


# Performance Analysis of Genetic Algorithms and KNN Using Several Different Datasets

<sup>1</sup>Enda Putri Atika, <sup>1\*</sup>Yudha Riwanto 

<sup>1</sup> Amikom University, Yogyakarta, Indonesia

\* Corresponding Author: yudha.riwanto@amikom.ac.id



**Citation:** E.P.Atika, Y.Riwanto, "Performance Analysis of Genetic Algorithms and KNN Using Several Different Datasets", *Iota*, 2024, ISSN 2774-4353, Vol.04, 03. <https://doi.org/10.31763/iota.v4i3.767>

Academic Editor : Adi, P.D.P

Received : July, 11 2024

Accepted : July, 22 2024

Published : August, 13 2024

**Publisher's Note:** ASCEE stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2024 by authors. Licensee ASCEE, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution Share Alike (CC BY SA) license(<https://creativecommons.org/licenses/by-sa/4.0/>)

**Abstract:** This research aims to increase the accuracy of the classification of mango, corn, and potato leaf types using an approach involving feature selection with a genetic algorithm (Genetic Algorithm), classification with K-Nearest Neighbors (KNN), and image processing in the HSV color space (Hue, Saturation). , Value). The dataset used consists of more than 1500 image samples for each type of leaf, with a total of 10 tests carried out. The research process begins with processing leaf images in HSV color space to extract more representative color information. Next, a genetic algorithm is applied to select the most relevant features from the processed image. The selected features are then used as input for the KNN model in the classification process. The test results show that the proposed method can achieve a classification accuracy of 87,9%. This shows that the combination of image processing in the HSV color space, feature selection using a genetic algorithm, and classification with KNN can improve performance in recognizing leaf types. This research makes significant contributions to the field of image processing and classification and shows the potential of using genetic algorithms for feature selection in pattern recognition applications.

**Keywords:** genetic algorithm; K-Nearest Neighbors; dataset; classification process; accuracy of classification

## 1. Introduction

The world of technology is currently developing increasingly rapidly in all fields, the use of artificial intelligence is also increasing because it makes human work easier. One of how artificial intelligence is used is to classify objects, such as types of diseases in plants, levels of fruit ripeness, types of fruit or vegetables, and many more. Classification is an important part of artificial intelligence because it is very helpful in grouping or categorizing various objects.

Various algorithms have been developed for this classification, one of which is KNN (K-Nearest Neighbors) which is known for its good performance in classifying various objects,[1], [2] but several journals[3], [4], [5], [6], [7] state that the performance of the algorithm KNN will be much better if combined with a feature selection algorithm such as a genetic algorithm[4]. In research [5], the use of genetic algorithms and KNN in classifying types of disease on oil palm leaves achieved an accuracy level of 100% or it could also be called a perfect level of accuracy[8], [9]. The level of perfect accuracy without the slightest error is worth asking, this result was obtained because the algorithm performance was very good or it could also be due to bias in the data.

Following the Journal, the amount and quality of data can affect the resulting level of accuracy. Therefore, in this research, we want to carry out further analysis regarding the use of genetic algorithms and KNN in classifying 3 classes in three different datasets[10]. These three datasets will be given the same treatment as previous research[7] to see the performance of the algorithm. If the same level of accuracy is obtained even though the datasets used are different, it can be concluded that the algorithm does have a good level of accuracy in classifying various objects and can be used for an artificial intelligence system for classifying objects because it has high accuracy in various types of datasets[11]. However, if the accuracy results obtained have significant differences, there is an influence of the dataset on the performance of the algorithm.

## 2. Literature Review

### 2.1 Dataset

A dataset is a collection of data that is stored to be processed or carried out further processing to analyze, predict, or classify the data. Datasets can be structured data stored in Excel, or can also be unstructured data such as video, audio, text, and images. Before the dataset is used, preprocessing is usually carried out first, such as removing emojis, case folding, tokenizing, stopword removal, and others for text data[1], or changing image size, segmentation, and removing noise for image data[5]. The purpose of preprocessing on a dataset is to correct missing, missing, or noisy data and also to delete data that is not needed [1], [5].

