

Classification of Arabica Coffee Green Beans Using Digital Image Processing Using the K-Nearest Neighbor Method

Nurun Najmi Amanina*¹, Galuh Wilujeng Saraswati²

Teknik Informatika, Universitas Dian Nuswantoro

Email : najmiamanina05@gmail.com*¹, galuhwilujengs@dsn.dinus.ac.id²

*Corresponding author

Abstract - Arabica coffee is the largest commodity produced by farmers in Pagergunung Village, Bulu District, Temanggung Regency. Coffee production in recent years has increased rapidly by 80% with the increasing lifestyle of the Indonesian people marked by the number of coffee shop buildings in various regions, and of course the demand for Arabica coffee has also increased, therefore it must improve the quality or quality of the coffee itself. However, in determining and classifying the quality of coffee beans, errors often occur due to the lack of understanding of the farmers in processing coffee. Based on this, the purpose of this research is to classify using the K- Nearest Neighbor method and feature extraction using the average value of Red-Green-Blue (RGB) color in determining the quality and quality of coffee beans according to grade so that they can get a high selling price. In this study using as many as 150 training image data and 150 testing image data, the results of this classification accuracy are 80% using k=1.

Keywords - KNN Algorithm, RGB, Coffee, Green Bean, Grade

1. INTRODUCTION

Arabica coffee is one of two types of coffee that are widely available in the market. Arabica coffee has a higher taste quality and lower caffeine content than other coffees. Based on this, the demand for Arabica coffee has also increased, so processing is needed to improve the quality of the Arabica coffee beans in order to get a higher price. However, in processing, determining and classifying the quality of coffee beans, errors often occur due to lack of understanding in processing coffee (Pamuji, 2019). Therefore, a study was conducted to classify Arabica coffee green beans based on grade using technology to make it more efficient and accurate.

Research by Nelly Oktavia Adiwijaya, Hammam Iqomatuddin Romadhon, Januar Adi Putra, and Dewangga Putra Kuswanto with title "The Quality of Coffee Bean Classification System Based on Color by Using K-Nearest Neighbor Method". In this study, researchers used image processing with the K-Nearest Neighbor method. This study discusses the K-Nearest Neighbor method in classifying the quality of coffee beans by class using 90 training data from 3 classes and 30 test data from each class. Accuracy results with k = 3, k = 5, and k = 7 are the same, namely 83% (Adiwijaya *et al.*, 2022).

Research by Siti Raysyah, Veri Arinal, and Dadang Iskandar Mulyana with title "Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode KNN dan PCA". This study uses a classification based on the maturity level of coffee cherries, namely raw, moderately ripe, and ripe. This study uses the RGB and HSV methods assisted by

the K-Nearest Neighbor method. Using 135 datasets divided into 90 training images and 45 test images, which resulted in an accuracy of 97.7% with $k=3$ (Raysyah, Veri Arinal and Dadang Iskandar Mulyana, 2021).

Research by Ariska Restu Ginanjar, and Enny Itje Sela with title “Sistem Deteksi Jenis Cacat Biji Kopi dengan Algoritma K-Nearest Neighbor”. This study discusses the defect detection system in coffee beans using the RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value) color model features. In his research Ariska, et al. used 60 coffee beans for training data and 40 coffee beans for test data. The resulting accuracy with $k=3$ in the data processing uses the RGB (Red, Green, Blue) color model feature of 95% (Ginanjar, 2019).

Research by Cinantya Paramita, Eko Hari Rachmawanto, Christy Atika Sari, and De Rosal Ignatius Moses Setiadi with title “Klasifikasi Jeruk Nipis Terhadap Tingkat Kematangan Buah Berdasarkan Fitur Warna Menggunakan K-Nearest Neighbor”. This study discusses the classification of limes with RGB color features (Red, Green, Blue) based on the level of skin color maturity with 5 categories, namely raw, slightly ripe, ripe, perfectly ripe, and rotten. The data used is 75 data which is divided into 50 training data and 25 testing data. The results of the classification accuracy of this study with $k = 3$ is 92% (Paramita *et al.*, 2019).

