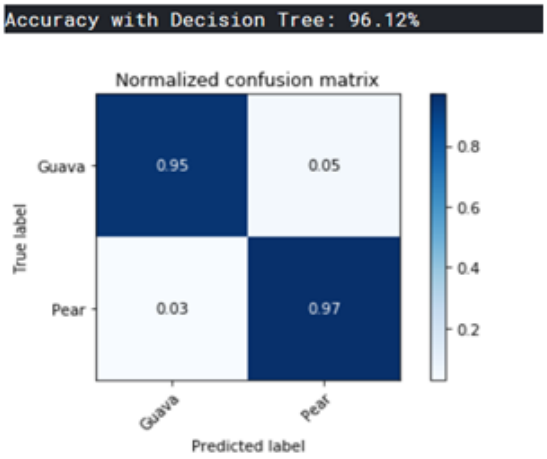

Figure 11. Accuracy using the Decision Tree

Based on **Figure 11** is Accuracy using the Decision Tree. Measuring the accuracy of training data and testing data can be seen in the following PCA visualization. The results of the Decision Tree calculation seem to decrease at the 2nd depth and the 7th depth. The Decision Tree looks up at the 3rd depth and the 8th depth.

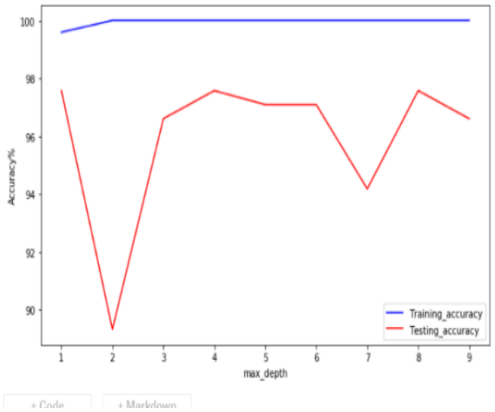

Figure 12. The results of the Decision Tree calculation

Based on **Figure 12** are The results of the Decision Tree calculation. Decision Tree + PCA functions for tree prediction for calculating the accuracy matrix score. Its accuracy was 98.00% with the prediction results of Guava and Pear having a close scope.

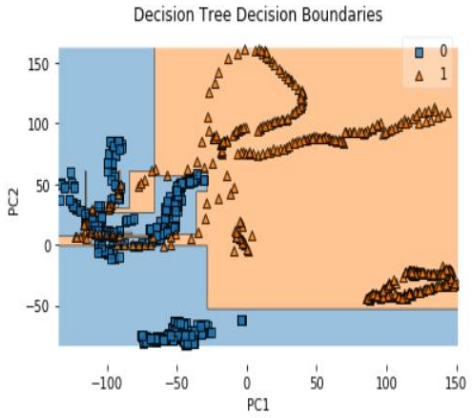

Figure 13. Decision Tree + PCA functions for tree prediction for calculating the accuracy matrix score

Based on **Figure 13** is decision tree + PCA functions for tree prediction for calculating the accuracy matrix score. From the comparison of the three methods, the predictive value for each method is obtained as follows:

1. The prediction result of SVM was 96.2%.
2. The prediction result of the KNN was 98.1%.
3. The prediction result of Decision Tree was 96.1%.

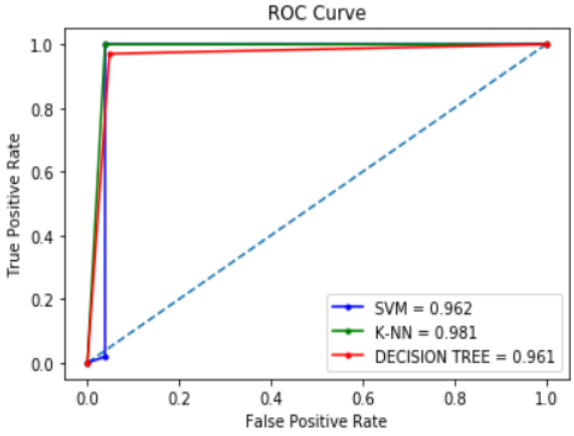

Figure 14. The comparison of the three methods

The results showed that the Support Vector Machine (SVM) method was able to detect with an accuracy of 98.06%, then the K-Nearest Neighbors (K-NN) method with an accuracy of 98.06%, and the Decision Tree method with an accuracy of 97.57% see in **Figure 14**. From the results of the accuracy test it can be concluded that basically these three classification methods have high accuracy with a difference of 0.49% and the overall average accuracy of the classification of the three methods was 97.89%. Cross validation can also be applied to measurement accuracy so that it can be seen the comparison of several test methods used. Cross validation is an additional method that aims to obtain maximum accuracy results, where the experiment is k times for one model with the same parameters.

