ISSN 2959-6157

Fruit and Vegetable Image Recognition Based on Multiple Tree Models: Applications of Random Forest, XGBoost and Decision Tree

Zihao Wang

School of Science, Rensselaer Polytechnic Institute, New York, America
wangz61@rpi.edu

Abstract:

The primary objective of this study is to evaluate and compare the performance of three machine learning models—Random Forest, XGBoost, and Decision Tree—in the context of fruit and vegetable image classification. This research aims to identify which model best handles the challenges associated with imbalanced datasets and complex data structures. The ultimate goal is to contribute to the development of more efficient and accurate automated systems for agricultural applications, thereby improving productivity and reducing operational costs in the industry. This study utilized a dataset of 3,825 images covering 36 fruit and vegetable classes. Images were resized, normalized, and augmented to enhance diversity. Three models—Random Forest, XGBoost, and Decision Tree—were trained on this dataset. Performance was evaluated using accuracy, precision, recall, and F1-score to assess classification effectiveness and handling of class imbalances. The evaluation revealed that XGBoost outperformed Random Forest and Decision Tree in fruit and vegetable image classification, achieving the highest accuracy of 96.66%. XGBoost demonstrated superior handling of class imbalances and complex data structures, reflected in its precision and recall scores across various classes. Random Forest also performed well, closely following XGBoost, while Decision Tree exhibited more variability in results, indicating potential overfitting in certain classes. In conclusion, this study highlights the effectiveness of ensemble methods, particularly XGBoost, in agricultural image classification tasks. These findings suggest that XGBoost is a robust model for similar applications, offering improved accuracy and reliability.

Keywords: Machine Learning Algorithms, Image Recognition, Ensemble Learning.

1. Introduction

The recognition and classification of fruits and vegetables in images have practical applications in automated retail checkout systems, food supply chain management, and dietary monitoring. Research in this area not only advances related technologies but also significantly improves operational efficiency across various industries. In automated retail checkout systems, accurate identification of fruits and vegetables can reduce manual intervention, speed up the checkout process, and enhance the consumer shopping experience. In food supply chain management, precise image recognition aids in tracking and managing inventory, reducing waste, and optimizing supply chain processes. In dietary monitoring, recognizing food types helps users better manage their diet and promote healthy living. This integration of advanced image recognition technologies offers substantial benefits, enhancing consumer experiences in retail, optimizing supply chains, and supporting health and wellness initiatives.

In recent years, Artificial Intelligence (AI) technology has made significant strides, with various representative algorithms such as Decision Tree [1], Random Forest [2], Gradient Boosting Decision Tree (GDBT) [3], and eXtreme Gradient Boosting (XGBoost) [4], being widely applied across multiple fields, including chemistry, biomedical sciences, and agriculture. In the agricultural domain, AI techniques have been utilized for various predictive and classification tasks. For instance, Muhammet Çakmak et al. used eXtreme Gradient Boosting (XGB), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) machine learning methods to determine the quality of apples, and the results showed that XGB and RF accuracy was significantly higher than other methods [5]. Additionally, Rocha et al. introduced a unified method combining multiple features and classifiers for the automatic classification of fruits and vegetables from images. It addresses the complexity of these tasks by employing feature fusion, which minimizes training data requirements and enhances classification accuracy. The technique is tested on a dataset featuring 15 categories of fruits and vegetables, significantly reducing classification errors by up to 15 percentage points compared to traditional methods [6]. Yang et al. presented the development of image recognition software based on artificial intelligence algorithms for efficient sorting of apples. The study introduces a low-cost machine vision system that employs deep learning, specifically Convolutional Neural Networks (CNN), for automated apple grading and sorting. The system achieved an average accuracy of 99.70% and a recognition accuracy of 99.38% for the CNN-based apple sorting system, demonstrating its feasibility for medium and large-scale enterprises [7].

