Article:

Guo, Z., Xiao, H., Dai, Z. et al. (5 more authors) (2025) Identification of apple variety using machine vision and deep learning with Multi-Head Attention mechanism and GLCM. Journal of Food Measurement and Characterization. ISSN 2193-4126

https://doi.org/10.1007/s11694-025-03385-5

Identification of apple variety using machine vision and deep learning with multi-head attention mechanism and GLCM

Zhiming Guo[a, b]*, Haidi Xiao[a], Zhiqiang Dai[c], Chen Wang[d], Chanjun Sun[a], Nicholas Watson[b], Megan Povey[b], Xiaobo Zou[a]

[a] *China Light Industry Key Laboratory of Food Intelligent Detection & Processing, School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China*

[b] *School of Food Science and Nutrition, University of Leeds, Leeds LS2 9JT, United Kingdom*

[c] *Beijing Key Laboratory of Multi-dimension & Multi-scale Computational Photography, LUSTER LightTech Co., Ltd. Beijing 100094, China*

[d] *Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, School of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China*

*Corresponding author at: School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China. Email addressed: guozhiming@ujs.edu.cn (Z. Guo)

Abstract: Apple variety identification plays a crucial role in pomology and agricultural sciences, as it could effectively assist growers in optimizing orchard management, enhancing product quality, and meeting consumer demand. Traditional identification methods based on visual observation are often influenced by various factors, including human subjective judgment and inter-cultivar variability. To address these challenges, with the support of the China Agriculture Research Systems for Apple Industry and Jiangsu University, we collected sample images of eleven common apple varieties in China, followed by image enhancement and dataset expansion to establish an apple sample database. Subsequently, Convolutional Neural Network (CNN), MobileNet Version 2 (MobileNetV2), and Visual Geometry Group 19 (VGG19) neural network models were utilized for apple variety classification using image-based data. Additionally, two optimization techniques, namely Multi-Head Attention and Gray-Level Co-occurrence Matrix (GLCM), were incorporated to further improve classification accuracy. Results demonstrated that the baseline CNN achieved an accuracy of 96.46%, while MobileNetV2 and VGG19 reached 97.78% and 97.25%, respectively. Multi-Head Attention improved feature extraction but sometimes reduced performance, as observed in MobileNetV2 (87.33%). In contrast, GLCM significantly improved model accuracy, with MobileNetV2 achieving the highest accuracy (98.25%) and the lowest Mean Absolute Error (MAE) (0.0571). GLCM consistently outperformed other techniques across all models, proving particularly effective for texture-rich image classification. These findings suggest that GLCM is a powerful enhancement for deep learning models, improving accuracy, precision, and recall in apple variety classification, with MobileNetV2 combined with GLCM yielding the best overall results.

Keywords: Apple variety classification; Deep learning; Optimization techniques; Convolutional neural network

44 # List of abbreviations

Abbreviation	Full Term
CNN	Convolutional Neural Network
MobileNetV2	Mobile Network Version 2
VGG19	Visual Geometry Group 19-layer
MAE	Mean Absolute Error
GI	Geographical Indication
CNNs	Convolutional Neural Networks
RH	Relative Humidity
TP	True Positive
TN	True Negatives
FP	False Positive
FN	False Negative
t-SNE	t-distributed stochastic neighbor embedding
PC	principal components
SSL	Self-Supervised Learning

45

1. Introduction

As one of the most widely cultivated fruits globally, apples hold substantial economic value, contributing billions of dollars annually to economies around the world. China, as one of the largest apple-producing regions, harvested 47.57 million tons of apples in 2023, accounting for more than half of the global apple production [1]. The long shelf life of apples and their suitability for various preservation techniques, such as refrigeration and canning, further enhance their economic importance.

Apples play a pivotal role in futures markets by facilitating price discovery and risk management, allowing producers, traders, and consumers to lock in prices and mitigate risks associated with price fluctuations. Accurate identification of apple varieties is fundamental for improving price discovery in futures markets, minimizing fraud during delivery, and enhancing agricultural efficiency.

With the growing promotion of Geographical Indication (GI) products, the connection between specific apple varieties and their production regions has become crucial for enhancing commercial value. GI products, protected by intellectual property laws, are associated with high quality and authenticity, fostering consumer trust. Accurate identification of apple varieties is vital in preventing fraud, such as the substitution or mixing of similar varieties, which compromises the integrity of GI products. Ensuring only the correct cultivar is labeled and sold protects the authenticity of GI apples and prevents misleading claims. For example, misrepresenting a premium cultivar like "Yanfu" with a lower-quality one like "ordinary Red Fuji" damages the market value and reputation of GI apples. Therefore, accurate cultivar identification is essential for improving trade transparency, promoting GI products, and protecting market integrity, thereby supporting the sustainable growth of the apple industry through fair competition and consumer trust [2].

However, identifying apple varieties is a significant challenge for farmers, traders, and consumers due to the large number of varieties that share similar appearances, especially within certain color ranges [3, 4]. Additionally, factors such as growing conditions, soil types, and cultivation practices influence the shape, color, and texture of apples, making the identification of apple varieties more complex [5]. In China, various apple varieties, such as Fuji, Red Delicious, and Gala, are cultivated for specific market segments. However, the introduction of new apple varieties closely resembling their parent varieties has made accurate classification challenging. Traditional methods, such as analyzing leaf and fruit characteristics, consulting experts, or using genetic testing, are time-consuming and struggle to balance efficiency with cost-effectiveness [6].

Since 2012, Convolutional Neural Networks (CNNs) have emerged as a leading technology in image processing and computer vision. Originally introduced by LeCun in the 1980s, CNNs serve as the foundational architecture in deep learning and have been widely applied to image classification and object detection tasks. The specific architecture is shown in Fig. 1. With local connectivity and weight sharing, CNNs enable efficient feature extraction and robust performance under varying conditions. In recent years, deep learning driven by CNNs has increasingly been utilized to address complex challenges in agricultural systems [7, 8]. Applications include detecting cucumber powdery mildew [9], classifying the maturity stages of custard apple fruits through image processing [10], and predicting fruit size and weight in apples using RGB-D cameras [11]. Additionally, CNNs have been employed for target detection, with models developed for object category recognition [12, 13].

Apple cultivar recognition involves extracting discriminative features from visual attributes such as shape, texture, and color [14-16]. A study [17] used VGG16, VGG19, and MobileNet to distinguish between ten apple varieties, with DenseNet201 achieving 97.48% accuracy. Another study [18] combined CNNs with a convolutional autoencoder to classify 26 fruits, including nine apple varieties. Additionally, a separate study [19] developed a shallow CNN to simplify deep neural networks for apple image recognition, achieving 92% accuracy.

