EMITTER International Journal of Engineering Technology
Vol. 9, No. 1, June 2021, pp. 126~136
DOI : 10.24003/emitter.v9i1.536

Comparison of Tree Method, Support Vector Machine, Naïve Bayes, and Logistic Regression on Coffee Bean Image

***Rahmat Robi Waliyansyah, Umar Hafidz Asy'ari Hasbullah**

Universitas PGRI Semarang
Jl. Sidodadi Timur No.24, Karangtempel, Kota Semarang, Jawa Tengah 50232
(024) 8316377
E-mail: *rahmat.robi.waliyansyah@upgris.ac.id, umarhafidzah@upgris.ac.id

Received February 10, 2021; Revised March 18, 2021; Accepted April 30, 2021

Abstract

Coffee is one of the many favorite drinks of Indonesians. In Indonesia there are 2 types of coffee, namely Arabica & Robusta. The classification of coffee beans is usually done in a traditional way & depends on the human senses. However, the human senses are often inconsistent, because it depends on the mental or physical condition in question at that time, and only qualitative measures can be determined. In this study, to classify coffee beans is done by digital image processing. The parameters used are texture analysis using the Gray Level Coocurrence Matrix (GLCM) method with 4 features, namely Energy, Correlation, Homogeneity & Contrast. For feature extraction using a classification algorithm, namely Naïve Bayes, Tree, Support Vector Machine (SVM) and Logistic Regression. The evaluation of the coffee bean classification model uses the following parameters: AUC, F1, CA, precision & recall. The dataset used is 29 images of Arabica coffee beans and 29 images of Robusta beans. To test the accuracy of the model using Cross Validation. The results obtained will be evaluated using the confusion Matrix. Based on the results of testing and evaluation of the model, it is obtained that the SVM method is the best with the value of AUC = 1, CA = 0.983, F1 = 0.983, Precision = 0.983 and Recall = 0.983.

Keywords: digital image processing, coffee beans, classification.

1. INTRODUCTION

Coffee is one of the most popular drinks in the community, consumers also vary, from teenagers to the elderly [1]. Coffee is the result of boiling coffee beans that have been roasted and processed into coffee powder [2]. Now coffee is the second most traded commodity after oil [3]. Liberica, arabica, and robusta coffee varieties are coffee varieties that are grown in Indonesia [4].

Coffee undergoes a long process before it can be drunk, starting from harvesting ripe coffee beans in a traditional or modern way using a machine then the coffee beans are dried and turned into coffee logs, then the roasting process is carried out with varying degrees of degree, then the coffee beans are ground & ground into a powder ready-to-brew coffee [2]. The competition of Indonesian coffee products against competing countries is decided by the quality of the processing results of the beans & variants of the coffee plant [5].

Arabica coffee varieties have lower caffeine content & high taste quality compared to robusta so that the price is more expensive, but robusta coffee has the advantage of being resistant to leaf rust disease [2]. The price per tonne for Arabica coffee is US $. 2,498, so that the coffee commodity becomes very prospective for the motor of agro-industry & agribusiness development in Indonesia [3].

Each area where coffee plants are grown has a different composition according to the processing method and the growing environment. So in the market coffee is more synonymous with the place where the coffee is grown because it has something unique in each region, from taste, color, and shape. The diversity of each area ultimately causes many consumers to find it difficult to identify with certainty the coffee traded in the market [2].

Arabica and Robusta coffee beans can be distinguished macroscopically. Robusta coffee beans are smaller than Arabica coffee beans. Arabica coffee beans have a width in the range of 6-8 mm & a length of 8-12 mm, a ratio of width & length of 7-6 mm with a ratio of 1.15-1.0. Coffee beans have a weight range in the range of 100-200 mg & a density between 1.42-1.15 [3].

The increasing number of food counterfeiting activities in Indonesia, especially for the coffee commodity. Counterfeiting is an attempt to change the appearance of food which is done deliberately by replacing / adding food ingredients with the aim of beautifying the appearance of the food in order to get a very large profit so that this has fatal consequences for the customer [4]. To classify coffee that has been faked we can use visuals (eyes), but this can only be used to distinguish coffee beans that have gone through the roasting process.

Coffee beans are generally classified depending on human sight & tradition. However, human vision is often inconsistent, depending on the mental or physical condition of the person at that time, and only qualitative measures can be classified [6]. Among them are ways that can be used, namely with digital images. Digital image processing has the ability to be more precise, objective & sensitive than the ability of human vision. In this way, the coffee classification results are more valid [7].