This project aims to develop a model using machine learning techniques to classify various fruits and vegetables. Specifically, different machine learning methods, including but not limited to Decision Trees and Random Forests, will be explored and applied to achieve high-precision fruit and vegetable image classification. The goal is to provide an efficient and accurate solution that can be widely applied in various practical applications. To achieve this goal, research and experiments will be conducted based on publicly available fruit and vegetable image datasets, particularly those provided by Kaggle. The research process will include image preprocessing, feature extraction, model training, and performance evaluation. The performance of different algorithms will be compared and analyzed in terms of classification accuracy, computational efficiency, and robustness to identify the optimal solution

2. Method

2.1 Dataset Preparation

The dataset utilized in this study was obtained from the Kaggle Fruit and Vegetable Image Recognition dataset [8]. This dataset comprises 3,825 images, categorized into 36 distinct classes of fruits and vegetables. The images vary in size, with each image provided in RGB format, offering a full-color spectrum ideal for image classification tasks. Given the diversity in image sizes, preprocessing was essential. All images were uniformly resized to 150×150 pixels, ensuring consistency across the dataset. Additionally, normalization was performed by scaling pixel values to a range between 0 and 1, a standard practice that facilitates faster and more stable convergence of the models. Fig. 1 provides the sample images of the collected dataset.

Fig. 1 The sample images of the collected dataset [8].

To further enhance the dataset, augmentation techniques such as rotation, flipping, and zooming were applied. These augmentations not only aimed to artificially increase the dataset's diversity—making the models more robust by exposing them to various transformations—but also to prevent underfitting by ensuring that the models are sufficiently complex to capture the underlying patterns in the data. This careful balance between avoiding overfitting, where the model performs well on training data but poorly on unseen data [9], and underfitting, where the model fails to capture the data's complexity [10], is crucial for developing effective machine learning models.

2.2 Machine Learning Models-based Prediction

Three machine learning models were employed in this study: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Decision Tree (DT). These models were implemented using the Scikit-learn (sklearn) library in Python, which provides a wide range of tools for data mining and data analysis. The models were evaluated based on several key metrics, including accuracy, precision, recall, and the confusion matrix, to thoroughly assess their performance in classifying fruits and vegetables. By leveraging these metrics, the study aimed to identify the most effective model for this specific classification task.

2.2.1 Random forest

Random Forest (RF) is a robust ensemble learning technique that aggregates the predictions of multiple decision trees to improve classification accuracy and reduce the risk of overfitting. Each tree in the forest is trained on a random subset of the training data, a process known as bootstrap sampling, and considers a random subset of features when splitting nodes, which introduces diversity among the trees. The final prediction is determined by majority voting across all trees. RF is particularly advantageous for handling large datasets with high dimensionality, as it efficiently manages missing data and maintains accuracy even with noisy datasets. Additionally, RF provides insights into feature importance, making it a valuable tool for identifying the most significant predictors in the dataset.

2.2.2 XGBoost

eXtreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm based on the gradient boosting framework. It builds trees sequentially, where each new tree attempts to correct the residual errors made by the previous trees. XGBoost is known for its scalability and speed, leveraging advanced optimization techniques such as parallel processing, tree pruning, and regularization to prevent overfitting. The model includes features like handling missing data internally, applying sparsity-aware algorithms, and using a weighted quantile sketch for approximate tree learning, making it particularly effective for complex, high-dimensional datasets. XGBoost's ability to generalize well across different data types and its fine-tuned control over model complexity have made it a preferred choice for many predictive modeling tasks.

2.2.3 Decision tree

Decision Tree (DT) is a straightforward yet effective machine learning model that classifies data by recursively splitting it into subsets based on the feature that maximizes information gain. The tree consists of internal nodes representing feature tests, branches representing the outcomes of these tests, and leaf nodes representing class labels. Decision Trees are highly interpretable, as the decision-making process is visualized in a tree-like structure, making them useful for understanding and explaining the relationships within data. However, they are prone to overfitting, particularly with complex datasets. To mitigate this, techniques such as pruning, setting a maximum depth, and defining a minimum number of samples required to split a node are often employed. Despite its simplicity, the Decision Tree serves as a foundational model, often used as a building block for more complex ensemble methods like Random Forest and XGBoost.

3. Results and Discussion

The overall training and testing accuracy for the three models—Random Forest, XGBoost, and Decision Tree—were evaluated shown in Table 1. All three models demonstrated the same high training accuracy (0.9942), reflecting their ability to fit the training data well. However, their testing accuracy varied slightly, with XGBoost achieving the highest accuracy of 0.9666, followed closely by Random Forest and Decision Tree, both at 0.9638. This indicates that while all models are effective, XGBoost has a slight edge in generalizing to unseen data.