For real-time applications and deployment on mobile or embedded devices, models with lower computational complexity are essential [20, 21]. While architectures like EfficientNet and Vision Transformers offer high performance, their high computational demands make them unsuitable for resource-constrained environments. Similarly, YOLO's emphasis on object localization limits its effectiveness in fine-grained classification, such as distinguishing subtle apple cultivar differences [22, 23]. In contrast, CNNs, including MobileNetV2 and VGG19, provide robust feature extraction. MobileNetV2, optimized for mobile environments, reduces computational costs through depth-wise separable convolutions and an inverted residual structure while maintaining strong representational power [24]. VGG19, developed by the Visual Geometry Group at Oxford in 2014, features a deeper architecture with 3×3 convolutional kernels, enabling better multi-level feature extraction and enhanced capability for distinguishing subtle morphological differences in apple varieties [25]. Both models balance efficiency, feature extraction, and robustness, making them ideal for apple cultivar recognition in practical, resource-limited settings [26, 27].

To improve model performance, two optimization techniques—Multi-Head Attention mechanism and GLCM—were used. Multi-Head Attention mechanism from the Transformer architecture enhances feature representation by capturing global dependencies [28]. GLCM, a texture analysis method, identifies pixel intensity co-occurrence patterns, aiding in the differentiation of similar apple varieties [29]. Its low computational complexity makes it ideal for lightweight models [25]. Together, these techniques improve accuracy and robustness in apple cultivar recognition.

The research approach is illustrated in Fig. 2. Our study utilizes a comprehensive dataset comprising eight varieties from the Fuji series, along with three additional widely cultivated apple varieties from major apple-producing regions in China. To address the challenge of new varieties closely resembling their parent varieties, we employ traditional CNN models, MobileNetV2 (efficiency), and VGG19 (feature extraction), incorporating the Multi-Head Attention mechanism and GLCM to improve feature extraction and classification accuracy. The objective is to enhance the accuracy and efficacy of apple cultivar classification by developing more efficient deep learning models and optimizing CNN architectures, evaluating the interaction of various components in real-world applications, and providing a comprehensive classification model for common apple varieties in China.

2. Material and methods

2.1 Apple samples

According to the 2023 data from the National Bureau of Statistics of China, the major apple varieties cultivated in China include the Fuji series, Red Delicious series, Gala series, and Golden Delicious series. Within the Fuji series, sub-varieties such as Red Fuji, Yanfu, and Miyakiji are widely cultivated. Fuji apples, including these sub-varieties, account for 69.8% of the total apple cultivation area in China, highlighting their dominance in the industry [30]. For this study, eleven apple varieties were selected, including eight from the Fuji series and three additional widely cultivated varieties, representing the major apple types in China.

With support from the China Agriculture Research Systems for Apple Industry, samples of eleven apple varieties were collected (Fig. 3). To enhance the generalizability of the apple sample images, samples of the same cultivar were sourced from different regions, improving the applicability of the apple classification model across diverse production areas and testing environments. The apple samples were gathered from major apple-producing regions across seven provinces in China and transported to the Apple Testing Center at Jiangsu University. Upon arrival, the apples were stored in the cold storage of the laboratory at a temperature of 0-2°C and a relative humidity (RH) of 85-90%.2.2

Image acquisition

The image collection for all evaluated apple varieties was completed within three days of receiving the samples at the Apple Testing Center at Jiangsu University. These samples were sourced from cooperatives, companies, and experimental stations in major apple-producing regions across China. The image collection process took place from September 18, 2023, to January 28, 2024.

Upon receipt of the samples, we classified them according to the GB/T 10651-2008 Chinese National Standard for Fresh Apples, categorizing them into Extra Class, Class I, Class II, and Substandard [31]. To minimize the impact of surface defects on feature extraction, apples classified as Class II or above were selected for image acquisition. The selection criteria were as follows: 1) A fruit diameter of at least 65 mm; 2) A shape index of 0.8 or higher; 3) A surface coloration rate of over 30%; 4) No more than two surface defects; 5) A total damage area smaller than one square centimeter.

Advanced hardware was utilized in the image acquisition process to ensure high-quality data collection. A BVC8350LC, a 3CMOS color area scan camera (Blue Vision

Corporation, Japan) served as the primary imaging device. This industrial-grade camera featured a resolution of 324 million pixels and was equipped with an f/1.2 maximum aperture lens to capture fine details. Full-spectrum industrial lighting was employed to provide uniform and consistent illumination, minimizing the influence of external lighting variations. The camera was equipped with a fixed-focus lens with a focal length of 20 mm and operated at a working distance of 650 mm. It operated at a shutter speed of 1/60 s to ensure sharp image capture. All images were saved in BMP format to preserve their quality and facilitate subsequent processing.

An average of 50 apple samples from each cultivar were selected for image collection. During this process, each apple sample was placed on a white-background platform for image capture, with images taken at 90-degree intervals along the apple's equator. Additionally, two images focusing on the apple's stem and calyx were captured, resulting in a total of six images per apple. Detailed information about the eleven photographed apple varieties was shown in Fig. 3. The dataset was then divided into training, validation, and test sets in a 6:2:2 ratio for model training.

2.3 Database construction

The original apple images were preprocessed to enhance data diversity and expand the dataset, thereby mitigating overfitting during model training. To reduce the influence of irrelevant information and highlight the apple as the primary subject, the images, initially with a resolution of 2448×1840 pixels, were cropped to 1324×1256 pixels. Four augmentation techniques—darkening, variation, Gaussian filtering, and

Gaussian noise—were applied to the images. For each cultivar, 50 images were randomly selected from the set and processed with each technique, ensuring a total of approximately 500 sample images per cultivar, as shown in Table 1.

2.4 Experimental design and algorithms

2.4.1. Experimental design

Image preprocessing: The quality of the apple images was affected by environmental factors, such as low light and vibration, as well as varietal characteristics. Preprocessing steps included cropping to focus on the apples, color correction for consistent lighting, noise reduction, brightness and contrast adjustment, and normalization. These measures improved image quality, ensuring reliable model training.

Model Training: Three base models (CNN, MobileNetV2, and VGG19) were employed for training. Additionally, GLCM features and Multi-Head Attention mechanisms were integrated with these models to create composite architectures. This combination harnesses both low-level texture information and high-level semantic features, enhancing classification accuracy and robustness. In total, nine distinct deep learning models were developed during the training process.

