ISSN 2959-6157

Machine Learning in Agriculture: KNN-Based Classification of Fruits and Vegetables

Siyang Yao

Computer Science, University of Iowa, Iowa City, The United States
syao12@uiowa.edu

Abstract:

In modern agriculture and the food industry, it is essential to classify fruits and vegetables accurately and efficiently to meet growing consumer demand and reduce post-harvest losses. Traditional manual methods are often labor-intensive and error prone, highlighting the need for automated solutions. This paper discusses the application of the k Nearest Neighbor (KNN) algorithm to fruit and vegetable recognition using image processing technology. In this study, image data sets are used for feature extraction using Directional Gradient Histogram (HOG) and dimensionality reduction using Principal Component Analysis (PCA). The accuracy of KNN model on verification set and test set reaches 97%, which proves its validity. Confusion matrix analysis and F1 score evaluation further revealed the strengths and areas of improvement of the model, particularly in distinguishing visually similar categories. The results show that the integration of artificial intelligence (especially KNN) offers great potential for the automation of agricultural classification tasks. Future studies could combine more advanced models, such as CNNS, with larger datasets to improve accuracy and robustness.

Keywords: K Nearest Neighbor (KNN); machine learning; image recognition

1. Introduction

In modern agriculture and the food industry, the importance of accurately and efficiently classifying fruits and vegetables cannot be overstated. As consumer demand for high-quality products continues to grow, so does the need for technology that can ensure these standards are always met. Traditional classification methods rely heavily on manual inspection, which is not only time-consuming, but also prone to human error and subjectivity. This inefficiency often leads to significant post-harvest losses and quality control issues. Therefore, there is an urgent need to adopt advanced technologies that can simplify this process and improve accuracy.

With the advent of Artificial Intelligence (AI), machine learning algorithms have revolutionized agricultural image recognition, enabling accurate classification of fruits and vegetables [1]. The K-nearest Neighbor (KNN) algorithm [2], known for its simplicity and effectiveness, is widely used for such tasks. However, other algorithms like decision trees and neural networks also play significant roles. Decision trees provide interpretable models by segmenting data based on key features, while neural networks, particularly Convolutional Neural Networks (CNNs), excel at recognizing complex patterns in images. These diverse algorithms collectively enhance the efficiency and accuracy of AI-driven agricultural systems. Among these algorithms, the KNN model stands out for its simplicity and effectiveness in classification tasks. KNN is an instance-based learning algorithm that classifies samples according to the majority votes of the nearest neighbors in the feature space. This approach is particularly beneficial for tasks involving small and medium-sized data sets, where it can provide powerful performance without the need for large computational resources or complex training processes. In the field of fruit and vegetable recognition, the integration of artificial intelligence and KNN models has shown great potential. By harnessing the power of image processing technology, AI can extract and analyze complex features from images, such as color, texture, and shape. These characteristics are crucial for distinguishing between different types of agricultural products. The KNN algorithm then uses these features to classify the images, making the process both efficient and highly accurate. Yalcin and Razavi et al. [3] have shown that combining image processing with KNN can significantly improve classification accuracy. Similarly, Shaik et al. [4] highlight the practical applications of these technologies in agriculture, highlighting their role in improving operational efficiency and accuracy.

The integration of AI in fruit and vegetable recognition is a major advance in agricultural technology, which can reduce human error, improve efficiency and ensure higher standards of agricultural product quality by automating the classification process. This paper will deeply discuss the specific application and technical realization of KNN in this field and provide valuable insights and references for future research and development in this field.

2. Methods

2.1 Dataset Preparation

The dataset utilized in this study originates from Kaggle and was meticulously divided into three primary sections [5]: the training set, validation set, and test set, to ensure a comprehensive approach to model training and evaluation. The training set, consisting of 100 images per category, was used to fit the model and enable it to learn underlying patterns within the data. The validation set, containing 10 images per category, played a crucial role in fine-tuning the hyperparameters, optimizing the model's performance without risking overfitting. The test set, also comprising 10 images per category, was reserved for the final evaluation, providing an unbiased assessment of the model's generalization capability. Each image in the dataset is a 200×200 pixel grayscale image, converted from its original RGB format during pre-processing to simplify the data and focus on the most relevant features for classification. The dataset includes a diverse range of categories, each representing a distinct type of fruit or vegetable, ensuring that the KNN algorithm's robustness and accuracy are thoroughly tested across different types of objects.