2. RELATED WORKS

Several studies related to coffee beans include: Testing the level of maturity of coffee beans using artificial neural networks & digital image processing. In this study using the Backpropagation Algorithm to test the ripeness of roasted coffee beans. This method is able to produce an accuracy of 97.5% [8]. Identification of disease in Euthopia coffee using image analysis. In this study, trials were carried out on each segmentation to obtain optimal results. Overall the results from the combined method trials are much better than Otsu, Fuzzy C Means, K-Means and Gaussian Distribution. The best test is on a combination of K-Means and Gaussian Distribution with an accuracy of 92.10% [9]. Identification of Euthophia coffee bean varieties using digital image processing. This research uses the Otsu Method, Fuzzy C Means and K-Mean for the segmentation process. The most optimal result is segmentation using Fuzzy C Mean and Backpropagation Neural Network recognition algorithm. The level of accuracy is 94.54% [10]. Scanning of coffee fakes using digital images and SPA-LDA (linear discriminant analysis). All models show accuracy above 90% [11]. Development of a model to predict the maturity of coffee beans using digital images. The method used is non-linear regression. The test results show an accuracy of above 80% [12]. Development of a system for classification of coffee beans using artificial intelligence methods and Naive Bayes. The parameter used is the color L * a * b. The test results with the Artificial Neural Network Algorithm, the error obtained is 1.15% & Naive Bayes accuracy is 100% for all samples. Identification of coffee counterfeiting using Ultra Violet light and SPA-LDA. From the test results, the identification accuracy rate is 100% using several SPA-LDA matrix models [13]. Development of a Non-Invasive Classification Method to measure the ripeness level of roasted coffee beans using hyperspectral imagery. The model used is the Least Squares Support Vector Machine (LS-SVM). The accuracy obtained is 90.3% [14].

Decision trees have been widely used in research and the results obtained also have a good level of accuracy. Here are some studies using the Decision Tree Method. Rockburst predictions in kimberlite using decision trees with incomplete data. 132 samples were used in this study and the accuracy rate showed 93% [15]. Effect of Flash Stimulation for Migraine Detection Using the Decision Tree Classifier. The experimental results show good accuracy with some noting that flash stimulation affects classification accuracy. In addition, the window length affects classification accuracy indicating that flash stimulation affects classification accuracy [16]. Decision Tree and Random Forest for prediction of outcome in kidney transplant incompatible antibodies. The Decision Tree and Random Forest classification developed in this work predicts early transplant rejection with an accuracy of 85%, which is helpful for clinicians tasked with predicting kidney transplant outcomes prior to clinical intervention [17].

Some research related to the Naïve Bayes Method is the use of the Naïve Bayes method to analyze feedback from students. The results of using this method are able to get good results and also this method can be used to determine the talents and placement of these students [18]. Naïve Bayes is also used to classify abnormalities on the EKG. The results of this method proved to be accurate in analyzing the patient's heart rate abnormalities [19].

There are several studies related to SVM, namely the SVM method used to measure waist circumference. The analysis error was obtained at 4.62 cm from the actual measurement. It is better than manual measurement [20].

Image classification of pigs while lying down. This research discusses the nature and habits of pigs. The level of accuracy obtained is 94% [21].

There are several logostic regression related studies, including the Increasing Accuracy of Classification with the Bootstrap Aggregating Method in Ordinal Logistic Regression. This study aims to improve the classification of ordinal logistic regression using bagging on birth weight. The classification results with bagging ordinal logistic regression were able to reduce classification errors by 20.237% with classification accuracy of 76.67% [22]. Optimization of Logistic Regression in Classification Process Using Genetic Algorithms. This method is an iterative method to get the global optimum. In its application, it uses septic tank data in the East Surabaya area where the results of classification accuracy are generated from Logistic Regression with 11 independent variables and the dependent variable is binary with an accuracy of 54.55%. However, when selected with a Genetic Algorithm, Binary Logistic Regression has a classification accuracy of 90.91% [23].

3. ORIGINALITY

In this study, the image used is an image of a cellphone with coffee beans that have been removed from the skin. The coffee beans used are arabica & robusta coffee beans. The distance to take the image is 20 cm considering the camera's pixel size is only 3.5 megapixels, so if it is taken too far it can reduce the level of image resolution and if it is taken too close the image becomes less focused. Texture analysis uses the Gray Level Coocurence Matrix (GLCM) and feature extraction uses the SVM Classification Algorithm, Tree, Logistic Regression and Naive Bayes. The four methods will be tested using Cross Validation and the results will be evaluated using the Confusion Matrix. The results of this validation test provide information on the four methods,

4. SYSTEM DESIGN

4.1 Framework of thinking

This study uses the GLCM method for texture analysis and the results of the analysis will be classified using the Tree, SVM, Naïve Bayes and Logistic Regression methods. This method is used in the classification of coffee bean types. Testing in this study uses the Confusion Matrix. The applications used in this study were Matlab 2013a and Orange Biolab. Figure 1 is a design thinking framework from a comparison of the classification of coffee beans using the Tree, SVM, Naïve Bayes and Logistic Regression methods.