### 2.2 Genetic Algorithms

Genetic algorithms are one algorithm that can be used for feature selection. The Genetic Algorithm will look for the best features from a set of existing features[12]. This algorithm has good performance in any search space [4], [13], [14], [15], [16]. In finding the best features, the genetic algorithm has several stages that must be carried out, namely [6], [7], [17], [18]:

1. Determine the initial population by randomly creating chromosomes
2. Evaluate chromosomes with a fitness function, chromosomes that have a high value have the opportunity to bring new individuals into the population.
3. Carrying out selection using the roulette wheel method aims to find parents who can later produce new offspring. The chromosome chosen to be the parent is determined by probability from the resulting fitness value.
4. Carrying out crossovers and mutations to produce new chromosomes from previously selected parent chromosomes. This process will stop when the algorithm has reached the maximum number of generations and the population does not show very significant changes.

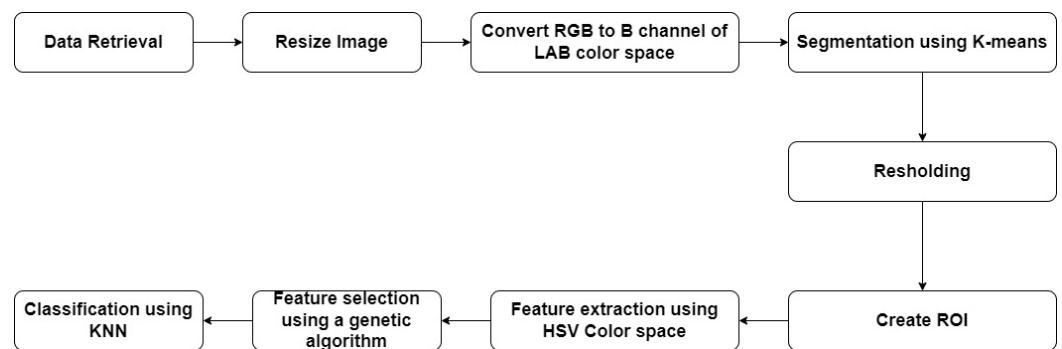
### 2.3 KNN (K-Nearest Neighbors)

KNN is the most widely used classification algorithm because it is simple and easy to apply to the machine learning process and can provide good accuracy results[2], [19]. In carrying out classification, the KNN algorithm [21],[22] will calculate the distance to the nearest neighbors using the K value, where the K value is a positive integer, the data will be grouped based on the value closest to the neighbor [10], [20]. Based on the journal[10], the KNN algorithm formula is as follows equation 1.

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

Based on equation 1, some parameters are explained as follows,  $d(x,y)$  is the distance between testing data and training data,  $x$  is testing data,  $y$  is training data, and  $n$  is the number of training data.

### 3. Method



**Figure 1.** Research flow diagram

In this research, the method used is as in Figure 1, around 8 stages will be carried out. The method used is the same as in the previous journal, only the difference is that feature extraction uses one color space, namely HSV, because based on the journal [5], the HSV color space provides the best level of accuracy. The method used is the same as before, aiming to determine the performance of the algorithm when using different datasets. By treating the dataset the same way, the results obtained will also be the same or different.

#### 3.1 Data retrieval

The data used in this research is secondary data, where the data is taken from open-source data provider websites such as Kaggle. There are 3 datasets used, namely the disease dataset on mango leaves, the disease on corn leaves, and the disease on banana leaves. There are around 3 classes in each dataset and the amount of data in each dataset is 500 data which will later be divided into 80% for training data and 20% for test data.

#### 3.2 Data pre-processing

The data pre-processing carried out in this research is the same as previous research [5] the data used is resized first, then the image is converted to color channel to b channel from the LAB color space, and then image segmentation will be carried out using the k-means clustering algorithm and do the final thresholding to form the ROI.

#### 3.3 Feature extraction

Feature extraction is carried out using the HSV color space because based on previous research [5], the HSV color space provides better results than other color spaces such as RGB, LAB, and HSI. The extraction results in the HSV color space will be selected again to find the best features from the HSV color space.

#### 3.4 Feature selection and classification

The next stage after the data has been extracted will be feature selection first using a Genetic Algorithm. Data with the best features from the feature selection results will later be classified using the KNN (K-Nearest Neighbors) classification algorithm to see the accuracy value of the data.

