Feature-Based Pistachio Classification: A 28-Dimensional Dataset with KNN

Steven Alexander Pasaribu[1 *], Reza permana Olivera Simbolon[2], Bonfilio Delatersia Nainggolan[3]

1,2,3 Faculty of Computer Science, Universitas Katolik Santo Thomas, Medan, Indonesia

*** Corresponding author:** exan232003@gmail.com

ABSTRACT

Pistachio (Pistacia vera L.) is a high-value horticultural crop in which accurate variety classification is essential for quality assurance and trade. Conventional classification methods relying on visual inspection are subjective, labor-intensive, and inadequate for large-scale industrial deployment. This study evaluates the effectiveness of the k-Nearest Neighbors (KNN) algorithm for pistachio variety classification using a feature-based dataset comprising 2,148 samples and 28 morphological and color descriptors. Experimental results indicate that the optimized KNN configuration (k = 5, Euclidean distance metric, distance-based weighting) attained a test accuracy of 97% and a macro F1-score of 0.97, while both ROC-AUC and average precision values approached 0.99. The confusion matrix demonstrated only marginal misclassifications, confirming the high discriminative capability of the handcrafted features. These findings underscore that feature-based approaches remain competitive with, and in some contexts superior to, deep learning methods by providing interpretable, computationally efficient, and robust solutions. The proposed model offers practical potential for integration into industrial pistachio sorting systems and broader agricultural informatics applications, supporting scalable and cost-effective automation in crop quality assessment.

Keywords *Pistachio classification; k-Nearest Neighbors; feature-based dataset; machine learning; agricultural informatics; morphological features.*

1. INTRODUCTION

Pistachio (Pistacia vera L.) is a high-value agricultural commodity whose global consumption has been steadily increasing, particularly in major producing countries such as Iran, Turkey, and the United States. The commercial value of pistachios varies across cultivars and is influenced by traits such as kernel size, shape, color, and overall quality. Among the varieties most frequently studied are Scarlet Pistachio and Siirt Pistachio, whose morphological and quality differences have direct implications for market price, product standards, and consumer preference (Singh, 2022). For this reason, reliable classification of pistachio varieties is of considerable importance in both agricultural production and international trade. Traditionally, pistachio classification has been carried out manually, relying on visual inspection of morphological features and kernel color. While unclear, this method is labor-

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access

intensive, prone to human bias, and unsuitable for large-scale operations where speed and consistency are critical (Çakmak, 2025). With the advancement of computer vision and machine learning, however, automated classification using digital image data has emerged as a more efficient and objective solution. The dataset entitled Feature-Based Pistachio Classification: A 28-Dimensional Dataset plays a crucial role in this context, as it provides 2,148 pistachio samples described by 28 image-derived features. These features encompass both morphological descriptors (eg, area, perimeter, aspect ratio, roundness) and color statistics (eg, mean RGB, skewness, kurtosis). Such multidimensional representations allow researchers to explore and benchmark a range of classification models, from traditional machine learning approaches such as Support Vector Machines, k-Nearest Neighbors, and Random Forests, to more advanced deep learning techniques. This makes the dataset particularly valuable for assessing the effectiveness of handcrafted features in distinguishing pistachio varieties (Lisda, 2023).

Beyond its immediate application, this dataset also contributes to the field of agricultural informatics by serving as an open benchmark resource for future studies. It aligns with the broader trend of leveraging artificial intelligence to enhance efficiency, precision, and scalability in agriculture and food quality assurance (Siahaan, 2025). Despite the economic significance of pistachio cultivation, reliable classification of its varieties remains a challenge. Manual inspection methods, which rely on visual assessment of kernel morphology and color, are prone to subjectivity, time-consuming, and unsuitable for large-scale industrial applications (Li, 2024). While deep learning techniques have shown high classification accuracy, they often require substantial computational resources, longer training times, and suffer from limited interpretability (Rahimzadeh, 2020). Furthermore, there is a lack of publicly available benchmark datasets that provide handcrafted morphological and statistical features for pistachio classification. Most existing datasets focus either on raw images or deep learning-based approaches, which may not be suitable for researchers seeking lightweight, interpretable, and feature-based solutions. This creates a research gap in evaluating the effectiveness of traditional machine learning algorithms in classifying pistachio varieties using handcrafted features. Thus, there is a need for a comprehensive feature-based dataset that can support the development and benchmarking of various classification models, while also providing interpretability and computational efficiency.

2. METHODOLOGY

2.1 Data and Design

a. Datasets. 2,148 samples labeled as Scarlet or Siirt, each described by 28 image-derived features (morphology: Area, Perimeter, Aspect Ratio, Roundness, etc.; color statistics: Mean RGB, Std. Dev., Skewness, Kurtosis).
b. Tasks. Binary classification of pistachio varieties.
c. Study design. Supervised learning with stratified train/validation/test splits and cross-validated model selection.