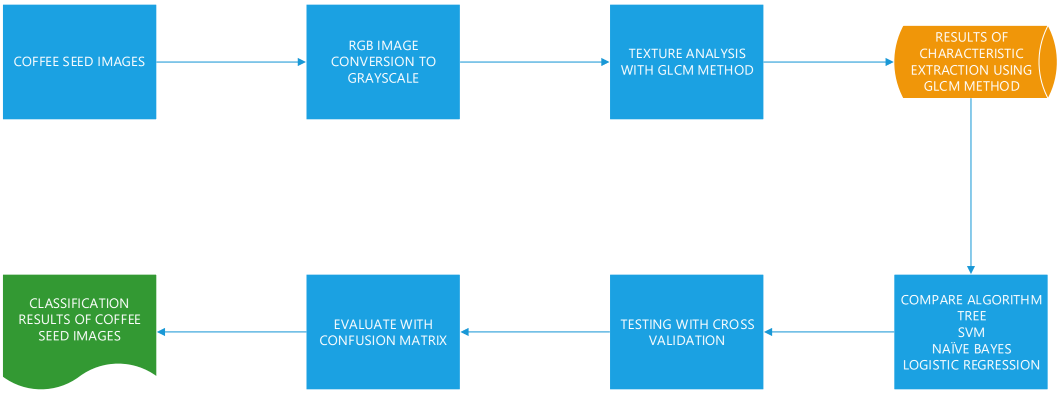
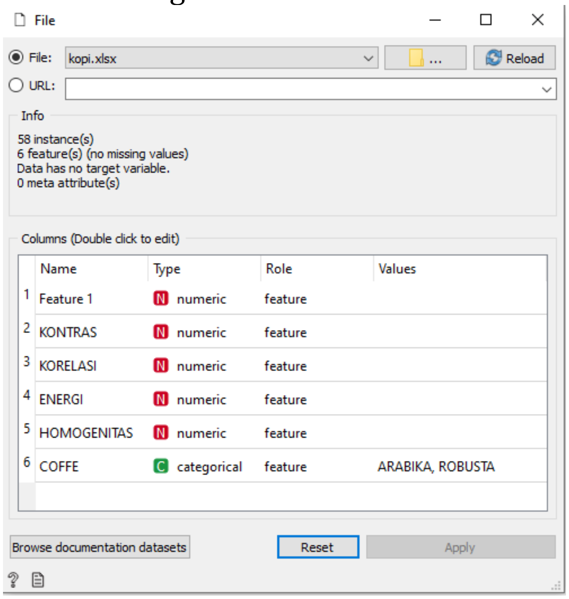

COFFEE SEED IMAGES
RGB IMAGE CONVERSION TO GRAYSCALE
TEXTURE ANALYSIS WITH GLCM METHOD
RESULTS OF CHARACTERISTIC EXTRACTION USING GLCM METHOD
COMPARE ALGORITHM TREE SVM NAÏVE BAYES LOGISTIC REGRESSION
TESTING WITH CROSS VALIDATION
EVALUATE WITH CONFUSION MATRIX
CLASSIFICATION RESULTS OF COFFEE SEED IMAGES

Figure 1. Thinking Framework

4.2 Data retrieval

The data taken is image data from robusta and arabica coffee beans. The coffee bean samples were taken from the UPGRIS Food Technology Laboratory. The coffee beans used are coffee beans that have been split into two parts, then the image is taken using the Andromax A mobile camera with 3.5 Megapixel resolution. The coffee beans that are selected are only green coffee beans & their condition has been removed from the skin. The coffee bean data set uses 4 parameters 1 target. The total number of images taken is 58 images (29 images of robusta coffee beans & 29 images of Arabica coffee beans) which can be seen in Figure 2.