The preparation of the data set involves several pre-processing steps. First, the image is resized to 100x100 pixels to ensure consistency across the data set. Each image is then converted from RGB to grayscale to reduce the complexity of the data and focus on essential features. The image is then normalized by scaling the pixel values to a range between 0 and 1, which helps improve the performance of the machine learning model by standardizing the input data.

Data enhancement techniques have also been employed to artificially increase the size of the training set to provide the model with more diverse data and improve its generalization ability. These techniques include random rotation, flipping, and shifting of images.

2.2 KNN-based Prediction

The K-nearest neighbor (KNN) algorithm was selected for the classification task due to its simplicity and efficiency, particularly when handling small to medium-sized datasets [6-8]. KNN is an instance-based, nonparametric learning algorithm that predicts the class of a new data point by assessing the distance between it and the nearest neighbors within the training set. The underlying principle of KNN is that similar data points are generally situated close to one another in the feature space, leading to the assumption that a new point is likely to belong to the same class as its closest neighbors.

For each image, features were extracted using the Histogram of Oriented Gradients (HOG) technique [9], which effectively captures edge and gradient information. The resulting feature vector serves as the input to the KNN classifier. To reduce computational complexity and enhance performance, Principal Component Analysis (PCA) was applied to the HOG features [10], thereby reducing dimensionality while retaining the most critical information.

The distance between the new data point and all points in the training set was computed. Several distance metrics were considered, with the most notable being Euclidean distance, Manhattan distance, and Minkowski distance. The algorithm then identified the "k" nearest neighbors and assigned the new data point to the most common class among these neighbors. The value of "k," a hyperparameter, was fine-tuned to achieve optimal performance, with a value of 10 being identified as yielding the highest accuracy in this model.

The KNN model was developed using Python libraries, with hyperparameters such as n_neighbors, weights, and distance measures being optimized through a random search over predefined parameter distributions. The best model was selected based on cross-validation performance on the training set. The model's performance was then evaluated using multiple metrics, with accuracy being the primary measure, indicating the proportion of correctly classified instances on both the validation and test sets. Additional evaluation metrics included the confusion matrix, which provided insights into the types of errors made by the model, as well as precision, recall, and F1 scores, which were particularly valuable for assessing the model's performance across different categories.

3. Results and Discussion

The results of this study show that the KNN model algorithm has excellent accuracy when applied to the classification of fruits and vegetables. This section focuses on the experimental results, along with a detailed analysis of the confusion matrix, F1 scores, and other performance metrics.

3.1 The Performance of the Model

The KNN model is trained and validated using carefully preprocessed data sets, including steps such as feature ex-

traction using directional gradient histograms (HOG) and dimensionality reduction through principal component analysis (PCA). After fine-tuning various hyperparameters, including number of neighbors (k), distance measures, and weighting schemes, the model achieved a stunning 97% accuracy on both the validation and test sets.

This high precision reflects the model's ability to generalize well from training data to new, unseen instances. The selected k value is determined to be 10, achieving the best balance between bias and variance. The lower the k value, the more sensitive the model is to the noise in the training data, resulting in overfitting; The higher the value of k, the higher the degree of model generalization and the failure to capture important patterns. The Euclidean distance measure combined with uniform weighting was found to be the most effective, further improving the robustness of the model.

The consistency of accuracy between validation set and test set is a strong indicator of model reliability. This consistency shows that the model is not only well-tuned, but also effective at learning the distinguishing features needed for accurate classification, even when faced with new data.

3.2 Confusion Matrix Analysis

To further analyze the performance of the model, this study generates a confusion matrix of the validation set shown in Fig. 1 and test set shown in Fig. 2. The confusion Matrix details the number of correct and incorrect predictions in all categories, providing valuable insights into the strengths and weaknesses of the model.