Model Evaluation and Deployment: Accuracy, precision, recall, and MAE were commonly used parameters for evaluating model performance [32]. The trained models were tested for prediction to identify issues related to overfitting or underfitting [2]. The validated models were uploaded to the server, and the local device accessed the server through software developed using PyCharm Professional (version 2023.1, JetBrains, Prague, Czech Republic) and Qt Creator (version 5.0.2, the Qt Company, Espoo, Finland) on a system running Windows 10 (version 22H2, Microsoft, Redmond, WA, USA) with Python (version 3.7.6, Python Software Foundation, Wilmington, DE, USA). This software was designed for downloading the models to perform image acquisition, preprocessing, and cultivar prediction.

2.4.2. Implementation description

The research was conducted on a Windows 10 (version 22H2, Microsoft, Redmond, WA, USA) operating system with Python (version 3.7.6, Python Software Foundation, Wilmington, DE, USA) and the TensorFlow framework, leveraging a Tesla P4 GPU (NVIDIA Corporation, Santa Clara, CA, USA)for computation. These computational resources ensured efficient task execution within the projected timeframe.

The apple cultivar recognition system was developed using TensorFlow (version 2.3.0, TensorFlow, Inc., Mountain View, CA, USA), a widely adopted Python deep learning framework. To maintain experimental rigor and fairness, a concise ten-layer CNN architecture, comprising convolutional and pooling layers, was employed. This design balanced simplicity and performance, minimizing overfitting risks. Both the MobileNetV2 and VGG19 architectures were integrated into the framework, maintaining uniform training parameters. Each model underwent preliminary training for 10 epochs to assess stability and determine the optimal number of training steps. The final training was conducted in 70 epochs, which accurately reflected the model's validation accuracy and loss throughout the process. The Multi-Head Attention mechanism employed consists of eight attention heads, with a hidden feature dimension of 256. The Adam optimizer was adopted for all models, with cross-entropy loss serving as the loss function to optimize classification performance.

2.4.3. Evaluation metrics

The confusion matrix is an essential tool for evaluating classification model performance and for assessing multi-class models. It provides a clear representation of predicted versus actual outcomes for each class, highlighting classification misclassifications [33]. The confusion matrix primarily consists of parameters such as accuracy, precision, recall, and F1 score, facilitating a comprehensive assessment of the model's performance.

These metrics depend on four fundamental values derived from the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Specifically, precision represents the ratio of true positives to all predicted positives, while recall indicates the proportion of true positives among all actual positives. The F1 score, calculated as the harmonic mean of precision and recall, effectively balances these two metrics. It is particularly advantageous in scenarios with class imbalance, as it considers both the precision and recall of correct positive predictions.

In addition, MAE is a regression metric that quantifies the average absolute difference between predicted and actual values. It is computed as the mean of the absolute residuals over all samples, offering an intuitive and interpretable measure of prediction accuracy [34].

These performance indicators can be mathematically defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

$$MAE = \frac{1}{n} \tag{5}$$

Moreover, five-fold cross-validation was employed to assess the model's performance, ensuring robustness given the sample size and specific experimental requirements. This approach divided the dataset into five subsets, with each subset serving as a validation set once, while the others were utilized for training. By systematically rotating through these subsets, this approach mitigated the risks of overfitting and underfitting, thereby improving the model's predictive performance. Furthermore, this robust evaluation method provided a comprehensive assessment of the model's generalization ability on unseen data, ensuring stable and reliable performance in real-world applications.

3. Results and discussion

3.1. Model training

3.1.1 Model performance evaluation

The training dataset comprised approximately 5,500 images, encompassing a diverse range of apple varieties widely cultivated in China, including the Fuji series and newer varieties such as Venus Gold and Yuhua Fushi. Because of the large size of the raw apple images, all images were resized to 224×224 pixels. The t-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction of the training samples was shown in Fig. 4. Before performing t-SNE, PCA was applied to reduce the dimensionality to 50 components, and the first two principal components (PC1 and PC2) explained 41.36% of the variance (PC1: 27.10%, PC2: 14.26%). Most apple varieties exhibited significant confusion in feature distribution, such as Yanfu No.10 and Yanfu No.3. Conversely, varieties such as Holstein and Venus Gold exhibited clearer clustering due to distinct differences in their color features compared to other varieties. In contrast, Nagafu No.2, was more dispersed across the clusters of other varieties, suggesting that it posed a greater challenge for identification.

To evaluate the performance of the CNN models, training accuracy and loss rate graphs were generated and illustrated in Fig. 5.

The CNN series models exhibited relatively smooth fluctuations overall in Fig. 5(a). However, significant variability was observed in the later stages of the CNN model incorporating Multi-Head Attention, which suggested overfitting due to excessive exposure to certain features. This was confirmed by its higher overall loss rate. Similar trends were observed in the MobileNetV2 and VGG19 models, highlighting the impact of the Multi-Head Attention mechanism on performance.

All three foundational models were found to perform well when integrating features with GLCM in Fig. 5(a) and Fig. 5(b). Among them, the combination of MobileNetV2 and GLCM demonstrated the best overall performance, as evidenced by a stable accuracy curve and a closely following loss rate curve. Additionally, integrating GLCM with the VGG19 model significantly reduced the variability of the validation loss curve, indicating that GLCM effectively mitigated the overfitting issues observed with this model.

3.1.2. Comparative analysis of the overall performance of deep learning models

The performance of CNN, MobileNetV2, and VGG19 model architectures for apple cultivar classification was evaluated, with the test results presented in Table 2. Classification accuracy, precision, recall, F1 score, and MAE were selected as the evaluation metrics for these models.

The baseline CNN model achieved excellent performance with an accuracy of 96.46%, precision of 96.48%, recall of 96.46%, and F-score of 96.41%. The MAE for this model was 0.1311, which is relatively high compared to other optimized models [35, 36]. Upon incorporating the Multi-Head Attention mechanism into the CNN model, the accuracy slightly increased to 97.08%, accompanied by improvements in precision, recall, and F1-score. The precision reached 97.10%, the recall was 97.08%, and the F-score was 97.04%. The MAE decreased to 0.1090, reflecting a reduction in error. This suggests that the inclusion of Multi-Head Attention enabled the model to focus on more relevant features, thereby enhancing its classification performance.

On the other hand, integrating GLCM with the CNN model resulted in even better performance. The accuracy increased to 97.92%, with precision, recall, and F-score all showing similar improvements, reaching 97.91%, 97.92%, and 97.86%, respectively. The MAE further decreased to 0.0980, demonstrating that the integration of GLCM not only enhanced the model's classification accuracy but also significantly reduced errors. The GLCM method, which extracts texture features from images, likely enabled the CNN to better capture subtle visual patterns specific to the apple varieties.