● File: kopi.xlsx
○ URL:

Info

58 instance(s)
6 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

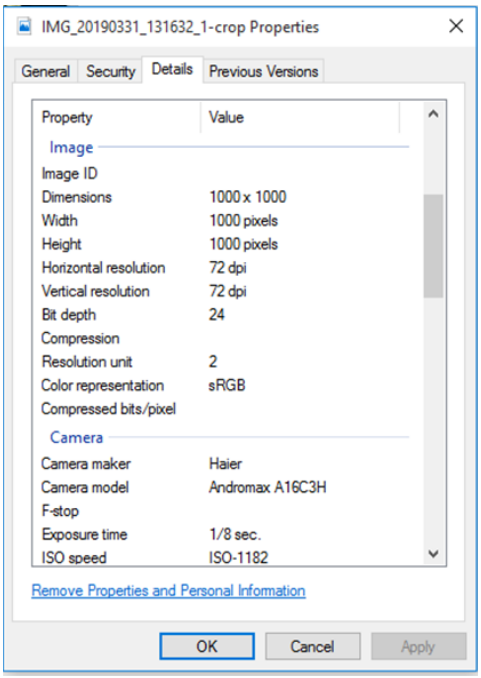
Columns (Double click to edit)

	Name	Type	Role	Values
1	Feature 1	N numeric	feature	
2	KONTRAS	N numeric	feature	
3	KORELASI	N numeric	feature	
4	ENERGI	N numeric	feature	
5	HOMOGENITAS	N numeric	feature	
6	COFFE	C categorical	feature	ARABIKA, ROBUSTA

Figure 2. Coffee bean data set

4.3 Preprocessing

The image of the coffee bean that is taken will be processed by the image cutting process. Unused portions of the image will be discarded. Image of coffee bean cut to 1000 x 1000 pixel size. Image cropping is done with the aim of speeding up the digital image processing by reducing the computational load shown in Figure 3.

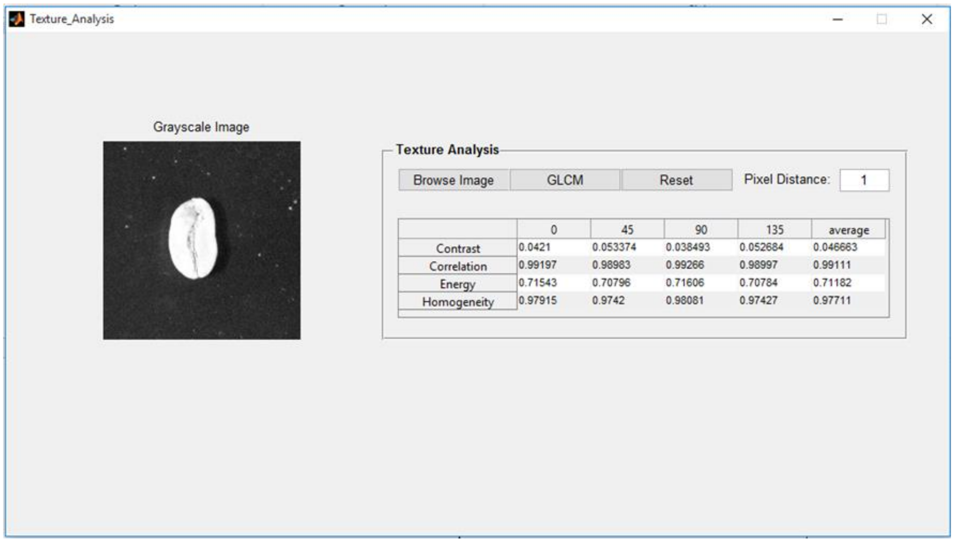

IMG_20190331_131632_1-crop Properties

Property	Value
Image	
Image ID	
Dimensions	1000 x 1000
Width	1000 pixels
Height	1000 pixels
Horizontal resolution	72 dpi
Vertical resolution	72 dpi
Bit depth	24
Compression	
Resolution unit	2
Color representation	sRGB
Compressed bits/pixel	
Camera	
Camera maker	Haier
Camera model	Andromax A16C3H
F-stop	
Exposure time	1/8 sec.
ISO speed	ISO-1182

Figure 3. The cropped image is 1000 x 1000 pixels in size

4.4 Extraction of Gray Level Co-ocurrence Matrix (GLCM) Texture

Prior to feature extraction using the GLCM method. The input image that is still in RGB format will first be converted into a gray level image. In this study, the GLCM method uses the Matlab 2013a application with 4 parameters, namely, Energy, Correlation, Homogeneity & Contrast and the pixel distance used is 1 pixel. After the texture feature extraction process using the GLCM method is carried out, the values of the four parameters (Energy, Correlation, Homogeneity & Contrast) will be stored as training and test data. Extraction of texture features using the GLCM method is shown in Figure 4.