Research by Andhika Ryan Pratama, Muhammad Mustajib, and Aryo Nugroho with title “Deteksi Citra Uang Kertas dengan Fitur RGB Menggunakan K-Nearest Neighbor”. This study discusses the detection of banknotes using color feature extraction, namely RGB (Red, Green, Blue). The number of datasets used is 40 images. This study uses old & new Rp. 2000 banknotes, and Rp. 5000 old & new banknotes. The results obtained that from the 16 test data obtained 15 data were detected correctly, and the accuracy obtained was 93.7% with $k = 5$ (Pratama, Mustajib and Nugroho, 2020).

Based on the research above, the research idea was obtained in classifying Arabica coffee green beans based on grade using the K-Nearest Neighbor method. With this research, it is hoped that farmers will understand the processing of Arabica coffee green beans based on their grade, and be able to increase the price of Arabica coffee.

1.1. Green Bean Arabica Coffee

Green coffee beans are raw coffee beans after going through the process of peeling from the skin and have not been roasted which is green in color. Most coffee farmers and experts distinguish coffee quality by looking at the color, shape and texture of green beans or raw coffee. (Nugroho and Sebatubun, 2020).

1.2. Grade Arabica Coffee

Grade is the level of coffee quality. Arabica coffee has several grades in classifying coffee quality, this is also included in the Indonesian National Standard (SNI) with SNI number 01-2907-2008 in order to adjust standards with other countries that also produce coffee. Arabica coffee has 6 quality grades consisting of grades 1, 2, 3, 4, 5, and 6 (Rizal, 2019).

1.3. Image Processing

Image processing is a processing method using images or images to get data from an image. There are many ways to process the image, one of which is the K-Nearest Neighbor method with RGB (Red, Green, Blue) color feature extraction. The RGB color feature (Red, Green, Blue) has 3 basic color components, namely red (R), green (G), and blue (B). Color feature extraction is done by calculating the average of each RGB value of an image or image. After the average value data is obtained, the results are used as input for further processing.

1.4. K-Nearest Neighbor

K-Nearest Neighbor is one of the supervised learning algorithms that processes training data based on the input and output information that has been given, then the system learns the pattern of the data which will be used as a reference to determine information from other data (Abijono, Santoso and Anggreini, 2021). The K-Nearest Neighbor method is used for data classification by determining the calculation of the nearest neighbor distance between training data and testing data using the Euclidean distance formula which depends on the value of k (Paramita *et al.*, 2019).

1.5. Confusion Matrix

To determine the size of the prediction accuracy (performance) of a classifier with more than 2 (two) target classes (multiclass) it is necessary to calculate the confusion matrix which will be used to determine the results of precision, recall, f1-score, and the level of accuracy.

2. RESEARCH METHOD

2.1. Research Object

The process of taking greenbean arabica coffee images is carried out using an OPPO Reno4F cellphone camera which has a 48 megapixel camera quality, which is assisted by a monopod as high as 10 cm from the object distance and is magnified or zoomed by 5x on the camera. The background used is white HVS paper, because it strengthens the color of the object. Taking pictures of each coffee bean is done back and forth or front and back so that it can be seen clearly because of the different front and back shapes. After the image capture process, preprocessing is carried out by removing the background and resizing the size from the original 3000x3000 pixels to 500x500 pixels.





Researchers took 50 alternating image data for each grade so that a total of 300 images for the training dataset, consisting of 50 grade 1 images, 50 grade 2 images, 50 grade 3 images, 50 grade 4 images, 50 grade 5 images, and 50 grade images. 6. For data testing, there are 50 alternating image data for each grade so that a total of 300 images for the training dataset, consisting of 50 grade 1 images, 50 grade 2 images, 50 grade 3 images, 50 grade 4 images, 50 grade 5 images, and 50 grade 6 images.

Based on interviews with middlemen in Pagergunung Village, there are defective coffee beans to distinguish each grade:

Table 1. Grade Green Bean Arabica Coffee

No.	Grade	Good	Defect	Total Green Beans
1.	Grade 1	25	0	25
2.	Grade 2	22	3	25
3.	Grade 3	20	5	25
4.	Grade 4	15	10	25
5.	Grade 5	7	18	25
6.	Grade 6	2	23	25

Table 2. Pre-Processing

Before pre-processing	After pre-processing
	
	

2.2. Research Stages

In conducting this research, the first to acquire Arabica coffee green bean images was divided into 2 parts, namely training and testing images. After that, preprocessing is done by removing the background. Then perform feature extraction by taking the average RGB color value (Red, Green, Blue). Then normalization is performed using decimal scaling, and the last step is to classify it using the K-Nearest Neighbor method by evaluating it using accuracy, precision, and recall.