The MobileNetV2 model, known for its computational efficiency, also demonstrated strong performance. In its baseline configuration, the model achieved an accuracy of 97.78%, with both precision and recall at 97.78%, and an F-score of 97.74%. The MAE was 0.0696, indicating the best error performance among the baseline models tested. When Multi-Head Attention was applied to MobileNetV2, the accuracy decreased to 87.33%, with precision, recall, and F-score following a similar decline. The MAE increased dramatically to 0.5490. This suggests that, in this case, the Multi-Head Attention mechanism did not provide the expected improvements and negatively impacted the model’s performance. Such complexity may have interfered with the added complexity from the attention mechanism interfered with MobileNetV2's efficient feature extraction, resulting in overfitting or poor generalization.

Integrating GLCM with MobileNetV2 resulted in a significant improvement. The accuracy increased to 98.25%, with precision and recall both reaching 98.29% and 98.25%, respectively. The F-score was 98.20%, while the MAE dropped to 0.0571. GLCM's texture features enhanced MobileNetV2's performance, leading to improvements in both classification accuracy and error reduction.

The VGG19 model, a deeper network known for its efficient feature extraction capabilities, achieved an accuracy of 97.25%, with both precision and recall at 97.25%, and an F-score of 97.22%. The MAE was 0.0992, indicating acceptable performance. However, when Multi-Head Attention was incorporated, the accuracy slightly decreased to 97.01%, with corresponding declines in precision, recall, and F1-score. The MAE decreased to 0.0975, suggesting that while the attention mechanism contributed to error reduction, it did not substantially enhance classification performance.

Consistent with the findings from other models, integrating GLCM with VGG19 led to improved performance. The accuracy reached 97.92%, with both precision and recall at 97.91%, and the F-score at 97.66%. The MAE decreased to 0.0921. These results demonstrate that the texture features extracted by GLCM enhanced VGG19's classification performance and reduced prediction errors.

To conclude, the application of GLCM demonstrated consistent improvements in classification performance across all models, resulting in significant advancements in accuracy and error reduction. The incorporation of Multi-Head Attention, however, produced mixed results, improving some models, such as CNN, but significantly degrading MobileNetV2's performance. Overall, GLCM was a valuable optimization technique, enhancing classification accuracy and reducing MAE, particularly for MobileNetV2 and CNN.

3.2. Model comparison of algorithms within the same series

Heatmap visualizations confirmed that models integrating GLCM and the Multi-Head Attention mechanism, built upon CNN, MobileNetV2, and VGG19 architectures, effectively focused on relevant image regions during feature extraction. These visualizations demonstrated that each model achieved satisfactory classification performance across most apple varieties.

3.2.1. CNN series cultivar detection models

The prediction results for eleven varieties of apples were shown in Fig. 6 from three models: (a) CNN model, (b) CNN+Multi-Head Attention model, and (c) CNN+GLCM model. The CNN model demonstrated generally high accuracy, with categories such as Changhong (95.93%) and Red Fuji (95.07%) achieving high classification accuracy. Other categories, such as Chengji No.1, Holstein, Lifu No.2, Miyakuj, Venus Gold, and Yuhua Fushi, also achieved near 100% accuracy, indicating solid performance in distinguishing these apple varieties. However, there were exhibited considerable misclassification, such as Nagafu No.2 (76.52%) and Yanfu No.3, which showed lower accuracy and higher misclassification rates. This suggested that the CNN model struggled to distinguish certain varieties which had visually similar features.

The addition of Multi-Head Attention improved the model's performance, particularly in categories such as Changhong (97.29%) and Red Fuji (96.06%), where accuracy improved compared to the CNN model. It also enhanced the overall prediction reliability for categories such as Chengji No.1 and Holstein, which now exhibited 100% accuracy. The Multi-Head Attention mechanism appears to have enabled the model to focus more effectively on key features, enhancing its classification performance, particularly in complex scenarios. However, despite the improved performance over the CNN model, some categories, such as Nagafu No.2 (79.13%), still showed confusion, though the performance was better than that of the CNN model.

The CNN + GLCM model demonstrated a significant performance improvement, particularly for categories such as Changhong (97.74%) and Red Fuji (95.07%), achieving high accuracy consistently. The integration of GLCM enhanced the model's ability to capture texture features, enabling it to more effectively differentiate between varieties with similar visual characteristics. The confusion for Nagafu No.2 (76.52%) persisted, but the accuracy improved slightly compared to the CNN model, indicating that GLCM contributed to better differentiation of such varieties. Overall, the combination of CNN and GLCM resulted in improved performance, particularly for texture-based features, although some categories still exhibited minor misclassifications.

In conclusion, the CNN model demonstrated strong classification performance, which was further enhanced by integrating Multi-Head Attention and GLCM, resulting in improved accuracy, especially for categories with similar visual features. However, some categories, such as Nagafu No.2, still posed classification challenges.

3.2.2. Analysis of MobileNetV2 series cultivar detection models

The original MobileNetV2 model demonstrated strong performance in predicting varieties, particularly for Changhong, Chengji No.1, and Holstein, achieving accuracies of 99.55%, 99.46%, and 100%, respectively (Fig. 7). However, the model struggled with varieties such as Nagafu No.2, Red Fuji, and Yanfu No.10, with accuracies ranging from 0% to 5%. The model excelled in identifying distinct varieties but struggled with those with less distinctive features, particularly Nagafu No.2. Despite these challenges, the model effectively distinguished the most distinctive apple varieties but faced difficulty classifying more complex or visually similar ones.

Integrating Multi-Head Attention into the MobileNetV2 model, its performance improved, particularly on more challenging varieties. This enhancement helped the model better capture intricate patterns, particularly for varieties such as Nagafu No.2 and Red Fuji, where accuracy increased. While the model's performance on Changhong decreased slightly to 97.74%, varieties such as Yanfu No.10 and Yanfu No.3 showed notable improvements, with accuracies of 91.79% and 83.33%, respectively. Nevertheless, the model continued to face challenges with certain varieties, such as Miyakuj and Venus Gold, where misclassifications persisted.

The final version, which combined MobileNetV2 with GLCM, showed the most significant improvements in classification accuracy. This hybrid model performed exceptionally well, achieving 100% accuracy on Changhong, 99.46% on Chengji No.1, and 98.98% on Holstein. It also excelled in identifying varieties such as Red Fuji (97.44%) and Venus Gold (100%). “The GLCM-based approach, focusing on texture features, helped differentiate visually similar varieties more effectively, improving precision, especially for challenging varieties like Yanfu No.3 (96.15%).

The combination of MobileNetV2 with the Multi-Head Attention mechanism and GLCM resulted in noticeable performance variations. While MobileNetV2 with GLCM demonstrated strong classification abilities, the inclusion of Multi-Head Attention resulted in a decline in overall performance. This decline was especially noticeable in the identification of varieties such as Nagafu No.2, Red Fuji, and Yanfu No.3. MobileNetV2, designed for efficiency with depthwise separable convolutions, is lightweight, but the addition of Multi-Head Attention, which captures global context, and increased computational complexity. This added complexity likely hindered performance, particularly on smaller datasets, and may have contributed to overfitting.