Texture_Analysis

Pixel Distance: 1

	0	45	90	135	average
Contrast	0.0421	0.053374	0.038493	0.052684	0.046663
Correlation	0.99197	0.98983	0.99266	0.98997	0.99111
Energy	0.71543	0.70796	0.71606	0.70784	0.71182
Homogeneity	0.97915	0.9742	0.98081	0.97427	0.97711

Figure 4. Feature extraction using the GLCM method

5. EXPERIMENT AND ANALYSIS

5.1 Image Classification of Coffee Beans

The classification process uses Algortima SVM, Tree, Logistic Regression and Naïve Bayes. These methods are quite good in the classification of an object. The classification process uses data from feature extraction resulting from the GLCM method, namely contrast, correlation, energy and homogeneity. This section uses the Orange Biolab application. Steps in coffee bean classification and parameters used in Figure 5.

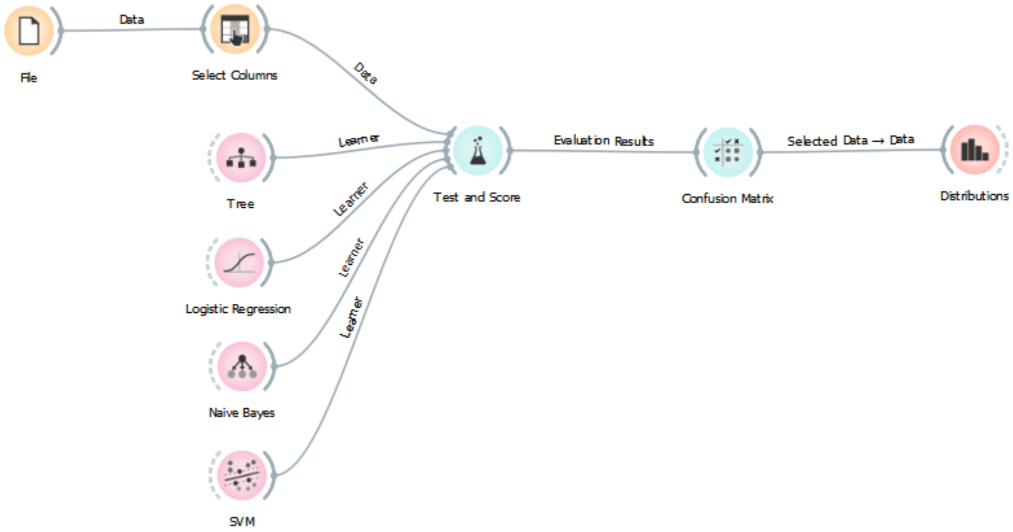
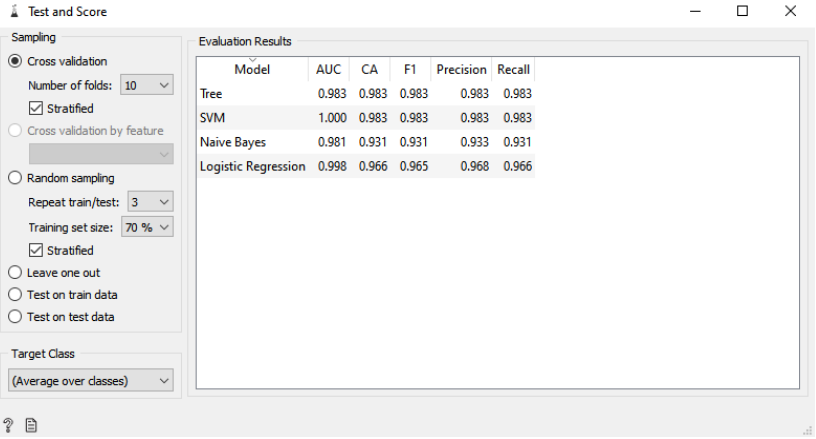


Figure 5. Comparison of Classification of Coffee Beans

5.2 Testing and Evaluation of Coffee Bean Image Classification Models

The testing process uses the Cross Validation Method with a number of folds of 10. A total of 58 samples are used in this section. Tests conducted using 5 parameters to measure the accuracy of the model, namely AUC, F1, CA, precison and recall. Figure 6 shows the results of the classification test, where the SVM method has the best results.