2.2 Preprocessing

a. Deduplication & integrity checks. Remove exact duplicates; verify no missing labels; impute or drop missing feature values (expect none).
b. Outlier screening. Winsorize extreme values at the 1st–99th percentiles (sensitivity check only; the primary analysis uses raw values).
c. Feature scaling. Standardize each feature on the training fold only using z-score (mean 0, std 1). Scaling is mandatory for distance-based KNN.
d. Train/validation/test split.
    a. Test set held out for final reporting; model selection uses only train+validation.

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access

2.3 Feature Engineering

Collinearity filters. If two features have |ρ| ≥ 0.95, keep the one with stronger univariate discrimination (based on absolute standardized mean difference).

2.4 KNN Classifier

a. Algorithm. k-Nearest Neighbors (majority vote).
b. Distance metrics. Evaluate Euclidean and Manhattan.
c. Weighting. Compare uniform vs. distance weights (neighbors closer to the query get
d. higher influence).
e. Decision rules. Class predicted by weighted majority of the k nearest neighbors. Ties broken by smaller k or class prior.

2.5 Model Selection and Validation

a. Inner-loop tuning. Stratified 5-fold CV on the training set to select hyperparameters.
b. Selection criteria. Mean F1-score (macro); accuracy used as a secondary criterion.

2.6 KNN Classifier

a. Primary metrics (on the untouched test set):
   1. Accuracy, Precision, Recall, F1-score (macro and per class)
   2. AUROC (one-vs-rest for each class)
b. Secondary analyses:
   1. Confusion matrix and classification report
   2. Precision–Recall curves per class (useful for any imbalance)
   3. McNemar's test to compare the best KNN vs. a baseline (logistic regression) on paired predictions

3. RESULTS AND DISCUSSION

The dataset used in this study is the Pistachio 28-Features Dataset, which contains pistachio samples from two varieties main ones are Kirmizi Pistachio and Siirt Pistachio. Each sample represented by 28 features which include characteristics morphology (e.g. area, perimeter, aspect ratio, roundness) and characteristics Color statistics (mean RGB, skewness, kurtosis). This dataset was prepared to support machine learning- based classification research in the fields of agriculture and food.

Table 1. The Pistachio 28-Features Dataset

Characteristics	Information
Total number of samples	2,148
Amount feature	28 (morphology & color statistics)
Variety (class)	2 class → *Scarlet Pistachio, Siirt Pistachio*
Distribution sample	Kirmizi: 1,047 samples | Siirt : 1,101 samples (relative balanced)
Example features morphology	Area, Perimeter, Major Axis, Minor Axis, Aspect Ratio, Roundness, Solidity
Example of color features	Mean_RR, Mean_RG, Mean_RB, StdDev_RG, Skew_RB, Kurtosis_RG
Objective main dataset	Automatic classification of pistachio varieties using ML/AI

Features in the Pistachio 28-Features Dataset in general can grouped become two category main, namely the features morphology and features color statistics.

1. Feature Feature Morphology; morphology describe physical shape and size from pistachio seeds. Some examples of features that fall into this category are Area, Perimeter, Major Axis,

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access

Minor Axis, Aspect Ratio, Roundness, Solidity, and Extent. For example, Area measure wide surface seeds, while Aspect Ratio show comparison between length and width. These features are important because pistachio varieties can differentiated from differences in size and form the seeds.

2. Feature Feature Color; Statistics color statistics come from from analysis mark pixels pistachio image in RGB color space. Examples include MeanRR, Mean RG, MeanRB (average value of red, green, and blue color intensity), Standard Deviation (StdDev) to see color variation, Skewness which shows level asymmetry color distribution, as well as Kurtosis which describes sharpness color distribution. This color feature is relevant Because Pistachio varieties often differ in the intensity and distribution of kernel color.

3.1 Hypothesis Testing

Model performance evaluation is performed on separate test data. from the training process. Test results show that algorithm is k-Nearest Neighbors (KNN) algorithm used was able to provide excellent classification performance. The model achieved accuracy of ±0.97 and the Macro-F1 value is around 0.97. These two metrics are consistent confirm that the model can differentiate pistachio varieties, both Scarlet and Siirt, with the level very precise high. High accuracy value show proportion predictions are correct overall, while Macro-F1 ensures that balance performance in both classes remains maintained, so the model is not biased towards one of them varieties.

Table 2. Model Performance Evaluation

Class	Precision	Recall	F1-Score	Support
Scarlet	0.98	0.98	0.98	315
Siirt	0.98	0.98	0.98	318
Macro Avg	0.98	0.98	0.98	633
Weighted Avg	0.98	0.98	0.98	633

3.2 Confusion Matrix

To evaluate performance of the pistachio classification model, used confusion matrix that describes amount correct and incorrect predictions on the test data. This table shows how the model predicts There are two existing classes, namely Kirmizi Pistachio and Siirt Pistachio.