### 4. Result and Discussion

The results of the research carried out were feature selection using HSV image processing and genetic algorithms for detecting leaf diseases in mango, corn, and potato plants. This research aims to detect leaf diseases with high accuracy. Predictive models that use features selected by the genetic algorithm are tested using cross-validation to ensure that overfitting does not occur. Testing was carried out 10 times to ensure consistency of results. The dataset used in testing consists of more than 1500 test samples for each type of leaf (mango, corn, and potato). The results of this research show that feature selection using HSV image processing and genetic algorithms can achieve

accuracy in getting an average of three types of leaves of 87.9%. Details can be seen in the following Table 1.

**Table 1.** Description of the index value of Leaf

Leaf	1	2	3	4	5	6	7	8	9	10
Mango	87	86	88	86	87	86	89	88	88	87
Corn	82	81	81	81	82	83	82	81	84	81
Potato	95	95	95	94	94	94	95	95	95	94

Extract important features from leaves using image processing in HSV (Hue, Saturation, Value) color space. The HSV color space was chosen because it is closer to human color perception than the RGB color space. A genetic algorithm is used to select a subset of traits that are most relevant for detecting leaf diseases.

Moreover, Genetic algorithms are optimization techniques that mimic the process of biological evolution. In this study, we use a genetic algorithm to find the optimal feature combination that provides the highest prediction accuracy.). The results of the research show that trait selection using a genetic algorithm has succeeded in increasing the accuracy of the prediction model for leaf diseases in mango, corn, and potato leaves. This shows that the genetic algorithm is effective in finding the optimal feature combination. The use of the HSV color space for feature extraction has also proven effective in detecting leaf diseases, as this color space is more relevant to human color perception and disease-related color changes.

5. Conclusion

Feature selection using HSV image processing and genetic algorithms has been proven to improve the performance of leaf disease prediction models with an accuracy of 87.9%. The selected features (H\_mean, S\_mean, V\_mean) provide significant contributions in detecting leaf diseases on mango, corn, and potato leaves. Genetic algorithms are effective in selecting optimal feature combinations, although they require higher computing time. This research shows the great potential of using feature selection and genetic algorithms in the development of more accurate and efficient plant disease detection systems.

**Acknowledgments:** I would like to thank all the academicians at Amikom University Yogyakarta who have helped in completing the manuscript of this article, I hope that this article can be an appropriate reference for the same discipline and be able to make a good contribution to the development of science, especially in the field of Artificial Intelligence.