Sampling

● Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 3

Training set size: 70 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.983	0.983	0.983	0.983	0.983
SVM	1.000	0.983	0.983	0.983	0.983
Naive Bayes	0.981	0.931	0.931	0.933	0.931
Logistic Regression	0.998	0.966	0.965	0.968	0.966

Figure 6. Model Testing

In evaluating the algorithm performance of Machine Learning, we use the Confusion Matrix reference. Confusion matrix is also often called error matrix. Basically confusion matrix provides information on the comparison of the classification results carried out by the system (model) with the actual classification results. The confusion matrix is in the form of a matrix table that describes the performance of the classification model on a series of test data whose true value is known. The confusion matrix results are shown in Figures 7 to 10.

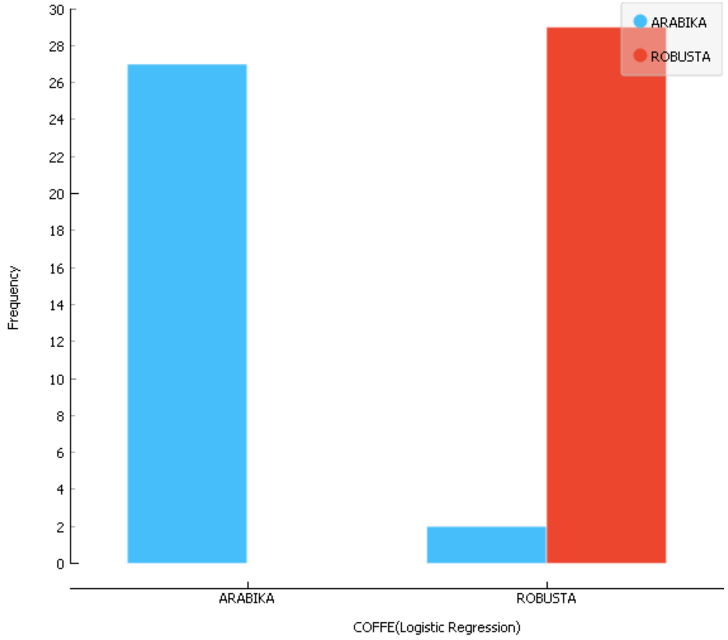


Figure 7. Confusion Matrix of Logistic Regression

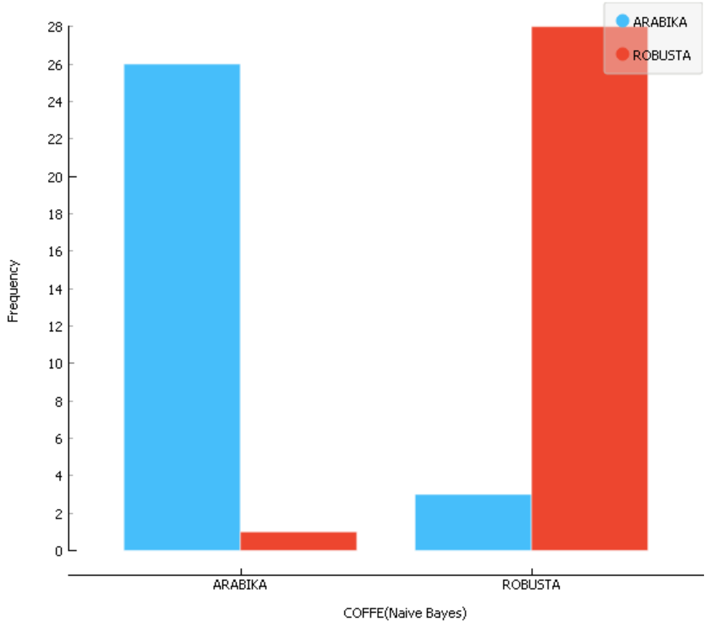


Figure 8. Confusion Matrix from Naïve Bayes

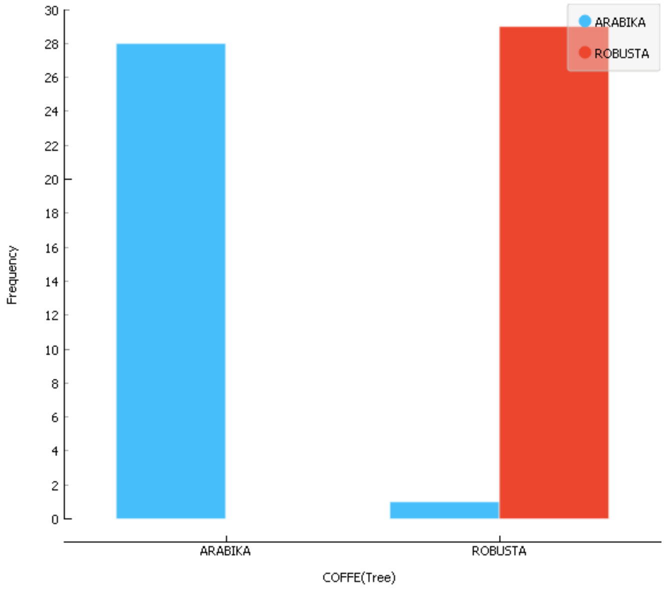


Figure 9. Confusion Matrix of the Tree

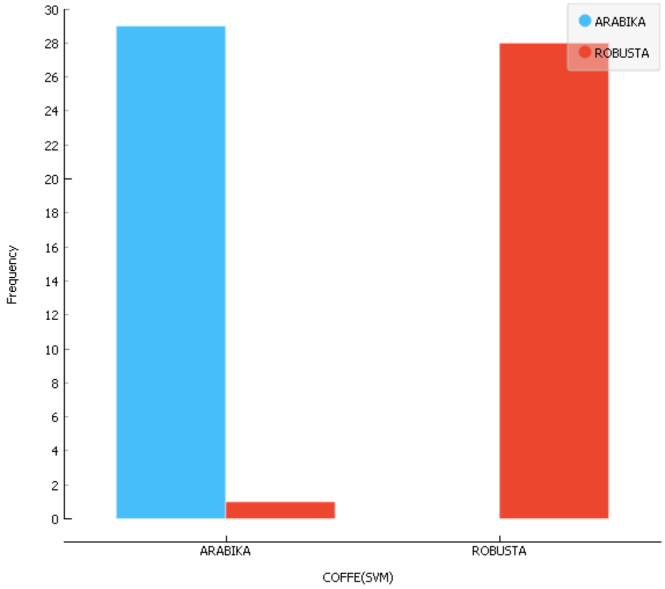


Figure 10. Confusion Matrix from SVM

6. CONCLUSION

Based on the results of testing and evaluation, for the Tree Method obtained AUC = 0.983, CA = 0.983, F1 = 0.983, Precision = 0.983 and Recall = 0.983. The SVM method obtained AUC = 1, CA = 0.983, F1 = 0.983, Precision = 0.983 and Recall = 0.983. The Naïve Bayes method obtained AUC = 0.981, CA = 0.931, F1 = 0.931, Precision = 0.933 and Recall = 0.931. Logistic Regression Method AUC = 0.998, CA = 0.966, F1 = 0.965, Precision = 0.968 and Recall = 0.966.