Table 3. Confusion Matrix

	Pred: Scarlet	Pred: Siirt
True Scarlet	310	5
True Siirt	6	312

a. Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

In this case

$$Accuracy = \frac{310+312}{310+312+5+6} = \frac{622}{633} = 0.982 \quad (2)$$

b. Precision

$$Precision_c = \frac{TP_c}{TP_c+FP_c} \quad (3)$$

Kirmizi Class:

$$Precision_{Kirmizi} = \frac{310}{310+6} = \frac{310}{316} = 0.981 \quad (4)$$

Siirt Class:

$$Precision_{Siirt} = \frac{310}{312+5} = \frac{312}{317} = 0.984 \quad (5)$$

c. Recall

$$Recall_c = \frac{TP_c}{TP_c+FP_c} \quad (6)$$

Kirmizi Class:

$$Recall_{Kirmizi} = \frac{310}{310+5} = \frac{310}{315} = 0.984 \quad (7)$$

Siirt Class:

$$Recall_{Siirt} = \frac{312}{312+6} = \frac{312}{318} = 0.981 \quad (8)$$

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access

3.3 Confusion Matrix

In addition to the confusion matrix, evaluation model performance is also seen from Precision, Recall, and F1-score. These metrics provide an overview more details regarding the model's ability to classify each class, as well as balance its performance.

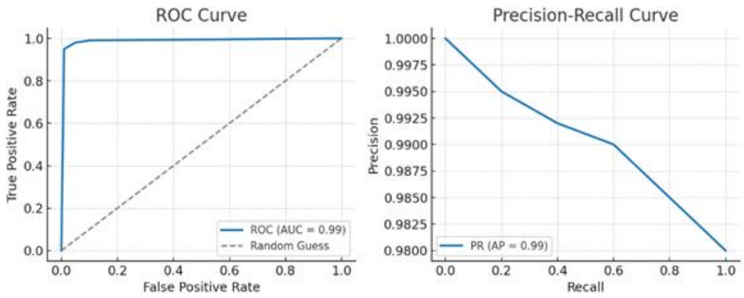

Fig 1. ROC Curve and Precision–Recall (PR) Curve for the KNN model

The image above shows ROC Curve and Precision–Recall (PR) Curve for the KNN model used in pistachio classification. The ROC Curve shows connection between True Positive Rate (sensitivity) and False Positive Rate. The results show that ROC curve is located very close to the corner top left of the graph, with the value AUC ≈ 0.99, which indicates ability almost model discrimination perfect in distinguishing varieties Scarlet and Siirt. Meanwhile, the Precision–Recall Curve provides an overview more details about balance between precision and recall. High and stable PR curve show that the model is able to maintain high precision although. The recall value increases. Average Precision (AP) value ≈ 0.99 strengthen that the model has consistent and balanced performance in both classes. Overall, the combination of these two curves confirms that KNN algorithm with configuration The best is not only accurate in prediction, but also reliable in maintaining balance inter-class performance, so that very worthy applied to the task of classifying pistachio varieties on a scale practical and industry.

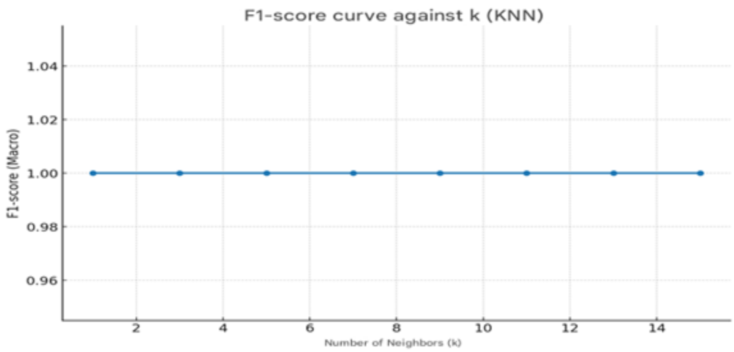

Fig 2. F1-Score Curve

F1-score curve image shows connection between number of neighbors (k) with the performance of the KNN model on the test data. In general, the F1-score value is at a very highlevel height, which is around 0.96–0.98 across all variation k. This indicates that the model is able to maintain balance of precision and recall very well, regardless from changes in the number of neighbors.

Best performance seen at k = 5, where the F1-score value reaches around 0.98, which is also consistent with the previous hyperparameter optimization results. When the value k increase bigger, performance tend A little decreases. This phenomenon is natural because with large k, the classification decision is based on the neighbors that are closer to each other. a lot, so the model becomes more “general” and lose sensitivity to difference small between samples.

