Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 1, June 2024
ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/js.v7i1.5090
(Page.42-47)

COMPARISON OF PINE SEED QUALITY CLASSIFICATION USING THE NAÏVE BAYES AND KNN METHODS

W. N. Permata[1], Istiadi[2,*], R. P. Putra[3]

[1,2,3]Department of Informatics Engineering, Universitas Widyagama Malang, Malang, Indonesia
**Email:istiadi@widyagama.ac.id*

Submitted : 25 October 2023; *Revision :* 4 June 2024; *Accepted :* 19 June 2024

ABSTRACT

Pine seeds are the seeds of pine trees, which are a type of open-seeded plant known as gymnosperms. Gymnosperm plants have seeds that are not protected by fruit, unlike flowering plants (angiosperms). Pine seeds are typically found inside hard cones. Pine seeds possess several distinctive characteristics, including their small, flat shape and are often equipped with thin wings that aid in their dispersal when released. The process of selecting pine seeds for planting must adhere to established standards of seed quality to enhance desired attributes such as color, texture, and shape in seedlings. Suitable pine seeds for use in planting or propagation are those in a new condition. Quality pine seeds cannot be distinguished by visual inspection alone; alternative tools are required. Given the challenge of differentiating seeds suitable for primary propagation, the researcher proposes a comparison of Pine seed classification using two different methods: the Naïve Bayes Method and K-Nearest Neighbors (KNN). This is expected to enable the accurate detection of pine seeds. The feature extraction method used is the Gray-Level Co-occurrence Matrix (GLCM). The dataset used consists of 165 pine seed samples, comprising 55 images of fresh pine seeds, 55 images of dry pine seeds, and 55 images of decayed pine seeds. Between the two methods, K-NN exhibits the highest percentage value compared to the Naïve Bayes method in the k-fold cross-validation, achieving an accuracy of 95%.

Keywords Pine Seeds, Naïve Bayes; K-Nearest Neighbor; Feature Extraction;
Paper type Research paper

INTRODUCTION

Indonesia is a country with vast natural resources and a wealth of spices. Among the plants found there is the pine tree, particularly pine merkusii, which bears flat, ovoid seeds with fruit scale wings. High-quality seeds have a dry, brownish skin and a firm, smooth, round shape without wrinkles [1]. Pine cultivation is often utilized for both its wood and other parts as required [2]. Pine seeds store carbon, remove organic waste, and improve water quality while reducing pollution in forests [3]. The selection of pinecones must adhere to quality standards to enhance color, texture, and shape characteristics in the nursery[4]. Pinecones that are suitable for planting or nursery purposes are new and of good quality; they cannot be distinguished just by looking at them, an alternative tool is needed [5]. Researchers are expected to offer a straightforward solution for selecting high-quality pine seeds with high accuracy based on the classification of seed quality [6].

Some other research references [7] It compares the prediction of Arabica coffee quality with an accuracy rate of 98%. Furthermore, it uses an algorithmic approach for object accuracy techniques [8]. The *Naïve Bayes* and *K-Nearest Neighbor* methods are compared using K-Fold *Cross Validation* to classify observed objects based on the potential classes of an input. This classification is done by processing clustering and training test data derived from stored image features[9][10]. Next, the research involved image processing of coffee beans using the *K-Nearest Neighbor* method, considering both color and texture. Three categories of bean defects were identified: young beans, moldy beans, and normal beans, with 60 training data samples and 15 test data samples. The data matching accuracy results for various K values are as follows: K=1: 66.67%, K=3: 60%, K=5: 66.67%, and K=7: 66.67%. The overall accuracy stands at 65%.[11]. The classification of coffee bean defects using the Naïve Bayes Classifier method for quality assessment involves identifying defects in coffee beans [12].