The SVM algorithm is superior to the other three methods. SVM has advantages including determining the distance using support vectors so that the computation process becomes fast. This is because SVM is a machine learning method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. The best separator hyperplane between the two classes can be found by measuring *margin* the hyperplane and find the maximum point.

Many data mining or machine learning techniques are developed with the assumption of linearity, so that the resulting algorithms are limited to linear cases. SVM can work on non-linear data by using a kernel approach to the initial data set feature. SVM has a basic principle of linear classifier, namely classification cases that can be separated linearly, but SVM has been developed to work on non-linear problems by incorporating the kernel concept in a high-dimensional workspace.

Acknowledgments

Thank you to the Head of the Food Technology Laboratory for helping to provide Arabica and Robusta coffee beans which have been split in half and separated from the skin.

REFERENCES

[1] TR Nanda, Zulhelmi, and M. Syaryadhi, **Designing Coffee Fruit Sorting System Based on Color Using Digital Image Technique Based on Atmega 328p Microcontroller**, *KITEKTRO*, vol. 3, no. 2, pp. 76–83, 2018.

[2] BD Argo and M. Andreane, **Identification of Robusta Coffee Bean and Powder Parameters Using Machine Vision and Artificial Neural Network (ANN) Methods**, *J. Agricultural Engineering. Trop. and Biosist.*, vol. 5, no. 2, pp. 150–162, 2017.

[3] DW Soedibyo, U. Ahmad, KB Seminar, and IDM Subrata, **Designing Smart Sorting System Based on Image Processing for Rice Coffee**, *J. Agricultural Engineering.*, vol. 24, no. 02, pp. 67–74, 2010.

[4] M. Yulia, R. Iriani, D. Suhandy, S. Waluyo, and C. Sugianti, **Study On The Use Of Uv-Vis Spectroscopy And Chemometrics To Quickly Identify The Falsification Of Arabica And Robusta Coffees**, *J. Tech. Question. Lampung*, vol. 6, no. 1, pp. 43–52, 2017.