Table 1. Model Performance Comparison

Model	Training Accuracy	Testing Accuracy
Random Forest	0.9942	0.9638
XGBoost	0.9942	0.9666
Decision Tree	0.9942	0.9638

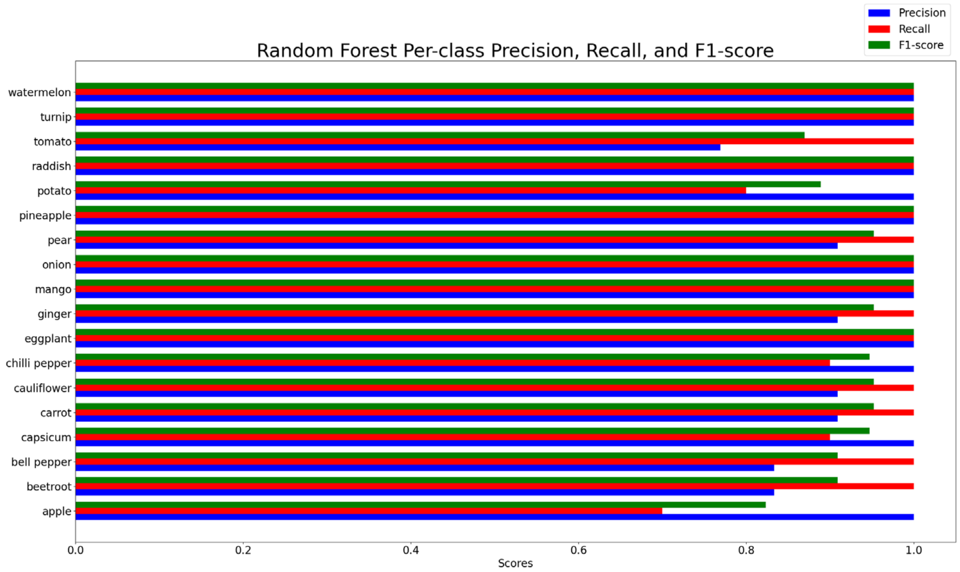


Fig. 2 Part of the classification report of the Random Forest models (Photo/Picture credit: Original)