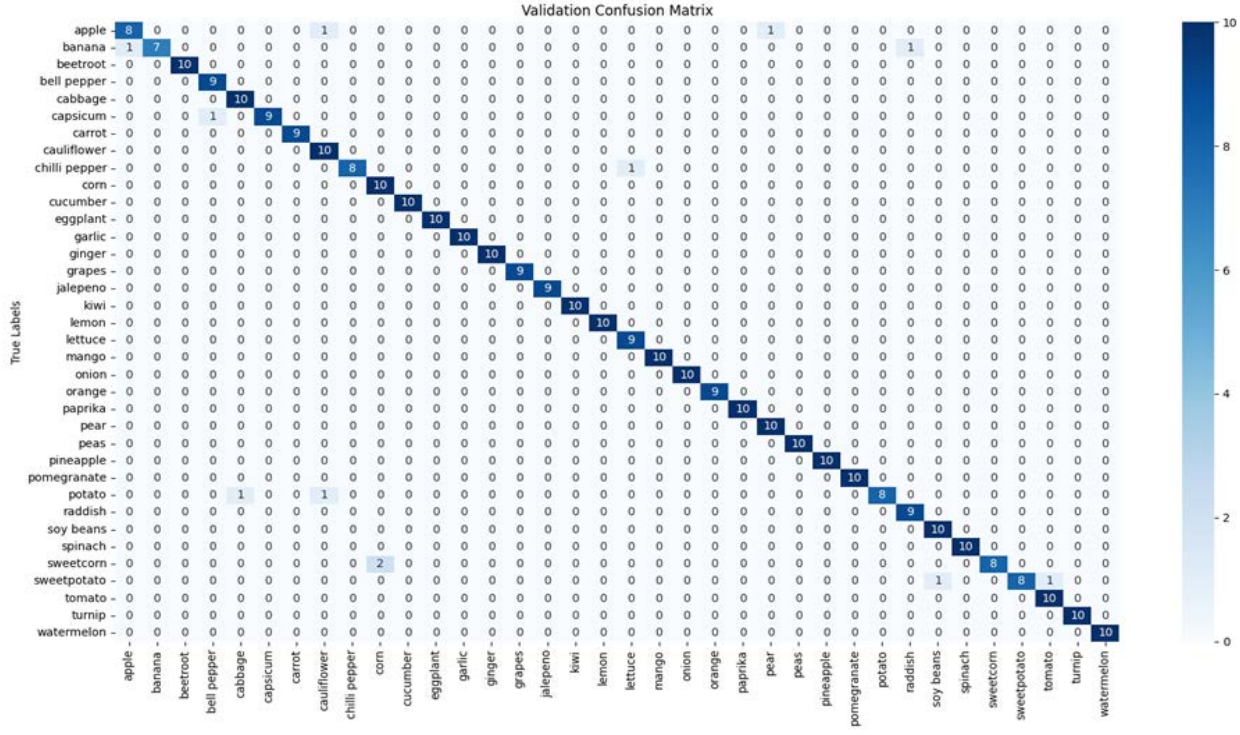

Fig. 1 Verification set confusion matrix (Photo/Picture credit: Original).

The confusion matrix of the validation set shows that most predictions are correct, as evidenced by the high values on the diagonal. This shows that the model successfully identifies the correct categories for most of the images. However, there are some cases of misclassification, especially between categories with similar visual characteristics. For example, different types of apples and citrus fruits can occasionally be confused with each other, possibly because of their similar shapes and textures. These misclassifications are relatively rare, but they highlight the potential for further improvements in the feature extraction process.