Conclusion

Receiver Operating Characteristics (ROC) is a performance measurement tool by determining a threshold value of a model. Theoretically, the higher the ROC value, the better the model will be. The value is low if the curve approaches the baseline line that crosses the ordinate point 0.0 whereas the value is good if the curve is close to point 0.1. Accuracy measurement using the confusion matrix in the Support Vector Machine (SVM) method was used to detect pear and guava objects with an accuracy of 98.06%. While using Receiver Operating Characteristics obtained an accuracy of 96.2%. The gap difference was 1.86% between the two methods. Then the measurement of accuracy using the confusion matrix on the K-Nearest Neighbors (K-NN) method obtained an accuracy of 98.06%, while using the Receiver Operating Characteristics obtained an accuracy of 98.1%. This indicated the difference of the two methods was 0.04%. Furthermore, the Decision Tree method obtained an accuracy of 97.57%. While using Receiver Operating Characteristics obtained an accuracy of 96.1%. The result indicated that gap difference of the two methods was 1.47%. Referring to the results of the accuracy test, it can be concluded that basically the three classification methods show good accuracy results with a difference of 0.49%, where the average the overall classification accuracy of the three methods is 97.89%.

References

[1] Gou, J., Du, L.Zhang, Y. & Xiong, T. “A new distance-weighted k-nearest Neighbor Classifier”, Journal of Information & Computational Science, 9 (6): 1429-1436. 2012

[2] Domeniconi, C., Peng, J. & Gunopulos, D. "Locally adaptive metric nearest-oriented classification", IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (9): 1281–1285. 2002

[3] Utku A, Hacer (Uke) Karacan, Yildiz O, Akcayol MA. Implementation of a New Recommendation System Based on Decision Tree Using Implicit Relevance Feedback. JSW. 2015 Dec 1; 10 (12): 1367-74.

[4] Jadhav SD, Channe HP. Efficient recommendation system using decision tree classifier and collaborative filtering. Int. Res. J. Eng. Technol. 2016; 3: 2113-8.

[5] Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security. 2009 Jun; 3 (3): 230-40.

[6] Gershman A, Meisels A, Lüke KH, Rokach L, Schclar A, Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).

[7] Pandey M, Sharma VK. A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications. 2013 Jan 1; 61 (13).

[8] Priyama A, Abhijeeta RG, Ratheeb A, Srivastavab S. Comparative analysis of decision tree classification algorithms. International Journal of Current Engineering and Technology. 2013 Jun; 3 (2): 334-7.

[9] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to Support Vector Classification". Deptt of Computer Sci. National Taiwan Union, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007.

[10] Banu GR. A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease. International Journal of Computer Sciences and Engineering. 2016; 4 (11): 111-5.

[11] C.-W. Hsu and CJ Lin. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13 (2): 415-425, 2002.

[12] Baoli, L., Shiwen, Y. & Qin, L. "An Improved k-Nearest Neighbor Algorithm for Text Categorization, ArXiv Computer Science e-prints.2003

[13] Bax, E. "Validation of nearest neighbor classifiers", IEEE Trans. Inform. Theory, 46: 2746–2752.2000

[14] Li Maokuan, Cheng Yusheng, Zhao Honghai "Unlabeled data classification via SVM and k-means Clustering". Proceeding of the International Conference on Computer Graphics, Image and Visualization (CGIV04), 2004 IEEE.

[15] Chitra, A. & Uma, S. "An Ensemble Model of Multiple Classifiers for Time Series Prediction", International Journal of Computer Theory and Engineering, 2 (3): 1793-8201.2010

[16] Gil-Garcia, R. & Pons-Porrata, A. "A New Nearest Neighbor Rule for Text Categorization", Lecture Notes in Computer Science 4225, Springer, New York, 814–823.2006

[17] Benetis, R., Jensen, C., Karciauskas, G. & Saltenis, S. "Nearest and Reverse Nearest Neighbor Queries for Moving Objects", The International Journal on Very Large Data Bases, 15 (3) : 229–250.2006

[18] Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K., "Using KNN model for automatic text categorization", Soft Computing –A Fusion of Foundations, Methodologies and Applications 10 (5): 423–430.2006

[19] Hastie, T., Tibshirani, R. & Friendman, J. "The elements of statistical learning: Data Mining, Inference and Prediction", Springer, Stanford, CA, USA, ISBN: 978-0- 387 -84858-7.2009.