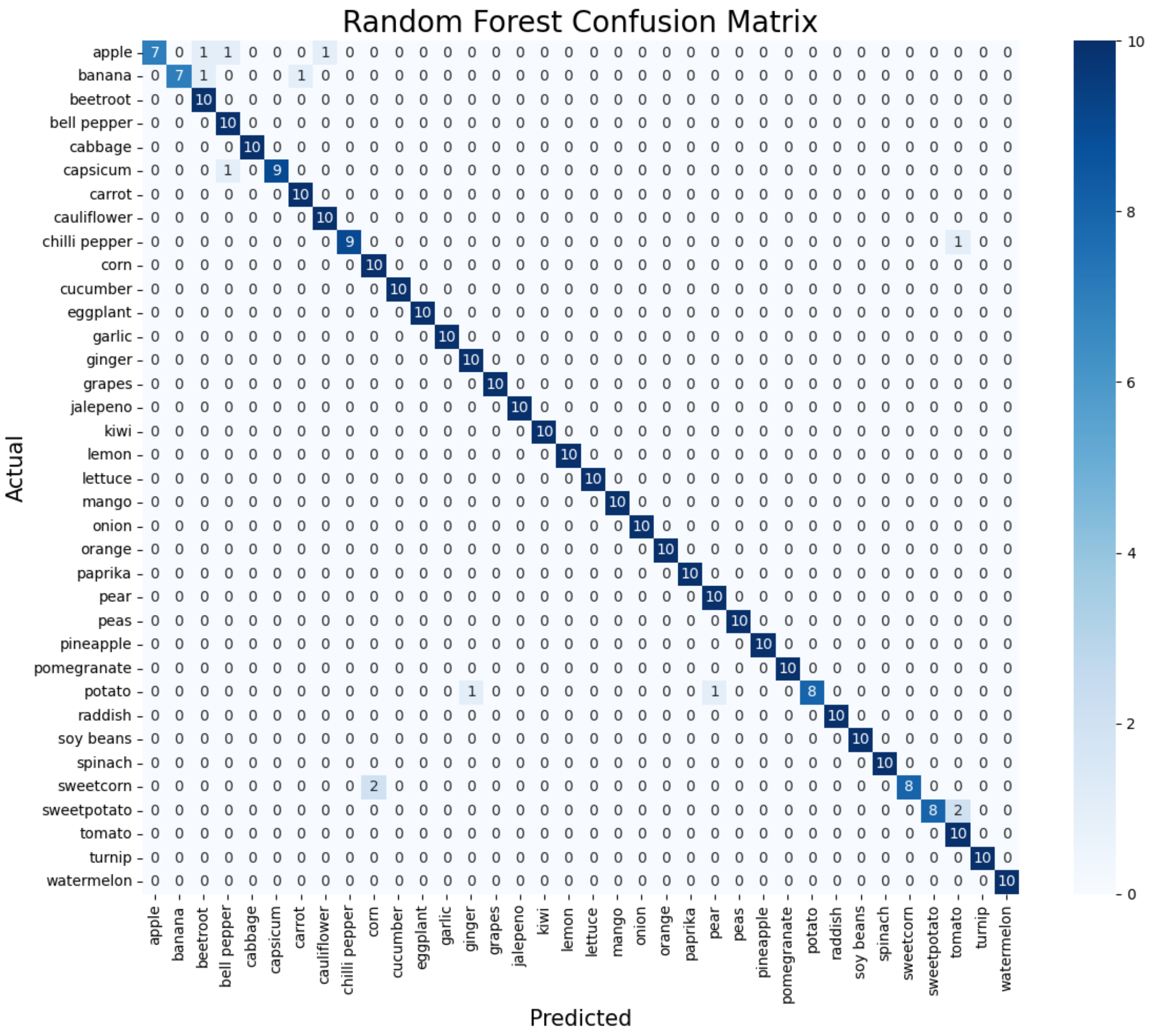


Fig. 3 The confusion matrix of the Random Forest model (Photo/Picture credit: Original)

The Random Forest model exhibited strong performance across most classes, particularly those with a large number of samples. Precision and recall were high for the majority of classes, with perfect scores in many cases such as “turnip,” “onion,” and “mango.” However, there were slight inconsistencies, for instance, “apple” had a precision of 1.00 but a lower recall of 0.70, resulting in an F1-score of 0.82. Fig. 2 provides the classification report of Random Forest. The confusion matrix shows that the Random Forest model was highly accurate in most classifications, with very few misclassifications. Fig. 3 provides the confusion matrix of Random Forest.

Random Forest performed well, showcasing its ability to generalize effectively across various classes. However, it struggled slightly in classes with fewer samples, suggesting potential areas for improvement, such as enhancing its handling of data imbalances. Random Forest also performed well, showcasing its ability to generalize effectively across various classes. However, it struggled slightly in classes with fewer samples, suggesting potential areas for improvement, such as enhancing its handling of data imbalances.