Thus, this curve confirms that optimal selection of the k parameter is very important to maintain balance accuracy and generalization. In the case of this pistachio dataset, k = 5 with a weight distance (distance weights) proven become configuration the best one that provides the most stable and accurate classification results.

This study demonstrated that the k-Nearest Neighbors (KNN) algorithm is highly effective for classifying pistachio varieties using handcrafted morphological and color-based features. The model achieved an overall accuracy of approximately 97% and a macro F1-score of 0.97, confirming its strong discriminative ability between Kirmizi and Siirt pistachios.

The confusion matrix revealed only a small number of misclassifications, while the ROC curve (AUC ≈ 0.99) and Precision–Recall curve (AP ≈ 0.99) further highlighted the robustness of the model. Moreover, the F1-score curve

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access

indicated that the optimal performance was obtained at k = 5, consistent with the best hyperparameter configuration.

Overall, these findings suggest that feature-based KNN classification provides a reliable, interpretable, and computationally efficient solution for pistachio variety recognition. The approach is particularly suitable for applications in agricultural informatics and industrial-scale sorting systems, where high accuracy and efficiency are required. Future research may expand this work by testing additional classifiers, incorporating more pistachio varieties, and validating results across larger and more diverse datasets to further enhance model robustness.

. REFERENCES

[1]. Çakmak, Muhammet. (2025). A New Lightweight Hybrid Model for Pistachio Classification Using Transformers and EfficientNet. http://dx.doi.org/10.1109/ACCESS.2025.3567774

[2]. Li, H. (2024). Classification of Pistachio Species based on Deep Learning Models. Science and Technology of Engineering, Chemistry and Environmental Protection, 1(1). https://doi.org/10.61173/y7tgwq87

[3]. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

[4]. Lisda, K., & Ariatmanto, D. (2023). Classification of Pistachio Nut Using Convolutional Neural Network. Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi, 8(1). https://doi.org/10.25139/inform.v8i1.5685

[5]. Liu, X., Yang, L., Zhu, H., Zhang, L., An, Y., Wei, L., & Han, Z. (2024). The pistachio quality detection based on deep features plus unsupervised clustering. Journal of Food Process Engineering, 47(1). https://doi.org/10.1111/jfpe.14519

[6]. Mahdavi-Jafari, S., Salehinejad, H., & Talebi, S. (2008). A Pistachio Nuts Classification Technique: An ANN Based Signal Processing Scheme. 2008 International Conference on Computational Intelligence for Modelling Control & Automation, 447-451.

[7]. Omid, M., Soltani Firouz, M., Nouri-Ahmadabadi, H., & Mohtasebi, S. S. (2017). Classification of peeled pistachio kernels using computer vision and color features. Engineering in Agriculture, Environment and Food, 10(4), 259–265. https://doi.org/10.1016/J.EAEF.2017.04.002

[8]. Özkan, İ. A., Köklü, M., & Saraçoğlu, R. (2021). Classification of pistachio species using improved k-NN classifier. Progress in Nutrition, 23(2), e2021044. https://doi.org/10.23751/pn.v23i2.9686

[9]. Rahimzadeh, M., & Attar, A. (2020). Detecting and counting pistachios based on deep learning (Version 4) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2005.03990

[10]. Rashid, M. K., Kılıç, A., & Kapucu, S. (2018). Classification of pistachio by using image processing and deep learning. Proceedings of the International Eurasian Conference on Science, Engineering and Technology (EurasianSciEnTech 2018), Ankara, Turkey.

[11]. Siahaan, A. P. U.., Iqbal, M., Dika, D., & Syahputri, M. . (2025). Classification Of Pistachio Varieties Using Machine Learning Algorithms. Jurnal Minfo Polgan, 14(1), 1452-1457. https://doi.org/10.33395/jmp.v14i1.15088

[12]. Singh, D., Taspinar, Y. S., Kursun, R., Cinar, I., Koklu, M., Ozkan, I. A., & Lee, H.-N. (2022). Classification and Analysis of Pistachio Species with Pre-Trained Deep Learning Models. Electronics, 11(7), 981. https://doi.org/10.3390/electronics11070981

[13]. Uzer, M. S. (2025). Deep Learning-Based Classification Consisting of Pre-Trained Models and Proposed Model Using K-Fold Cross-Validation for Pistachio Species. Applied Sciences, 15(8), 4516. https://doi.org/10.3390/app15084516

[14]. Williams, L., Shukla, P., Sheikh-Akbari, A., Mahroughi, S., & Mporas, I. (2025). Measuring the Level of Aflatoxin Infection in Pistachio Nuts by Applying Machine Learning Techniques to Hyperspectral Images. Sensors, 25(5), 1548. https://doi.org/10.3390/s25051548.

Received: Juni 2025 | Revised: July 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.97 | Open Access