The research in [13], Seven feature extraction methods were implemented, including energy, contrast, dissimilarity, homogeneity, correlation, and maximum probability. The training dataset for

ISSN Print : 2621-3745
ISSN Online : 2621-3753
Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 1, June 2024
(Page.42-48)
DOI : 10.31328/js.v7i1.5090

this study consists of 100 images per category, while the test dataset comprises 30 images per category. The accuracy rates for the categories were as follows: early-stage disease 90%, late-stage disease 90%, and non-disease 83.33%[14]. In the research [15] The Naive Bayes method and image feature extraction are employed to classify the ripeness level of Manalagi apples, categorized into unripe and ripe apples. The accuracy of Manalagi apple recognition is 63%, based on 130 images (30 test and 100 training). Using the K-Nearest Neighbor method for classification results in an overall precision of 94%, recall of 100%, and accuracy of 98%, with a dataset comprising 600 training images and 200 images[16]. The K-Nearest Neighbor (KNN) method is a straightforward classification approach that relies on proximity, offering high accuracy [17]. On the other hand, the Naïve Bayes method is effective for merging data [18][19]. The use of K-Nearest Neighbor and Naïve Bayes comparisons increases the level of accuracy [20].

Based on the above description, computational image processing capabilities are needed to compare the classification of pine seed quality. This research aims to use the Naive Bayes and K-Nearest Neighbor (KNN) methods to compare the classification of pine seed quality based on their image classes, in order to determine the best method.

METHOD

The research process during the research phase commences with the preparation and gathering of data from pine seeds as a reference. Subsequently, system design is conducted according to the research problem. Following this, the data is put into action to test the quality classification of pine seeds in the research

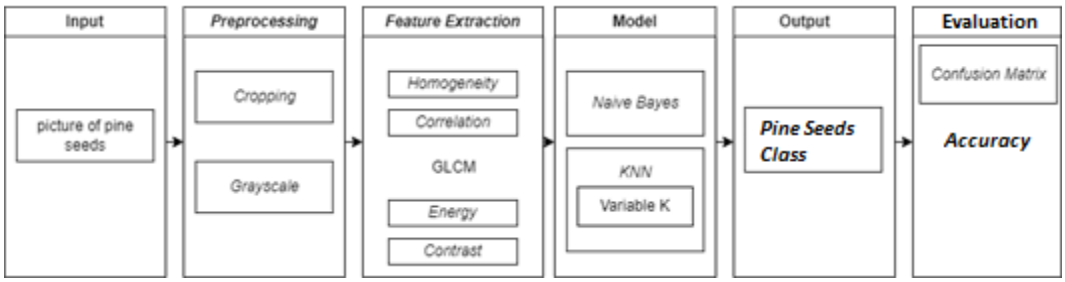


Figure 1. Research Stages

In Figure 1, the research process is depicted, comprising several structured stages. First, input data in the form of images of pine seeds are categorized into different classes for testing. The second stage involves preprocessing, which includes cropping the test data, enhancing pinecone images, reducing noise, and converting them to grayscale. The third stage focuses on feature extraction, where color attributes are assessed using GLCM to identify pine seed texture images and evaluate pine seed quality. For classifying pine seeds, the K-Nearest Neighbor and Naïve Bayes methods are employed. The final stage is evaluation, which entails assessing the classification process, making necessary revisions, and improving the system.

Pine seed image data

The researchers employed three categories to assess the classification of pine seeds using the Naïve Bayes and K-NN algorithms. Each category includes 55 data points for wet pine seeds, 55 for rotten pine seeds, and 55 for dry seeds, totaling 165 data points. Regarding the category of pine seeds suitable for planting, there are three categories of pine seeds as follows.

Table I is an example of an image of a pinecone, which is one of the most common and significant plant species in forest ecosystems. Pine trees are found all over the world and constitute some of the most extensive forests globally. Pines play a crucial role in preserving the natural balance of forests. In the example table image, there are three categories of pinecones that are suitable for seeding [3].

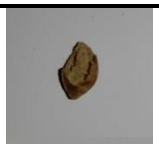
Pinecone seed collection

Data collection involves capturing images. In addition, the steps in data collection include designing tools for the process of capturing seed images in a way that prevents light from entering, both during

Journal of Science and Applied Engineering (JSAE)
Vol. 7, No. 1, June 2024
ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/js.v7i1.5090
(Page.42-47)

the day and at night. After that, selecting an interesting shooting angle, whether it's from a high, low, or side perspective, to offer a unique viewpoint.