The inclusion of GLCM notably enhanced classification performance, particularly for varieties such as Nagafu No.2 and Yanfu No.3. By extracting textural features, GLCM improved MobileNetV2's ability to capture subtle texture variations that might have been overlooked by the model, leading to better classification accuracy.

3.2.3. Analysis of VGG19 series cultivar detection models

VGG19 exhibited robust performance in classifying apple varieties, consistently achieving high accuracy, as illustrated in Fig. 8. The model did not exhibit overfitting for most varieties, achieving prediction accuracies of 98.64% for Changhong and 99.46% for Chengji No. 1, demonstrating excellent performance. Holstein reached 98.98%

accuracy, while Miyakuj attained 99.48%, further showcasing the robustness of VGG19. Although the classification accuracy for Lifu No. 2 slightly decreased to 96.02%, it remained high overall. However, Nagafu No. 2 showed a relatively lower accuracy of 81.74%, revealing challenges in recognizing this cultivar. Despite this, VGG19 continued to perform strongly across most varieties, handling classification tasks with only minor difficulties for a few specific cases.

The addition of the Multi-Head Attention mechanism significantly enhanced the performance of VGG19 in classifying apple varieties. While the accuracy for Changhong decreased slightly to 97.32%, it remained high. Chengji No. 1 and Holstein maintained perfect classification at 100%, and Miyakuj also achieved 100%, highlighting the effectiveness of the Multi-Head Attention mechanism. Lifu No. 2 showed stability with an accuracy of 99.50%. However, Nagafu No. 2 experienced a 4% drop in accuracy to 74.00%, indicating that the mechanism might have introduced interference for some varieties. Other varieties, such as Red Fuji and Venus Gold, saw modest improvements, with accuracies of 98.70% and 98.00%, respectively. The classification accuracy for Yanfu No. 10 and Yanfu No. 3 remained mostly unchanged at 98.50% and 92.50%. Yuhua Fushi continued to perform flawlessly with an accuracy of 100%.

After incorporating GLCM into the VGG19 model, the performance for many varieties showed improvement. Nagafu No.2 saw a significant increase to 84.17%, up from 81.74% in the original VGG19 model. The model's ability to correctly identify

"Red Fuji" remained consistent at 98.22%. Yanfu No.3 experienced a modest improvement to 94.06%, with only 0.60% misclassified. Meanwhile, varieties such as Chengji No.1 (100%) and Holstein (100%) showed no major changes, as the addition of texture-based features from GLCM did not impact these already high-performing categories. This indicates that GLCM had a substantial effect on the more challenging varieties.

The performance of the VGG19 model changed notably with the addition of Multi-Head Attention and GLCM, highlighting the interplay between these mechanisms and the network's structure. VGG19's deep architecture, with its multiple convolutional layers and pooling operations, effectively captured hierarchical image features. However, the introduction of Multi-Head Attention sometimes caused interference, particularly with varieties such as Nagafu No. 2, where the model struggled to distinguish subtle differences. This may have been due to an overemphasis on less relevant features. On the other hand, GLCM improved texture-based feature extraction, boosting Nagafu No. 2's accuracy. This improvement likely stemmed from GLCM's ability to capture subtle texture differences that VGG19 alone might have missed, which was especially beneficial for varieties with more intricate surface textures where color alone was insufficient for accurate classification.

GLCM provides complementary information about the spatial relationship between pixels, helping the model discern finer details that distinguish varieties such as Nagafu No. 2. This texture-based enhancement aids the model in focusing on important patterns that might be overlooked in color-based feature extraction, resulting in improved classification accuracy for varieties with complex textures.

3.3 Model comparison of algorithms across different series

Among the nine models assessed, MobileNetV2+GLCM achieved the highest performance, achieving an overall classification accuracy of 98.25% with an MAE of 0.0571. It was followed by VGG19+GLCM, which achieved a classification accuracy of 97.92% and an MAE of 0.0921. The training-related charts show that the introduction of GLCM positively impacted all three baseline models, significantly enhancing their performance. For instance, the inclusion of GLCM improved the MobileNetV2 model's accuracy to 98.25%, outperforming the baseline model. Similarly, the CNN model saw an increase in accuracy to 97.92%, maintaining strong performance compared to the version without optimization techniques.

In contrast, the Multi-Head Attention mechanism exhibited a selective effect. After its introduction, slight improvements were observed in the performance of CNN, with accuracy rising from 96.46% to 97.08%. However, the Multi-Head Attention mechanism had a pronounced negative impact on the MobileNetV2 model, leading to a decline in performance. A slight decrease in performance was also noted when Multi-Head Attention was applied to the VGG19 model. The confusion matrix for MobileNetV2 with Multi-Head Attention revealed a significant drop in prediction accuracy across several varieties, resulting in distorted predictions. Notable misclassifications occurred, particularly among closely related varieties such as Yanfu

No. 10 and Yanfu No. 3. This suggested that Multi-Head Attention did not effectively enhance MobileNetV2's performance, potentially causing overfitting or feature loss.

To analyze why the Multi-Head Attention mechanism had adverse effects on MobileNetV2, we considered both the model's parameter count and the nature of the attention mechanism. MobileNetV2 is a lightweight model with only 2,263,108 parameters, far fewer than CNN (53,767,748) and VGG19 (26,585,163). Designed for efficiency, MobileNetV2 was optimized to work with limited computational resources, making it highly effective for tasks that require fewer parameters. However, the introduction of Multi-Head Attention, which adds complexity and increases computational demands, may have disrupted this balance, negatively impacting performance.

The introduction of the Multi-Head Attention mechanism added computational complexity and increased the model's capacity, which may have led to overfitting or instability, especially in lightweight models such as MobileNetV2. These effects were particularly noticeable in cases where the dataset or training process was sensitive to parameter adjustments. The observed training fluctuations and the decline in recognition accuracy indicated that the added complexity disrupted MobileNetV2's balance, ultimately reducing its efficiency and performance. In contrast, models with higher parameter capacities, such as CNN, better accommodated the additional layers introduced by attention mechanisms, which explains their improved performance.

Finally, we observed that the recognition accuracy for the Nagafu No. 2 cultivar remained around 80% across the nine models, with frequent misclassifications as Lifu No. 2, Miyakuj, or Venus Gold. Similarly, the recognition accuracy for Yanfu No. 3 clustered around 92%, indicating a high degree of similarity between Yanfu No. 3 and other apple varieties. The persistent misclassification of Nagafu No. 2, a key maternal parent in Fuji-lineage breeding, is likely attributable to its similar genomic traits, which lead to phenotypic ambiguities in standard RGB imaging and, consequently, contribute to classification errors.