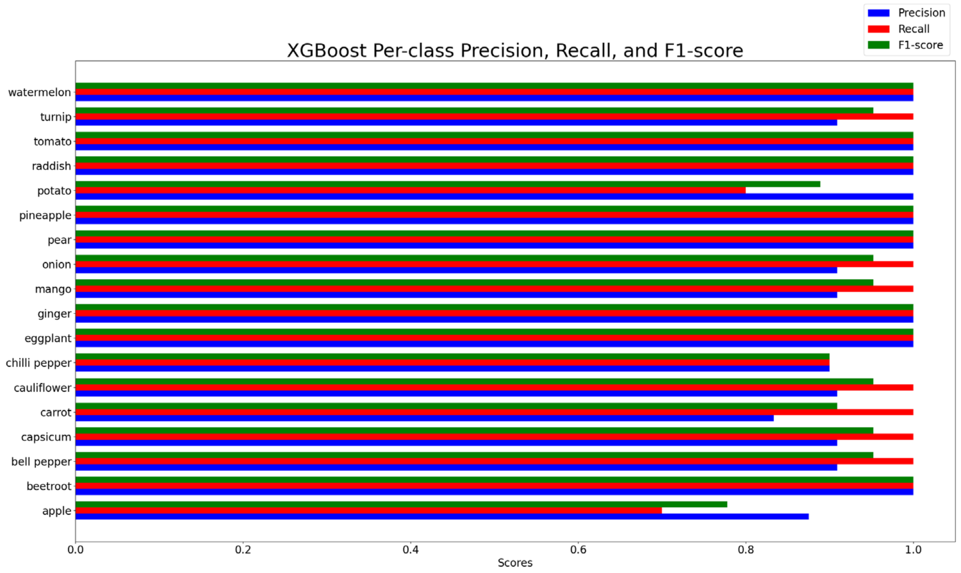
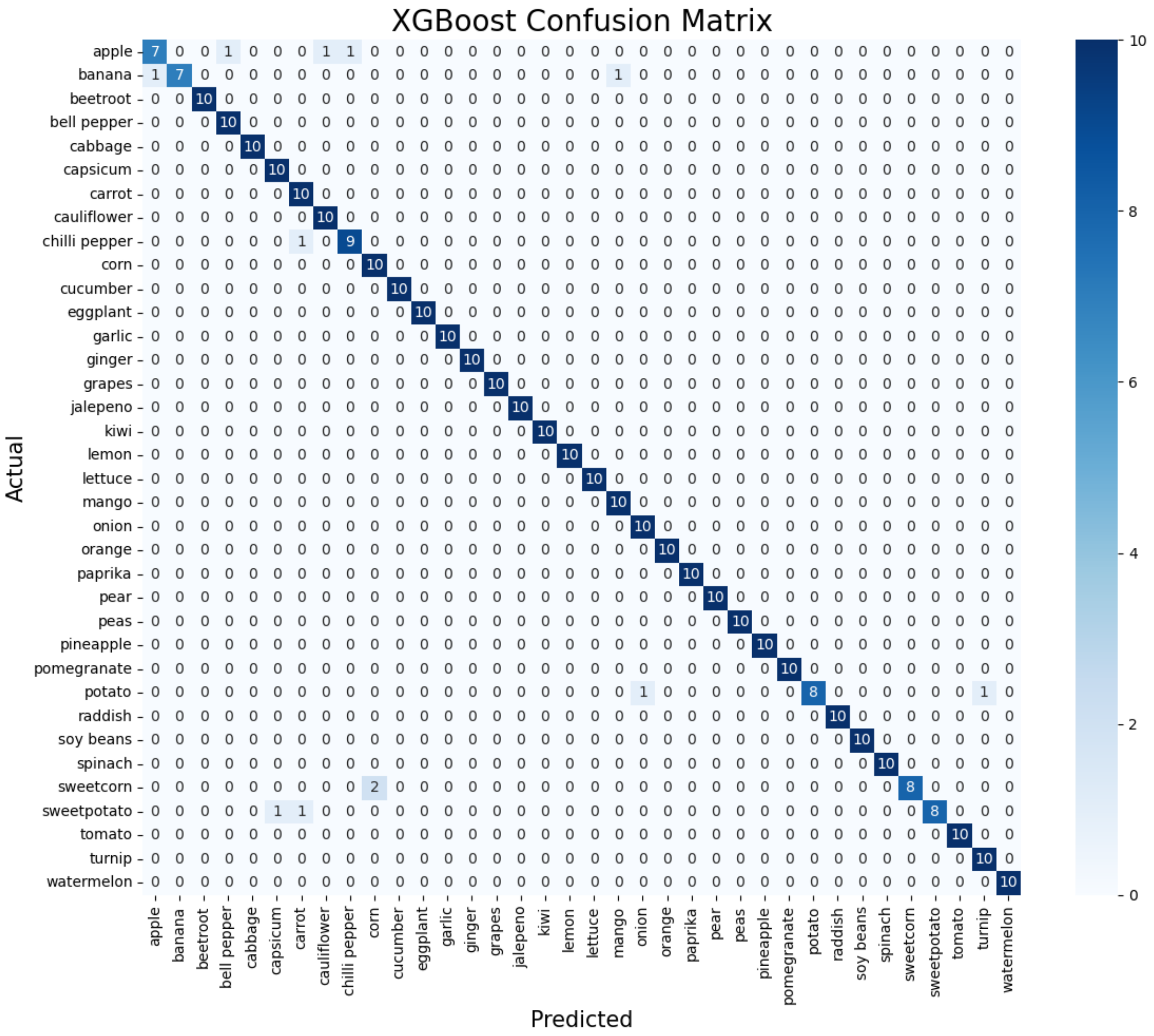