TABLE I. EXAMPLE PICTURE

Class	Example Seeds		
Rotten			
Wet			
Dry			

Preprocessing

At this point, two stages are performed: image size normalization and image conversion. The image normalization process is executed to establish a specific size by resizing the image and cropping it at the pine seed image coordinates. Color image conversion is carried out to simplify the feature extraction process because the RGB color space has a wide color spectrum, making it challenging to convert to other color spaces, such as grayscale. This conversion turns the image into grayscale using a specific equation (1).

$$Grayscale = \frac{R+G+B}{3} \quad (1)$$

Where R = red component image, G = green component image, and B = blue component image.

TABLE II. *GRAYSCALE*

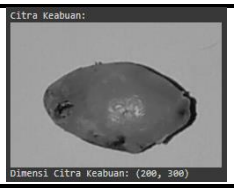
	Original Image	***Grayscale Image***
Image of Pine Seeds		

Table II presents the stage of converting color images (in RGB format) into grayscale images. The subsequent step is the feature extraction stage, during which the data transitions from image files into numerical data, essential for the Google Colaboratory application process to analyze texture. This information will later be employed for classification comparison. Below, you will find the results of feature extraction for characterizing texture in images of pine seeds

Feature extraction

Feature extraction is a process designed to capture distinctive characteristics specific to the research subject. In this reaserch, texture features are employed. The texture features from the GLCM (Gray-Level Co-occurrence Matrix) used include contrast, homogeneity, energy, and correlation at each angle of 0°, 45°, 90°, and 135°, computed using specific formulas in (2)(3)(4)(5).

$$Contrast = \sum_k k^2 \left[\sum_i \sum_j p\,(i,j)\right] \quad (2)$$

$$Corelation = \sum i,j \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma i \sigma j} \quad (3)$$

$$Energy = \sum_{i,j} P(i,j)^2 \quad (4)$$

$$Homogeneity = \sum_{i,j} \frac{p(i,j)}{1+[i-j)} \quad (5)$$

Naïve Bayes Classification

Naive Bayes is a probabilistic classification method that relies on the assumption of independence among input variables. It leverages data and probability theory to estimate the likelihood of various classes for a given input. The Naive Bayes classifier uses probabilities to estimate the probability of an input belonging to different classes. This is achieved using the equation (6).

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(Xi|Y)}{P(X)} \quad (6)$$

When it comes to X and Y, simplifying it, Naive Bayes assumes that the value of a feature is unrelated to the presence of other features once the class variable has been determined. It assumes that each feature has an independent chance of contributing, irrespective of the presence or absence of other features.

K-Nearest Neighbor Classification

K-Nearest Neighbors (K-NN) is a classification method that employs a voting approach to forecast the output class based on input data. This method relies on an algorithm that computes the distance between a data point and its neighboring data points. The algorithm utilizes this distance as a parameter for predicting the output class of the input data. K-NN identifies the K nearest data points to the input data point and subsequently tallies the occurrences of different classes among these data points. We use a specific formula (7) to find the class with the highest frequency, and this class becomes the predicted output.

$$d(Xi,Xj) = \sqrt{\sum_{l=1}^{N}(diff(XiI,XjI))^2} \quad (7)$$

Where X1, l = 1, 2, represents the category attribute, and n1j, n1 represents the corresponding frequency. Based on the category that appears most frequently in K-Nearest Neighbor, it will generate a class prediction for new data. In general, the best k-value weights are relative to the available data.

Training

Training is needed to form a classification model according to the algorithm used, namely Naive Bayes and KNN. The training process uses a certain amount of data so that the model is able to recognize data patterns. The resulting model will then be used for testing.

Testing

Testing involves processing image extraction features, followed by calculating similarity comparisons with these features. The aim of data testing is to improve data comprehension and facilitate better decision-making. This study compares two classification algorithms, namely, K-Nearest Neighbor and Naive Bayes Classifier, and employs k-fold cross-validation for data validation.

Evaluation

In the evaluation stage, the measure or parameter used is the confusion matrix. The confusion matrix is a performance measurement for machine learning classification problems where the output can consist of two or more classes. Based on the confusion matrix, the accuracy calculation of equation (8) can be derived.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN)} \quad (8)$$

ISSN Print : 2621-3745
ISSN Online : 2621-3753
DOI : 10.31328/js.v7i1.5090
(Page.42-47)

DISCUSSION

The data used for classification is obtained from KPH Kedu Utara for gathering both training and testing data. Data collection involves taking photographs. There are 55 data points for each class, including wet pine seeds, rotten pine seeds, and dry seeds, making a total of 165 data points. Researchers employ four feature extraction methods: homogeneity, correlation, energy, and contrast for classifying pine seeds, as these four features are sufficient for distinguishing suitable pine seeds for planting. Here, researchers present the results of GLCM feature extraction in numerical form, as shown in the Table III.