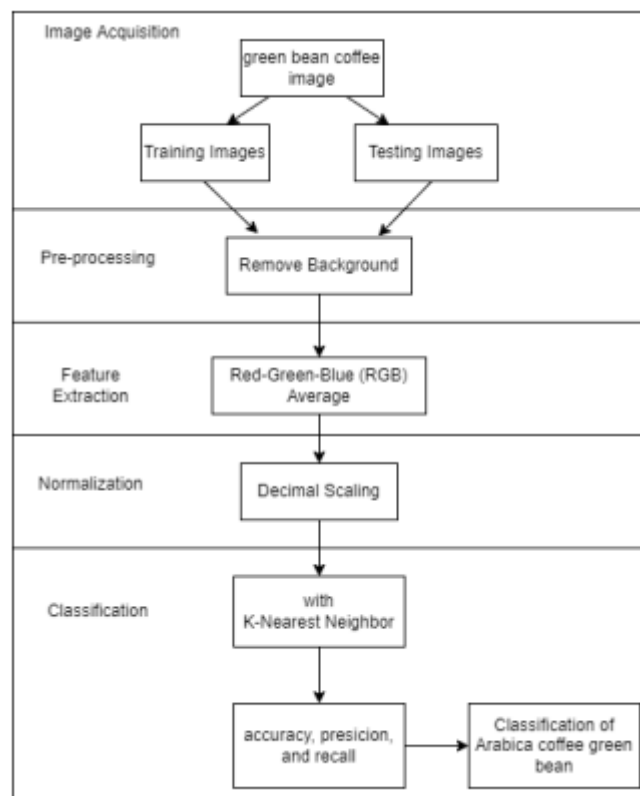


Figure 1. Research Methods





3. RESULTS AND DISCUSSION

This study uses a Google Collaboratory notebook with the Python programming language to create a system for classifying Arabica coffee green beans based on grade. This

classification uses the K-Nearest Neighbor method with feature extraction of the average RGB (Red, Green, Blue) value. The grades used are grade 1, grade 2, grade 3, grade 4, grade 5, and grade 6 with defects in each grade.

The following are examples of good quality coffee images and defective quality in this study:

Table 3. Green Bean Arabica Coffee

Good Quality Coffee Beans (Front)	Good Quality Coffee Beans (Back)	Defect Quality Coffee Beans (Front)	Defective Quality Coffee Beans (Back)
			

There are several stages in this research, namely the search for the average value of RGB (Red, Green, Blue), the search for Euclidean distance, and classification.

3.1. Means RGB

The first stage in this research is extracting features by taking the average RGB (Red, Green, Blue) value. At this stage the system calculates the average value for each training and testing image of 300 images (front and back). The results of these calculations will be used as a reference in the classification of Arabica coffee green beans based on grade. The following is a sample table of RGB average values from 10 green bean images (front, back) :

Table 4. Mean green bean

No	Label	meanRD	meanRB	meanGD	meanGB	meanBD	meanBB
1	grade1	13,870833	14,092221	15,434829	15,93794	16,13433	16,733355
2	grade1	14,143049	14,139851	15,998287	15,809836	16,796919	16,31332
3	grade1	13,791307	13,813822	15,207637	15,404626	15,808628	16,07488
4	grade1	14,362682	14,486639	16,015822	16,248656	16,612214	16,84597
5	grade1	16,171922	16,100894	18,144538	18,172087	18,875724	18,948602

3.2. Euclidean Distance

The classification method using K-Nearest Neighbor in this study uses the Euclidean Distance formula in calculating the distance between neighbors. Here is the formula for Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Description: d = proximity distance
x = training data
y = testing data
n = number of attributes between 1 to n
i = individual attribute between 1 to n

(Raysyah, Veri Arinal and Dadang Iskandar Mulyana, 2021)

The process of calculating Euclidean Distance between 5 training data image samples (x) and 1 testing image sample (y) can be seen below:

Figure (1,1)

$$= \sqrt{(13,870833 - 16,475036)^2 + (14,092221 - 16,316254)^2 + (15,434829 - 18,478607)^2 + (15,93794 - 18,273906)^2 + (16,13433 - 19,160383)^2 + (16,733355 - 18,930663)^2} = 6,358827$$