Fig. 4 Part of the classification report of the XGBoost models (Photo/Picture credit: Original)



	apple	banana	beetroot	bell pepper	cabbage	capsicum	carrot	cauliflower	chilli pepper	corn	cucumber	eggplant	garlic	ginger	grapes	jalepeno	kiwi	lemon	lettuce	mango	onion	orange	paprika	pear	peas	pineapple	pomegranate	potato	raddish	soy beans	spinach	sweetcorn	sweetpotato	tomato	turnip	watermelon
apple	7	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
banana	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
beetroot	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bell pepper	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cabbage	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
capsicum	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
carrot	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cauliflower	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chilli pepper	0	0	0	0	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cucumber	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eggplant	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
garlic	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ginger	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
jalepeno	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kiwi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lemon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lettuce	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
onion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
paprika	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
pear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
peas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
pineapple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0
potato	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	8	0	0	0	0	0	0	1	0
raddish	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
soy beans	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
spinach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
sweetcorn	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
sweetpotato	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0
tomato	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
turnip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
watermelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Fig. 5 The confusion matrix of the XGBoost model (Photo/Picture credit: Original)

XGBoost slightly outperformed Random Forest, especially in handling imbalanced classes. It maintained high precision, recall, and F1-scores across all classes, with perfect scores in several, such as "beetroot" and "cabbage." XGBoost's ability to handle complex data and regularize effectively contributes to its superior performance. Fig. 4 provides the classification report of XGBoost. The confusion matrix for XGBoost reflects its robustness, showing fewer misclassifications compared to Random Forest, particularly in challenging classes. Fig. 5 provides the confusion matrix of XGBoost.

XGBoost emerged as the top performer, achieving the highest accuracy and demonstrating robust performance across most classes. Its advanced handling of class imbalances and regularization techniques were key factors in its success. This model is particularly well-suited for applications requiring high accuracy and stability, even in the presence of complex and imbalanced datasets.