TABLE III. EXAMPLE FEATURE EXTRACTION DATA

(Project_ID)	*Features Extraction*				
	Contrast	*Energy*	*Homogeneity*	*Correlation*	*Result*
IMG_20230604_020144.jpg	0.533053	0.482212	0.209375	0.206090	Rotten
IMG_20230704_231634.jpg	0.516026	0.520793	0.397436	0.275721	Wet
IMG_20230604_020144.jpg	0.536538	0.563301	0.230769	0.293269	Dry

Table III displays the outcomes of converting various image files into numerical data, as features extraction, for the purpose of conducting texture analysis using the Naïve Bayes and KNN methods. The subsequent table presents the results of texture characteristics derived from GLCM (Gray-Level Co-occurrence Matrix). Below, the results of feature extraction used for characterizing the texture in images of pine seeds.

TABLE IV. ACCURACY RATES BASED ON CROSS-VALIDATION

(Cross Validation)	**CLASIFICATION METHODS**					
	KNN 1	*KNN 2*	*KNN 3*	*KNN 4*	*KNN 5*	*Naïve Bayes*
Kfold 1	96%	83%	79%	94%	93%	62%
Kfold 2	93%	93%	89%	83%	86%	92%
Kfold 3	100%	83%	78%	94%	83%	85%
Kfold 4	93%	86%	82%	100%	78%	75%
Kfold 5	93%	78%	82%	75%	93%	85%
Average Accuracy	95%	85%	82%	89%	87%	80%
Best Accuracy			95%			

Based on Table IV, the results of the cross-validation process are explained. Two types of models were used: K-Nearest Neighbors (KNN) with various K values (KNN 1, KNN 2, KNN 3, KNN 4, KNN 5) and the Naïve Bayes model. Cross-validation is a process in which the data is divided into 5 folds to objectively measure the model's performance. In this case, K-Fold Cross Validation was employed with five folds (Kfold 1 to Kfold 5). Each cell in the table represents the accuracy of the model in that particular fold. For instance, in Kfold 1, KNN 1 achieved an accuracy of 96%, indicating that the KNN model with K=1 made correct predictions for 96% of the data in that fold. The highest accuracy achieved among all the models in this cross-validation is 95%, which was attained by KNN 1.

CONCLUSION

Based on the results of research on pine seed quality using the KNN and Naïve Bayes methods, it can be concluded that the conducted research has yielded a fairly effective pine seed detection process when using the K-NN method. In the dataset created for K-NN classification, the researcher observed that the research results in the classification of pine seeds into their respective classes are quite satisfactory. The accuracy of pine seed recognition greatly depends on the quality of pine seed classification, aiding in the identification of wet, dry, or rotten pine seeds. The feature extraction process using the GLCM method has proven to be highly beneficial for classification, with the value of pine seed texture feature extraction being contingent on image quality. In the classification, it can be concluded that among the two methods, K-NN is the method that exhibits the highest percentage

value when compared to the Naïve Bayes method in the k-fold testing, achieving an accuracy of 95%.