To further improve cultivar recognition accuracy, plan to implement data augmentation techniques such as rotation, scaling, flipping, and color adjustments. These methods will introduce variability, increase dataset diversity, and enhance the model's generalization capability. Additionally, the loss function will be modified by incorporating class weights to place greater emphasis on reducing misclassifications of underrepresented varieties.

We will also integrate additional features, such as color histograms, shape descriptors, and metadata, to better differentiate between varieties. Furthermore, to enhance the model's feature extraction capabilities, we will explore advanced techniques such as Adaptive Feature Fusion and Self-Supervised Learning (SSL). These methods will enable more effective high-level feature learning, ultimately improving the model’s accuracy in recognizing apple varieties with genetic relationships, such as Nagafu No. 2.

4. Conclusion

This research presented an innovative approach to classifying and recognizing apple varieties using deep learning techniques, investigating the integration of the Multi-Head Attention mechanism and GLCM optimization across three distinct architectures: CNN, MobileNetV2, and VGG19. By combining image processing with machine learning, the study significantly improved the accuracy and efficiency of apple cultivar identification. Focusing on eleven popular apple varieties in China, the optimized models consistently achieved classification accuracies above 95%, with some varieties exceeding 98%.

MobileNetV2+GLCM achieved the highest accuracy of 98.25%, demonstrating the effectiveness of combining traditional methods with advanced image processing techniques. Compared to the study in [17], which trained seven types of CNN models on a dataset comprising 5,808 images from 10 different Turkish apple varieties and identified DenseNet as the best-performing model with an accuracy of 97.48%, and the study in [37], which employed MobileNetV2 and EfficientNetV2B0 to classify six Turkish apple varieties from a dataset of 120 images, where EfficientNetV2B0 combined with GLCM and Color-Space achieved the highest accuracy of 98.33%, our research covered a broader range of apple varieties and utilized a significantly larger dataset. Moreover, the MobileNetV2 model offers advantages over EfficientNet and DenseNet in terms of computational efficiency and suitability for deployment in resource-constrained environments.

However, integrating Multi-Head Attention into lightweight models such as MobileNetV2 resulted in training instability and potential overfitting. In contrast, its incorporation into CNN and VGG19 resulted in moderate improvements, with accuracies exceeding 97.08%, demonstrating the selective advantages of advanced feature enhancement[38, 39]. These findings provide valuable insights for developing models for fruit species recognition and surface defect detection in agricultural products such as citrus, pears, and peaches. By enhancing the precision and efficiency of agricultural technology, these models have the potential to be applied across a wide range of agricultural applications, from quality control to automated sorting.

Our research focuses on identifying 11 apple varieties commonly cultivated in China, contributing to the development of deep learning models for apple recognition. The study holds significant potential for integration into commercial apple sorting lines. These models can complement existing systems for quality recognition, thereby enhancing both sorting efficiency and accuracy. Furthermore, the system is d designed for deployment on mobile devices, enabling farmers, traders, and consumers to conveniently use the software and models for real-time cultivar identification and quality assessment.

Despite these advantages, the apple images used in our study were collected under controlled conditions using a professional acquisition platform. In practical applications, factors such as lighting, vibrations, and the presence of foreign objects may affect image quality. To improve model robustness and generalizability, it is essential to incorporate images captured in real-world conditions, such as those obtained from conveyor belts, storage environments, and varying lighting scenarios. Additionally, apple phenotypes vary with maturity, causing certain varieties to exhibit striking similarities at specific growth stages while diverging significantly at others, which poses challenges for classification, sorting, and grading models.

To further validate the efficacy of the proposed approach, future research should expand both the dataset and the diversity of varieties to enhance broader applicability. We will focus on optimizing the baseline models to improve robustness, building upon the current findings. Additionally, we will explore the incorporation of additional features, such as color, shape, and temporal information, to enhance grading accuracy. Furthermore, real-time applications will be developed to improve automation and efficiency in apple cultivar recognition systems.

Acknowledgments

This research was funded by the Key R&D Project of Jiangsu Province (BE2022363), the National Key R&D Program of China (2022YFD2100604), the National Natural Science Foundation of China (W2412103), the Natural Science Foundation of Jiangsu Province (BK20220524), and the Senior Talent Program of Jiangsu University (22JDG015).

References

Fan, S., Liang, X., Huang, W., Zhang, V. J., Pang, Q., He, X., Li, L., & Zhang, C. 2022. Real-time defects detection for apple sorting using NIR cameras with pruning-based YOLOV4 network. *Comput. Electron. Agric.*, *193*, Article 106715. https://doi.org/10.1016/j.compag.2022.106715

Musacchi, S., & Serra, S. 2018. Apple fruit quality: Overview on pre-harvest factors. *Scientia Hortic.*, *234*, 409-430. https://doi.org/10.1016/j.scienta.2017.12.057

Wu, D., Lv, S., Jiang, M., & Song, H. 2020. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.*, *178*, Article 105742. https://doi.org/10.1016/j.compag.2020.105742

Adhinata, F. D., Wahyono, & Sumiharto, R. 2024. A comprehensive survey on weed and crop classi fi cation using machine learning and deep learning. *Artif. Intell. Agric*, *13*, 45-63. https://doi.org/10.1016/j.aiia.2024.06.005

Liang, X., Zhang, R., Gleason, M. L., & Sun, G. 2022. Sustainable Apple Disease Management in China: Challenges and Future Directions for a Transforming Industry. *Plant Dis.*, *106*(3), 786-799. https://doi.org/10.1094/pdis-06-21-1190-fe

Gao, F., Fu, L., Zhang, X., Majeed, Y., Li, R., Karkee, M., & Zhang, Q. 2020. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Comput. Electron. Agric.*, *176*, Article 105634. https://doi.org/10.1016/j.compag.2020.105634

Knott, M., Perez-Cruz, F., & Defraeye, T. 2023. Facilitated machine learning for image-based fruit quality assessment. *J. Food Eng.*, *345*, Article 111401. https://doi.org/10.1016/j.jfoodeng.2022.111401

Zheng, S., Gao, P., Zhang, J., Ma, Z., & Chen, S. 2024. A precise grape yield prediction method based on a modified DCNN model. *Comput. Electron. Agric.*, *225*, Article 109338. https://doi.org/10.1016/j.compag.2024.109338

Yag, I., & Altan, A. 2022. Artificial Intelligence-Based Robust Hybrid Algorithm Design and Implementation for Real-Time Detection of Plant Diseases in Agricultural Environments. *Biology*, *11*(12), Article 1732. https://doi.org/10.3390/biology11121732