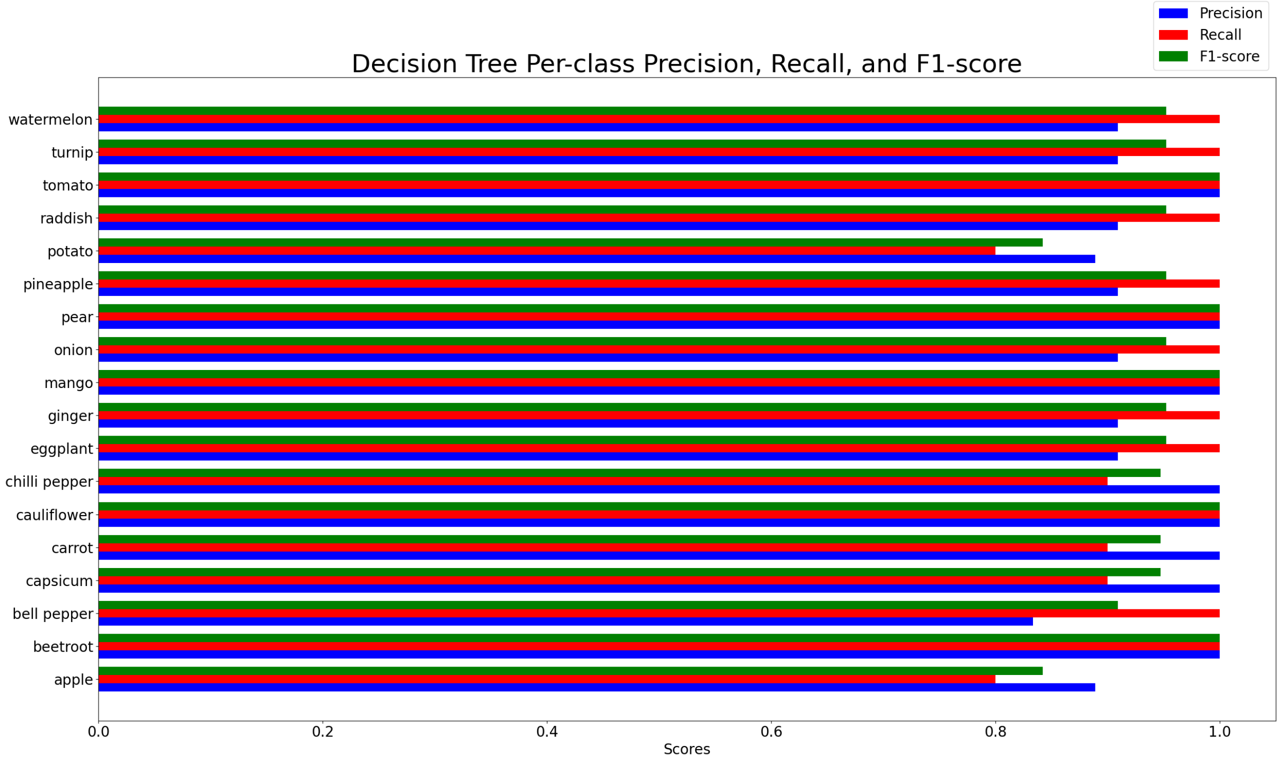
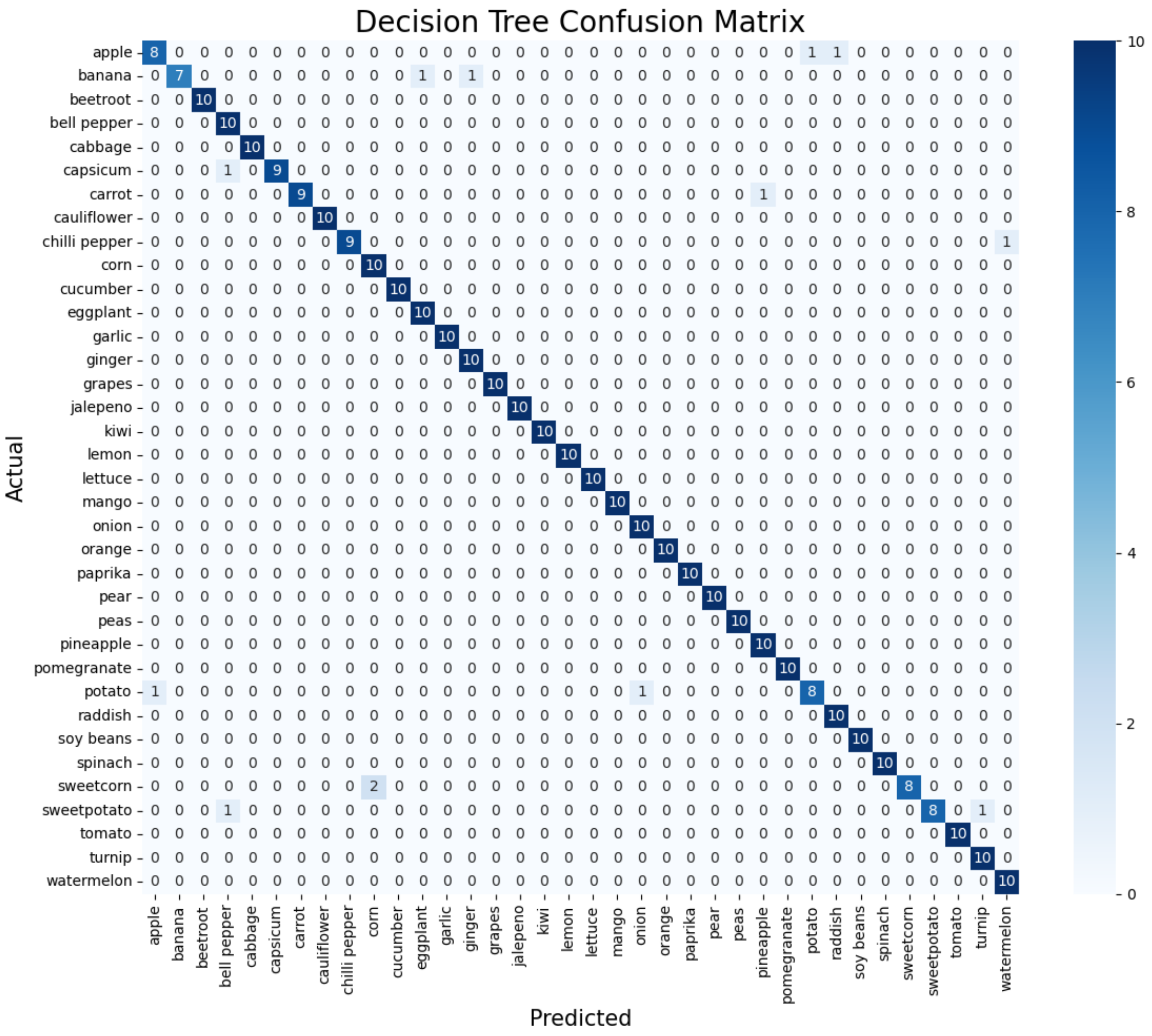