[5] PS Maria and M. Rivai, **Classification of Quality of Coffee Beans Using Image Processing and Fuzzy Logic**, in *National Seminar: Initiating the Awakening of Local Superior Commodities for Agriculture and Maritime Affairs, Faculty of Agriculture, Trunojoyo University, Madura*, 2013, pp. 773–780.

[6] N. Ulum, IGLPE Prismana, and RAJ Firdaus, **Identification of Coffee Beans Using Digital Images Using City Block Distance Classification Methods**, *INOVATE*, vol. 03, no. 01, pp. 30–37, 2018.

[7] SA Mutallib, J. Nugroho, and N. Bintoro, **Identify the Blending of Arabica and Robusta Coffee with Electronic Nose Using a Pattern Recognition System**, in *PERTETA National Seminar Proceedings*, 2012, pp. 154–163.

[8] TH Nasution and U. Andayani, **Recognition of Roasted Coffee Bean Levels using Image Processing and Neural Network**, in *Annual Applied Science and Engineering Conference*, 2017, vol. 1, pp. 1–8.

[9] AD Mengistu, SG Mengistu, and DM Alemayehu, **Image analysis for**

Ethiopian Coffee Plant Diseases Identification Abrham, *Int. J. Biometrics Bioinforma.*, vol. 10, no. 1, pp. 1–11, 2016.

[10] AD Mengistu, **The Effects of Segmentation Techniques in Digital Image Based Identification of Ethiopian Coffee Variety**, *TELKOMNIKA*, vol. 16, no. 2, pp. 713–717, 2018.

[11] UT de CP Souto *et al.*, **Screening for Coffee Adulteration Using Digital Images and SPA-LDA**, Food Anal. Methods, vol. 8, no. 6, pp. 1515–1521, 2015.

[12] JDB Vanegas *et al.*, **Developing predictive models for determining physical properties of coffee beans during the roasting process**, Ind. Crops Prod., Vol. 112, pp. 839–845, 2018.

[13] UT de CP Souto *et al.*, **Identification of adulteration in ground roasted coffees using UV-Vis spectroscopy and SPA-LDA**, LWT - Food Sci. Technol., Vol. 63, no. 2, pp. 1037–1041, 2015.

[14] B. Chu, K. Yu, Y. Zhao, and Y. He, **Development of Noninvasive Classification Methods for Different Roasting Degrees of Coffee Beans Using Hyperspectral Imaging**, sensors, vol. 18, no. 4, pp. 1–15, 2018.

[15] Y. Pu, DB Apel, and B. Lingga, **Rockburst prediction in kimberlite using decision tree with incomplete data**, *J. Sustain. Min.*, vol. 17, no. 3, pp. 158–165, 2018.

[16] A. Subasi, A. Ahmed, and E. Alickovic, **Effect of flash stimulation for migraine detection using decision tree classifiers**, in *Procedia Computer Science*, 2018, vol. 140, pp. 223–229.

[17] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, **Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation**, *Biomed. Signal Process. Control*, vol. 52, pp. 456–462, 2019.

[18] S. Maitra, S. Madan, R. Kandwal, and P. Mahajan, **Mining authentic student feedback for faculty using the Naïve Bayes classifier**, in *Procedia Computer Science*, 2018, vol. 132, pp. 1171–1183.

[19] S. Padmavathi and E. Ramanujam, **Naïve Bayes Classifier for ECG abnormalities using Multivariate Maximal Time Series Motif**, *Procedia Comput. Sci.*, vol. 47, no. C, pp. 222–228, 2015.

[20] D. Seo, E. Kang, Y. mi Kim, SY Kim, IS Oh, and MG Kim, **SVM-based waist circumference estimation using Kinect**, *Comput. Biomed Methods Programs.*, vol. 191, pp. 1-6, 2020.

[21] A. Nasirahmadi *et al.*, **Automatic scoring of lateral and sternal lying posture in grouped pigs using image processing and Support Vector Machine**, Comput. Electron. Agric., Vol. 156, December 2018, pp. 475–481, 2019.

[22] IKP Suniantara, IGEW Putra, and G. Suwardika, **Improving Accuracy of Classification with the Bootstrap Aggregating Method in Ordinal Logistic Regression**, *INTENSIVE J. Ilm. Researcher. and Application of Technol. Sist. Inf.*, vol. 3, no. 1, p. 32, 2019.

[23] A. Salim and MR Alfian, **Optimizing Logistic Regression in the Classification Process Using Genetic Algorithms**, *J. Technol. Inf. and Applied.*, vol. 6, no. 2, pp. 50–55, 2019.