Figure (2,1)

$$= \sqrt{(14,143049 - 16,475036)^2 + (14,139851 - 16,316254)^2 + (15,998287 - 18,478607)^2 + (15,809836 - 18,273906)^2 + (16,796919 - 19,160383)^2 + (16,31332 - 18,930663)^2} = 5,902116$$

Figure (3,1)

$$= \sqrt{(13,791307 - 16,475036)^2 + (13,813822 - 16,316254)^2 + (15,207637 - 18,478607)^2 + (15,404626 - 18,273906)^2 + (15,808628 - 19,160383)^2 + (16,07488 - 18,930663)^2} = 7,196272$$

Figure (4,1)

$$= \sqrt{(14,362682 - 16,475036)^2 + (14,486639 - 16,316254)^2 + (16,015822 - 18,478607)^2 + (16,248656 - 18,273906)^2 + (16,612214 - 19,160383)^2 + (16,84597 - 18,930663)^2} = 5,368015$$

Figure (5,1)

$$= \sqrt{(16,171922 - 16,475036)^2 + (16,100894 - 16,316254)^2 + (18,144538 - 18,478607)^2 + (18,172087 - 18,273906)^2 + (18,875724 - 19,160383)^2 + (18,948602 - 18,930663)^2} = 0,584448$$

3.3. KNN Classification

The following are the results of the classification of grade 1 using the Euclidean Distance K- Nearest Neighbor method using the value k = 1 to k = 10.

Table 5. Classification Results k=1

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1

Table 6. Classification Results k=2

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1

Table 7. Classification Results k=3

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2

Table 8. Classification Results k=4

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2

Table 9. Classification Results k=5

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3

Table 10. Classification Results k=6

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3
(16, 1)	2,00415	grade1	4

Table 11. Classification Results k=7

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3
(16, 1)	2,00415	grade1	4
(50, 1)	2,107586	grade2	4

Table 12. Classification Results k=8

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3
(16, 1)	2,00415	grade1	4
(50, 1)	2,107586	grade2	4
(73, 1)	2,195615	grade3	4

Table 13. Classification Results k=9

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3
(16, 1)	2,00415	grade1	4
(50, 1)	2,107586	grade2	4
(73, 1)	2,195615	grade3	4
(11, 1)	2,716227	grade1	5

Table 14. Classification Results k=10

Figure per grade	Euclidean Distance	Label	Results
(5, 1)	0,584448	grade1	1
(48, 1)	0,73016	grade2	1
(24, 1)	0,762383	grade1	2
(75, 1)	1,575665	grade3	2
(6, 1)	1,85098	grade1	3
(16, 1)	2,00415	grade1	4
(50, 1)	2,107586	grade2	4
(73, 1)	2,195615	grade3	4
(11, 1)	2,716227	grade1	5
(15, 1)	3,552712	grade1	6

The results of the green bean classification of Arabica coffee based on the grade of 150 testing data of Arabica coffee green beans against 150 Arabica coffee green bean training data using a value of k = 1, there were 30 data that did not match the target. Accuracy can be obtained by :

$$\begin{aligned}
 \text{Accuracy} &= \frac{\text{correct classification results}}{\text{testing data}} \times 100\% \\
 &= \frac{120}{150} \times 100\% \\
 &= 80\%
 \end{aligned}$$

The accuracy obtained by using k=1 is 80%.

4. CONCLUSION

Based on research on the Arabica coffee green bean classification system based on grade with the K-Nearest Neighbor method using the Python Google Colab programming

language, it was found that the accuracy obtained by using the value of $k = 1$ was 80%, $k = 2$ was 73%, $k = 3$ was 76,67%, $k = 4$ was 70%, $k = 5$ was 70,67%, $k = 6$ was 68,667%, $k = 7$ was 69,3%, $k = 8$ was 70%, $k = 9$ was 68%, and $k = 10$ was 71,33%. The results of this study are expected to assist coffee farmers in classifying Arabica coffee green beans based on grade in order to improve the quality and price of coffee, and the economic welfare of coffee farmers.