Fig. 6 Part of the classification report of the Decision Tree models (Photo/ Picture credit: Original)



	apple	banana	beetroot	bell pepper	cabbage	capsicum	carrot	cauliflower	chilli pepper	corn	cucumber	eggplant	garlic	ginger	grapes	jalepeno	kiwi	lemon	lettuce	mango	onion	orange	paprika	pear	peas	pineapple	pomegranate	potato	raddish	soy beans	spinach	sweetcorn	sweetpotato	tomato	turnip	watermelon
apple	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
banana	0	7	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
beetroot	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bell pepper	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cabbage	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
capsicum	0	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
carrot	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
cauliflower	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chilli pepper	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
corn	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cucumber	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eggplant	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
garlic	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ginger	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
jalepeno	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kiwi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lemon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lettuce	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
onion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
paprika	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
pear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
peas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
pineapple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0
potato	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0
raddish	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
soy beans	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
spinach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
sweetcorn	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
sweetpotato	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	1	0
tomato	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
turnip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
watermelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Fig. 7 The confusion matrix of the Decision Tree model (Photo/Picture credit: Original)

The Decision Tree model showed good performance, similar to Random Forest, but with more variability in precision and recall across classes. Some classes like “apple” and “potato” had lower recall values, leading to slightly lower F1-scores compared to the ensemble methods. This variability suggests that while Decision Trees are effective, they are more prone to overfitting on certain classes. Fig. 6 provides the classification report of Decision Tree. The confusion matrix for the Decision Tree model shows that while it performed well overall, there were more instances of misclassification compared to XGBoost and Random Forest. Fig. 7 provides the confusion matrix of Decision Tree.

Decision Tree, while easy to interpret and implement, underperformed relative to the ensemble methods. The higher variance in its performance metrics and lower overall accuracy indicate that it may not be the best choice for tasks involving complex datasets with multiple classes.

4. Conclusion

This study evaluated the performance of three machine learning models—Random Forest, XGBoost, and Decision Tree—in fruit and vegetable image classification. The findings show that while all models achieved high accuracy, XGBoost outperformed the others, especially in handling class imbalances and complex data. The research highlights the strengths of ensemble methods in image classification, though it is limited by its focus on classical machine learning techniques. Future work could explore deep learning models to enhance accuracy and generalization across broader datasets. The study aims to provide effective technical means for fruit and vegetable image recognition and serve as a reference for related research and applications. As machine learning technologies advance, their use in automating fruit and vegetable image recognition can enhance efficiency, reduce costs, and promote the modernization of the industry. Further exploration of this study's findings may offer new ideas and methods for optimizing and applying machine learning algorithms in practical scenarios, driving their broader application.

References

[1] Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar;1:81-106.

[2] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

[3] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001 Oct 1:1189-232.

[4] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[5] Çakmak M. Classification of Apple Quality Using XGBoost Machine Learning Model. InKonya: 4th International Conference on Innovative Academic Studies 2024 Mar (pp. 607-615).

[6] Rocha A, Hauagge DC, Wainer J, Goldenstein S. Automatic fruit and vegetable classification from images. Computers and Electronics in Agriculture. 2010 Jan 1;70(1):96-104.

[7] Yang M, Kumar P, Bhola J, Shabaz M. Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. International Journal of System Assurance Engineering and Management. 2022 Mar;13(Suppl 1):322-30.

[8] Seth K. Fruits and Vegetables Image Recognition Dataset. Kaggle; 2020. Available from: https://www.kaggle.com/kritikseth/fruit-and-vegetable-image-recognition.

[9] Ying X. An overview of overfitting and its solutions. In Journal of physics: Conference series 2019 Feb (Vol. 1168, p. 022022). IOP Publishing.

[10] Wang H, Lei Z, Zhang X, Zhou B, Peng J. Machine learning basics. Deep learning. 2016:98-164.