REFERENCES

[1] M. F. Remeli, N. E. Bakaruddin, S. Shawal, H. Husin, M. F. Othman, and B. Singh, "Experimental study of a mini cooler by using Peltier thermoelectric cell," IOP Publishing, 2020, p. 12076.
[2] Aurelia, "Journal of Fashion & Textile Design Unesa," vol. 1, pp. 128–137, 2020.
[3] S. Rizqa Fethiananda and M. Sigit Ramadhan, "Application of Block Printing Techniques Using the Direct Print Method with Merkusii Pine Inspiration on Textile Materials," *Agustus*, vol. 7, no. 2, p. 3091, 2020.
[4] P. U. Rakhmawati, Y. M. Pranoto, and E. Setyati, "Classification of Potato Leaf Diseases Based on Texture Features and Color Features Using Support Vector Machine," *Semin. Nas. Technol. and Engineering*, pp. 1–8, 2018.
[5] E. J. Hantono, "Utilization of Pine Flowers in Making Particle Board as a Material in Making Craft Products (Case Study: Pine Forest in Karangpucung, Cilacap, Central Java) Utilization of Fine Flowers in the Manufacture of Particle Board.," *Art Des.*, vol. 3, no. 3, pp. 1–7, 2016.
[6] M. Safrizal and A. Harjoko, "Comparison of Grayscale Image Coloring Using K-Means Clustering and Agglomerative Hierarchical Clustering Methods," *Bimipa*, vol. 23, no. 3, pp. 255–263, 2013.
[7] V. Sari, F. Firdausi, and Y. Azhar, "Comparison of Arabica Coffee Quality Predictions Using the SGD, Random Forest and Naive Bayes Algorithms," *Edumatic J. Educator. Inform.*, vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.
[8] Muhammad Romzi and B. Kurniawan, "Learning Python Programming Using an Algorithmic Logic Approach," *JTIM J. Tek. Inform. Masterpieces*, vol. 3, no. 2, pp. 37–44, 2020.
[9] F. Tempola, M. Muhammad, and A. Khairan, "Comparison of Classification Between KNN and Naive Bayes in Determining Volcano Status with K-Fold Cross Validation," *J. Teknol. Inf. and Computer Science.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
[10] Karsito and S. Susanti, "Classification of Eligibility of Home Credit Applicants Using the Naïve Bayes Algorithm at Azzura Residencia Housing," *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.
[11] M. A. Rizal, "Coffee Bean Quality Classification Using the K – Nearest Neighbor Method Based on Color and Texture," *Univ. Technol. Yogyakarta*, vol. 1, no. 1, pp. 1–8, 2019.
[12] J. Aramiko, "Using the Naïve Bayes Method Thesis by : Juni Aramiko," 2020.
[13] I. G. R. A. Sugiartha, M. Sudarma, and I. M. O. Widyantara, "Color, Texture and Shape Feature Extraction for Clustered-Based Retrieval of Images (CLUE)," *Maj. Ilm. Technol. Electro*, vol. 16, no. 1, p. 85, 2016, doi: 10.24843/mite.1601.12.
[14] R. A. Asmara, B. S. Andjani, U. D. Rosiani, and P. Choirina, "Gender Classification in Facial Images Using the Naive Bayes Method," *J. Inform. Polynema*, vol. 4, no. 3, p. 212, 2018, doi: 10.33795/jip.v4i3.209.
[15] A. Ciputra, D. R. I. M. Setiadi, E. H. Rachmawanto, and A. Susanto, "Classification of Manalagi Apple Ripeness Levels Using the Naive Bayes Algorithm and Digital Image Feature Extraction," *Simetri J. Tek. Mechanical, Electrical and Computer Science.*, vol. 9, no. 1, pp. 465–472, 2018, doi: 10.24176/simet.v9i1.2000.
[16] N. Wijaya and A. Ridwan, "Classification of Apple Types Using the K-Nearest Neighbors Method," *J. SISFOKOM*, vol. 8, no. 1, pp. 74–78, 2019.
[17] M. R. Alghifari and A. P. Wibowo, "Application of the K-Nearest Neighbor Method for Web-Based Security Guard Performance Classification," *J. Teknol. dan Manaj. Inform.*, vol. 5, no. 1, 2019, doi: 10.26905/jtmi.v5i1.3074.
[18] W. Muslehatin and M. Ibnu, "Application of Naïve Bayes Classification to Classify the Possible Level of Obesity for Information Systems Students at UIN Suska Riau," *Semin. Nas. Technol. Information, Commun. and Ind.* , pp. 2579–5406, 2017.
[19] H. Muhamad *et al.*, "Optimization of Naïve Bayes Classifier Using Particles," *J. Teknol. Inf. and Computer Science.* , vol. 4, no. 3, pp. 180–184, 2017
[20] D. S. Wisdayani, "Comparison of the K-Nearest Neighbor and Naive Bayes Algorithms for Classifying the Severity Level of Traffic Accident Victims in Pati Regency, Central Java," *J. Dr. Diss. Muhammadiyah Univ. Semarang , pp. 9–25, 2019.*