REFERENCES

- [1] Abijono, H., Santoso, P. and Anggreini, N. L. (2021) 'Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data', *Jurnal Teknologi Terapan: G-Tech*, 4(2), pp. 315–318. doi: 10.33379/gtech.v4i2.635.
- [2] Adiwijaya, N. O. *et al.* (2022) 'The quality of coffee bean classification system based on color by using k-nearest neighbor method', *Journal of Physics: Conference Series*, 2157(1). doi: 10.1088/1742-6596/2157/1/012034.
- [3] Arboleda, E. R., Fajardo, A. C. and Medina, R. P. (2018) 'Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors', in *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 1–5. doi: 10.1109/ICIRD.2018.8376326.
- [4] Ginanjar, A. R. (2019) 'Sistem Deteksi Jenis Cacat Biji Kopi dengan Algoritma K-Nearest Neighbor'.
- [5] Ikhsan, D., Utami, E. and Wibowo, F. W. (2020) 'Metode Klasifikasi Mutu Greenbean Kopi Arabika Lanang Dan Biasa Menggunakan K-Nearest Neighbor Berdasarkan Bentuk', *Jurnal Ilmiah SINUS*, 18(2), p. 1. doi: 10.30646/sinus.v18i2.456.
- [6] Kohn, T. (2017) 'Teaching Python Programming to Novices: Addressing Misconceptions and Creating a Development Environment ETH Library', (24076), p. 166. Available at: <https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/129666/eth-50720-02.pdf>.
- [7] Linge, S. and Langtangen, H. P. (2020) *Programming for Computations - Python*, Springer Open. Available at: <http://link.springer.com/10.1007/978-3-030-16877-3>.
- [8] Nugraha, D. A. and Wiguna, A. S. (2018) 'Klasifikasi Tingkat Roasting Biji Kopi Menggunakan Jaringan Syaraf Tiruan Backpropagation Berbasis Citra Digital', *SMARTICS Journal*, 4(1), pp. 1–4. doi: 10.21067/smartics.v4i1.2165.
- [9] Nugroho, M. A. and Sebatubun, M. M. (2020) 'Klasifikasi Varietas Kopi Berdasarkan Green Bean Coffee Menggunakan Metode Machine Learning', *Journal of Information System Management (JOISM)*, 1(2), pp. 1–5. doi: 10.24076/joism.2020v1i2.24.
- [10] Pamuji, R. (2019) 'Identifikasi Citra Biji Kopi Arabika dan Robusta Menggunakan Learning Vector Quantization', *Naskah Publikasi Program Studi Teknik Informatika. Universitas Mercu buana Yogyakarta*, (November), pp. 1–7. Available at: <http://eprints.mercubuana-yogya.ac.id/6648/>.
- [11] Paramita, C. *et al.* (2019) 'Klasifikasi Jeruk Nipis Terhadap Tingkat Kematangan Buah Berdasarkan Fitur Warna Menggunakan K-Nearest Neighbor', *Jurnal Informatika: Jurnal Pengembangan IT*, 4(1), pp. 1–6. doi: 10.30591/jpit.v4i1.1267.
- [12] Pratama, A. R., Mustajib, M. and Nugroho, A. (2020) 'Deteksi Citra Uang Kertas dengan Fitur RGB Menggunakan K-Nearest Neighbor', *Jurnal Eksplora Informatika*, 9(2), pp. 163–172. doi: 10.30864/eksplora.v9i2.336.
- [13] Raysyah, S. R., Veri Arinal and Dadang Iskandar Mulyana (2021) 'Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode Knn Dan Pca',

JSil (Jurnal Sistem Informasi), 8(2), pp. 88–95. doi: 10.30656/jsii.v8i2.3638.

- [14] Rizal, M. A. (2019) 'Klasifikasi Mutu Biji Kopi Menggunakan Metode K – Nearest Neighbor Berdasarkan Warna Dan Tekstur', *Universitas Teknologi Yogyakarta*, pp. 1–8.
- [15] Sulistyaningtyas, A. R. (2017) 'Pentingnya Pengolahan basah (Wet Processing) Buah kopi Robusta (*Coffea var. robusta*) untuk menurunkan resiko kecacatan biji hijau saat coffe grading', *Prosiding Seminar Nasional Publikasi Hasil-Hasil Penelitian dan Pengabdian Masyarakat*, pp. 90–94.
- [16] Temanggung, P. K. (2017) *DESA PAGERGUNUNG KECAMATAN BULU*. Available at: <https://laman.temanggungkab.go.id/info/detail/82/212/desa-pagergunung.html>.