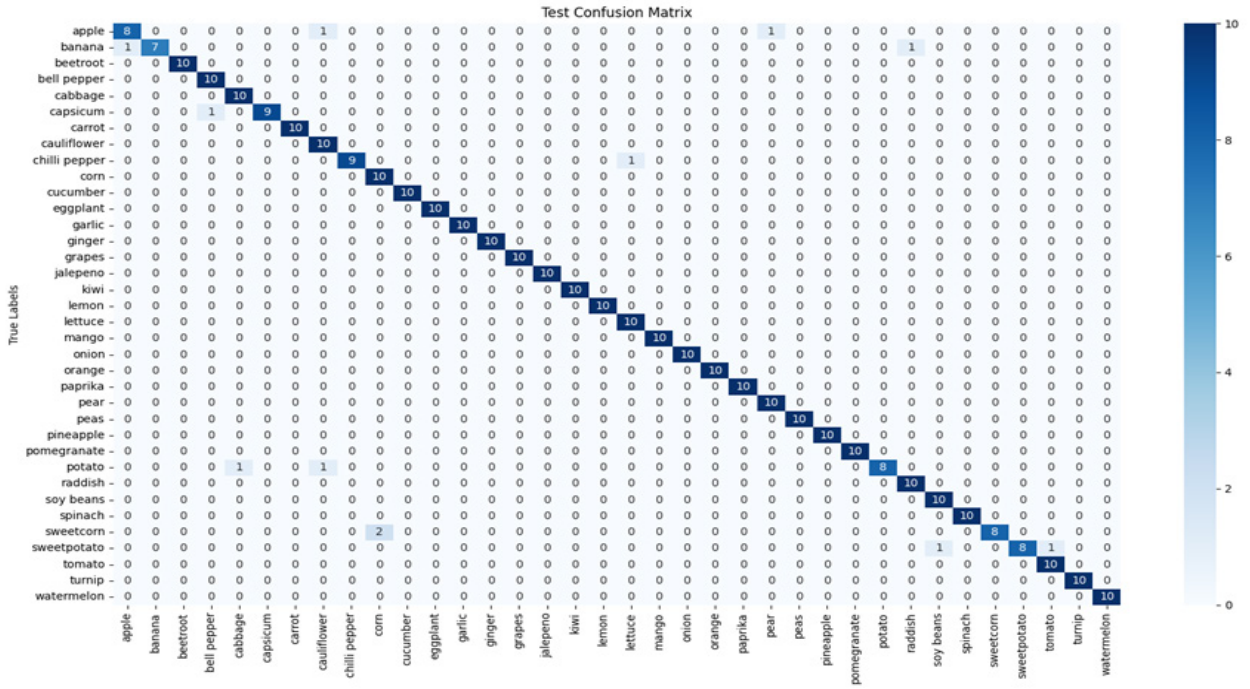

Fig. 2 Test set confusion matrix (Photo/Picture credit: Original).

The test set confusion matrix shows a similar pattern to the validation set, with most predictions being correct. The consistency between the results of the validation and test sets indicates that the model is robust and generalizes well to new data. However, as with validation sets, some misclassifications have been observed, especially in visually similar categories. For example, certain varieties of chili peppers are occasionally misclassified, possibly because of their similar shape and size. Despite these errors, the overall accuracy is still high, which enhances the validity of the model.

3.3 Analysis of F1 Scores

The F1 score is a key metric to assess the balance between accuracy and recall in a model's performance. In this study, an F1 score plot was provided to give a comprehensive overview of how KNN models perform across all categories in the validation and test sets.