**Author contributions:** All authors are responsible for building Conceptualization, Methodology, analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision of project administration, funding acquisition, and have read and agreed to the published version of the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. C. Raras, A. Widiawati, and P. Korespondensi, "Pengaruh Dataset Terhadap Performa Convolutional Neural Network Pada Klasifikasi X-Ray Pasien COVID-19", doi: 10.25126/jtiik.202295645.
2. R. A. Saputra, Suharyanto, S. Wasiyanti, D. F. Saefudin, A. Supriyatna, and A. Wibowo, "Rice Leaf Disease Image Classifications Using KNN Based on GLCM Feature Extraction," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1641/1/012080.
3. N. Benayad, Z. Soumaya, B. D. Taoufiq, and A. Abdelkrim, "Features selection by genetic algorithm optimization with a k-nearest neighbor and learning ensemble to predict Parkinson's disease," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1982–1989, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1982-1989.
4. Y. Riwanto, M. T. Nuruzzaman, S. Uyun, and B. Sugiantoro, "Data Search Process Optimization using Brute Force and Genetic Algorithm Hybrid Method," *IJID (International Journal on Informatics for Development)*, vol. 11, no. 2, pp. 222–231, Jan. 2023, doi: 10.14421/ijid.2022.3743.
5. E. P. Atika, A. Sunyoto, and E. T. Luthfi, "Genetic Algorithm and K-Nearest Neighbors for Oil Palm Leaf Disease Classification," in *ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era*, Proceeding, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 447–451. doi: 10.1109/ICOIACT55506.2022.9971854.
6. Yasin KAYA, "Comparison of Using the Genetic Algorithm and Cuckoo Search for Feature Selection," 2018.
7. R. Sehly and M. Mezher, "Performance Impact of Genetic Operators in a Hybrid GA-KNN Algorithm," 2020. [Online]. Available: [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)
8. A. H. Suherman, N. Ibrahim, H. Syahrian, V. P. Rahadi, and M. K. Prayoga, "Klasifikasi Daun Teh Gambung Varietas Assamica Menggunakan Convolutional Neural Network Dengan Arsitektur LENET-5," *Journal of Electrical and System Control Engineering*, vol. 4, no. 2, pp. 63–71, Feb. 2021, doi: 10.31289/jesce.v4i2.4136.
9. D. Pitaloka, "Hortikultura: Potensi, Pengembangan dan Tantangan."
10. S. Napitu, R. Paramita Panjaitan, P. A. Nulhakim, and M. Khalik Lubis, "Klasifikasi Buah Jeruk Segar dan Busuk Berdasarkan RGB dan HSV Menggunakan Metode KNN," *Jurnal SAINTEKOM*, vol. 13, no. 2, pp. 214–221, Sep. 2023, doi: 10.33020/saintekom.v13i2.420.
11. S. Sanjaya, "Aplikasi Pengenalan Tingkat Kematangan Buah Tomat Menggunakan Fitur Warna HSV Berbasis Android," 2022.
12. Y. Sari, M. Alkaff, E. S. Wijaya, S. Soraya, and D. P. Kartikasari, "Optimasi Penjadwalan Mata Kuliah Menggunakan Metode Algoritma Genetika dengan Teknik Tournament Selection," vol. 6, no. 1, pp. 85–92, 2019, doi: 10.25126/jtiik.201961262.
13. I. D. Lesmono and A. Dwi Praba, "Optimasi K-Nearest Neighbour dengan Algoritma Genetika," *AGUSTUS*, no. 2, p. 143, 2017.
14. F. O. Awalullaili, D. Ispriyanti, and T. Widiarihari, "Klasifikasi Penyakit Hipertensi Menggunakan Metode SVM Grid Search dan SVM Genetic Algorithm (GA)," *Jurnal Gaussian*, vol. 11, no. 4, pp. 488–498, Feb. 2023, doi: 10.14710/j.gauss.11.4.488-498.
15. A. B. Hassanat, V. B. S. Prasath, M. A. Abbadi, S. A. Abu-Qdari, and H. Faris, "An improved Genetic Algorithm with a new initialization mechanism based on Regression techniques," *Information (Switzerland)*, vol. 9, no. 7, Jul. 2018, doi: 10.3390/info9070167.
16. Dr. R. S. R. R. N. Vaishali R, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians Diabetes dataset," 2018.
17. Dr. V. A. T.D.Srividya, *Feature Selection Classification of Skin Cancer using Genetic Algorithm*. IEEE, 2018.
18. I. O. Suzanti and F. A. Mufarroha, "Implementasi Relevant Feedback Menggunakan Algoritma Genetika pada Dokumen Bahasa Indonesia Implementation of Relevant Feedback Using Genetic Algorithm in Indonesian Documents," *Jurnal Ilmu Pengetahuan dan Teknologi Komunikasi*, vol. 23, no. 2, 2021.

- 
19. K. Nakata, Y. Ng, D. Miyashita, A. Maki, Y.-C. Lin, and J. Deguchi, "Revisiting a kNN-based Image Classification System with High-capacity Storage," Apr. 2022, [Online]. Available: <http://arxiv.org/abs/2204.01186>
  20. I. N. Simbolon, "Prediksi Kualitas Air Sungai di Jakarta Menggunakan KNN yang Dioptimalisasi dengan PSO," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4191.
  21. I.Kurniawan, P. B.Santoso, "Design of K-Nearest Neighbor Algorithm For Classification of Credit Loan Eligibility At Senarak Dana Purwakarta Cooperative", *Iota*, 2024, ISSN 2774-4353, Vol.04,02.<https://doi.org/10.31763/iota.v4i2.742>
  22. F.Hia, V.Sihombing, A.P.Juledi, "The Relationship of Teacher Activity in the Teaching and Learning Process to Elementary Student Learning Outcomes Using Machine Learning", *Iota*, 2023, ISSN 2774-4353, Vol.03, 04.<https://doi.org/10.31763/iota.v3i4.669>