Wakchaure, G. C., Nikam, S. B., Barge, K. R., Kumar, S., Meena, K. K., Nagalkar, V. J., Choudhari, J. D., Kad, V. P., & Reddy, K. S. 2024. Maturity stages detection prototype device for classifying custard apple (Annona squamosa L) fruit using image processing approach. *Smart Agric. Technol.*, *7*, Article 100394. https://doi.org/10.1016/j.atech.2023.100394

Miranda, J. C., Arno, J., Gene-Mola, J., Lordan, J., Asin, L., & Gregorio, E. 2023. Assessing automatic data processing algorithms for RGB-D cameras to predict fruit size and weight in apples. *Comput. Electron. Agric.*, *214*, Article 108302. https://doi.org/10.1016/j.compag.2023.108302

Wang, C., Liu, S., Wang, Y., Xiong, J., Zhang, Z., Zhao, B., Luo, L., Lin, G., & He, P. 2022. Application of Convolutional Neural Network-Based Detection Methods in Fresh Fruit Production: A Comprehensive Review. *Front. Plant Sci.*, *13*, Article 868745. https://doi.org/10.3389/fpls.2022.868745

Lei, L., Yang, Q., Yang, L., Shen, T., Wang, R., & Fu, C. 2024. Deep learning implementation of image segmentation in agricultural applications: a comprehensive review. *Artif. Intell. Rev*, *57*(6), Article 149. https://doi.org/10.1007/s10462-024-10775-6

Ji, W., Wang, J., Xu, B., & Zhang, T. 2023. Apple Grading Based on Multi-Dimensional View Processing and Deep Learning. *Foods*, *12*(11), Article 2117. https://doi.org/10.3390/foods12112117

Wang, Z., Jin, L., Wang, S., & Xu, H. 2022. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.*, *185*, Article 111808. https://doi.org/10.1016/j.postharvbio.2021.111808

Ji, W., Zhang, T., Xu, B., & He, G. 2024. Apple recognition and picking sequence planning for harvesting robot in a complex environment. *J. Agric. Eng.*, *55*(1), Article 1549. https://doi.org/10.4081/jae.2024.1549

Taner, A., Mengstu, M. T., Selvi, K. C., Duran, H., Gur, I., & Ungureanu, N. 2024. Apple Varieties Classification Using Deep Features and Machine Learning. *Agriculture-Basel*, *14*(2), Article 252. https://doi.org/10.3390/agriculture14020252

Xue, G., Liu, S., & Ma, Y. 2020. A hybrid deep learning-based fruit classification using attention model and convolution autoencoder. *Complex & Intelligent Systems*, *9*, 2209–2219. https://doi.org/10.1007/s40747-020-00192-x

Li, J., Xie, S., Chen, Z., Liu, H., Kang, J., Fan, Z., & Li, W. 2020. A Shallow Convolutional Neural Network for Apple Classification. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3002882 (IEEE)

Guo, Z., Zou, Y., Sun, C., Jayan, H., Jiang, S., El-Seedi, H. R., & Zou, X. 2024. Nondestructive determination of edible quality and watercore degree of apples by portable Vis/NIR transmittance system combined with CARS-CNN. *J. Food Meas. Charact.*, 4058–4073, (2024). https://doi.org/10.1007/s11694-024-02476-z

Xu, B., Cui, X., Ji, W., Yuan, H., & Wang, J. 2023. Apple Grading Method Design and Implementation for Automatic Grader Based on Improved YOLOv5. *Agriculture-Basel*, *13*(1), Article 124. https://doi.org/10.3390/agriculture13010124

Wang, A., Qian, W., Li, A., Xu, Y., Hu, J., Xie, Y., & Zhang, L. 2024. NVW-YOLOv8s: An improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages. *Comput. Electron. Agric.*, *219*, Article 108833. https://doi.org/10.1016/j.compag.2024.108833

Lu, S., Chen, W., Zhang, X., & Karkee, M. 2022. Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation. *Comput. Electron. Agric.*, *193*, Article 106696. https://doi.org/10.1016/j.compag.2022.106696

Liu, Y., Wang, Z., Wang, R., Chen, J., & Gao, H. 2023. Flooding-based MobileNet to identify cucumber diseases from leaf images in natural scenes. *Comput. Electron. Agric.*, *213*, Article 108166. https://doi.org/10.1016/j.compag.2023.108166

Bansal, M., Kumar, M., Sachdeva, M., & Mittal, A. 2021. Transfer learning for image classification using VGG19: Caltech-101 image data set. *J. Ambient Intell. Humaniz. Comput.*, *14*, 3609–3620. https://doi.org/10.1007/s12652-021-03488-z

Sun, J., Zhang, L., Zhou, X., Yao, K., Tian, Y., & Nirere, A. 2021. A method of information fusion for identification of rice seed varieties based on hyperspectral imaging technology. *J. Food Process Eng.*, *44*(9), Article e13797. https://doi.org/10.1111/jfpe.13797

Thakur, P. S., Sheorey, T., & Ojha, A. 2023. VGG-ICNN: A Lightweight CNN model for crop disease identification. *Multimed. Tools Appl.*, *82*(1), 497-520. https://doi.org/10.1007/s11042-022-13144-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, *30*, Article 1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Iqbal, N., Mumtaz, R., Shafi, U., & Zaidi, S. M. H. 2021. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. *PeerJ Comput. Sci.*, Article e536. https://doi.org/10.7717/peerj-cs.536

BEEDATA. (2024). *Yunguo: 2024 China Apple Industry Data Analysis Report*. https://www.weihengag.com/home/article/detail/id/23912

(SAC), S. A. o. C. (2008). *Fresh apples (GB/T 10651-2008).*

Jia, W., Zhang, Z., Shao, W., Hou, S., Ji, Z., Liu, G., & Yin, X. 2021. FoveaMask: A fast and accurate deep learning model for green fruit instance segmentation. *Comput. Electron. Agric.*, *191*, Article 106488. https://doi.org/10.1016/j.compag.2021.106488

Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., & Zheng, Y. 2020. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.*, *172*, Article 105380. https://doi.org/10.1016/j.compag.2020.105380

Shafik, W., Tufail, A., Liyanage, C. D. S., & Apong, R. A. A. H. M. 2023. Using a novel convolutional neural network for plant pests detection and disease classification. *J. Sci. Food Agric.*, *103*(12), 5849-5861. https://doi.org/10.1002/jsfa.12700

Rai, N., Zhang, Y., Ram, B. G., Schumacher, L., Yellavajjala, R. K., Bajwa, S., & Sun, X. 2023. Applications of deep learning in precision weed management: A review. *Comput. Electron. Agric.*, *206*, Article 107698. https://doi.org/10.1016/j.compag.2023.107698