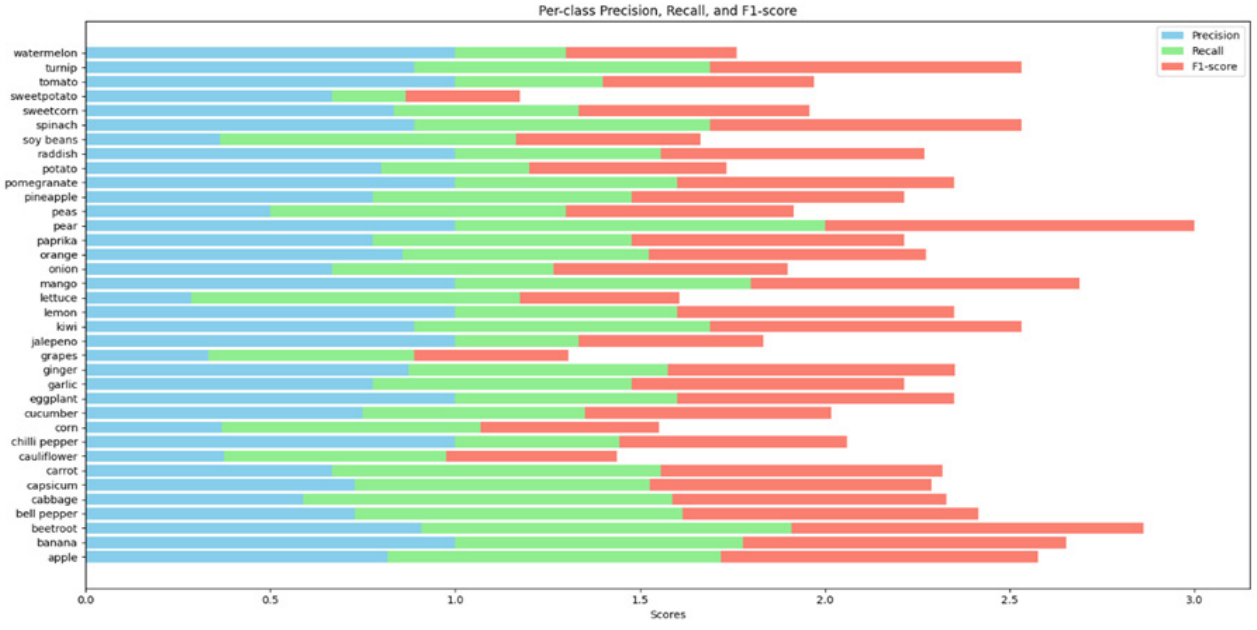

Fig. 3 F1 total results (Photo/Picture credit: Original).

The F1 score chart shown in Fig. 3 shows the scores for each category, clearly and intuitively showing the strengths and weaknesses of the model. F1 scores for most categories were close to 1.0, indicating that the model was consistently able to make accurate predictions with high accuracy and recall rates. This shows that the model can effectively identify the correct category for most images, minimizing false positives and missed positives.

However, the chart also highlights some categories with slightly lower F1 scores. These lower scores correspond to categories that are harder for the model to classify correctly, as shown in the confusion matrix. For example, categories with visually similar characteristics, such as different types of apples or peppers, had slightly lower F1 scores. This shows that while the model performs well overall, there is still room for improvement in distinguishing categories with subtle visual differences.

The F1 score map is a very useful tool for quickly identifying areas of strength in a model and areas that may need further improvement. By focusing on these specific categories, future iterations of the model can be improved, either by enhancing the feature extraction process or by increasing the diversity and size of the training dataset.

4. Conclusion

This study successfully applied the KNN model algorithm to the classification of fruits and vegetables, demonstrating the potential of machine learning in enhancing agricultural image recognition. By utilizing preprocessed image data sets and implementing feature extraction techniques such as HOG and PCA, the KNN model achieves a significant accuracy of 97% on both the validation and test sets.

The results highlight the effectiveness of KNN models in automating the classification process, reducing human error and improving efficiency. However, there are still challenges in distinguishing visually similar categories, suggesting that further refinement of feature extraction or inclusion of additional data can improve accuracy. Future research could explore integrating advanced models such as Convolutional Neural Networks (CNNS) or integrating more diverse datasets to address these limitations. And more models are soft together so that the accuracy and application rate can be improved more significantly. Overall, this study provides valuable insights into the application of KNN models in agricultural technology, laying the foundation for further development of automated classification systems in the food industry.

References

[1] Patel K, Patel HB. A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. In2021 fourth international conference on electrical, computer and communication technologies (ICECCT) 2021 Sep 15 (pp. 1-5). IEEE.

[2] Yamaç SS. Reference evapotranspiration estimation with kNN and ANN models using different climate input combinations in the semi-arid environment. Journal of Agricultural Sciences. 2021 Apr 1.

[3] Razavi S, Yalçın H. Plant classification using group of features. In2016 24th Signal Processing and Communication Application Conference (SIU) 2016 May 16 (pp. 1957-1960). IEEE.

[4] Shaik MA, Manoharan G, Prashanth B, Akhil N, Akash A, Reddy TR. Prediction of crop yield using machine learning. InAIP Conference Proceedings 2022 May 24 (Vol. 2418, No. 1). AIP Publishing.

[5] Kaggle, Fruit and Vegetable Image Recognition, 2021, https://www.kaggle.com/datasets/kritikseth/fruit-and-vegetable-image-recognition

[6] Peterson LE. K-nearest neighbor. Scholarpedia. 2009 Feb 21;4(2):1883.

[7] Laaksonen J, Oja E. Classification with learning k-nearest neighbors. InProceedings of international conference on neural networks (ICNN'96) 1996 Jun 3 (Vol. 3, pp. 1480-1483). IEEE.

[8] Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics. 1985 Jul(4):580-5.

[9] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) 2005 Jun 20 (Vol. 1, pp. 886-893). Ieee.

[10] Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Computers & Geosciences. 1993 Mar 1;19(3):303-42.