Pacal, I., Kunduracioglu, I., Alma, M. H., Deveci, M., Kadry, S., Nedoma, J., Slany, V., & Martinek, R. 2024. A systematic review of deep learning techniques for plant diseases. *Artif. Intell. Rev*, *57*(11), Article 304. https://doi.org/10.1007/s10462-024-10944-7

Kilicarslan, S., Donmez, E., & Kilicarslan, S. 2024. Identification of apple varieties using hybrid transfer learning and multi-level feature extraction. *Eur. Food Res. Technol.*, *250*(3), 895-909. https://doi.org/10.1007/s00217-023-04436-1

Li, Y., Yao, T., Pan, Y., & Mei, T. 2023. Contextual Transformer Networks for Visual Recognition. *IEEE Trans. Geosci. Remote Sens.*, *45*(2), 1489-1500.

https://doi.org/10.1109/tpami.2022.3164083

Zhou, Y., Lao, C., Yang, Y., Zhang, Z., Chen, H., Chen, Y., Chen, J., Ning, J., & Yang, N. 2021. Diagnosis of winter-wheat water stress based on UAV-borne multispectral image texture and vegetation indices. *Agric. Water Manag.*, *256*, Article 107076. https://doi.org/10.1016/j.agwat.2021.107076

Figure Captions

Fig. 1. Schematic of the steps involved in intelligent identification of apples based on deep learning.

Fig. 2. Images of the eleven major apple cultivars in the detection dataset created by Jiangsu University and supported by China Agriculture Research Systems for the Apple Industry.

Fig. 3. Apple classification algorithm based on deep learning, including (a) Apple dataset, (b) Multi-Head Attention mechanism, (c) Detection result, (d) CNN algorithm, (e) VGG19 algorithm, and (f) MobileNet-V2 algorithm.

Fig. 4. t-SNE Dimensionality Reduction Distribution of Apple Samples.

Fig. 5. Comparison of Training Accuracy and Validation Accuracy of the Eight Proposed Models.

Fig. 6. Prediction of eleven apple cultivars using three CNN-based models, including (a) CNN model, (b) CNN+Multi-Head Attention model, and (c) CNN+GLCM model.

Fig. 7. Prediction of eleven apple cultivars using three MobileNet V2-based models, including (a) MobileNet V2 model, (b) MobileNet V2+Multi-Head Attention model, and (c) MobileNet V2+GLCM model.

Fig. 8. Prediction of eleven apple cultivars using two VGG19-based models, including (a) VGG19 model, (b) VGG19+GLCM model, and (c) VGG19+Multi-Head Attention model.

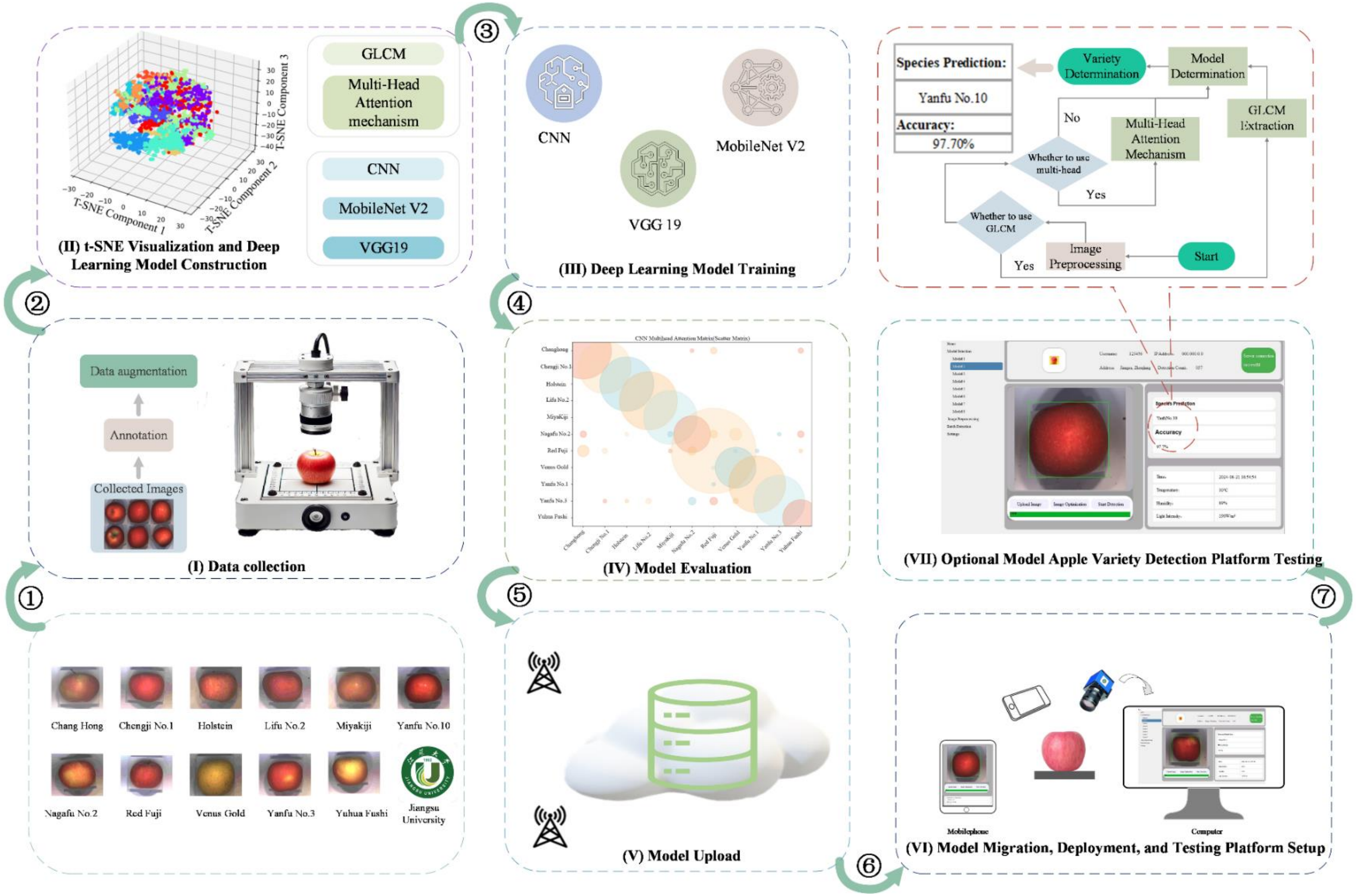


Fig. 1. Schematic of the steps involved in intelligent identification of apples based on deep learning.

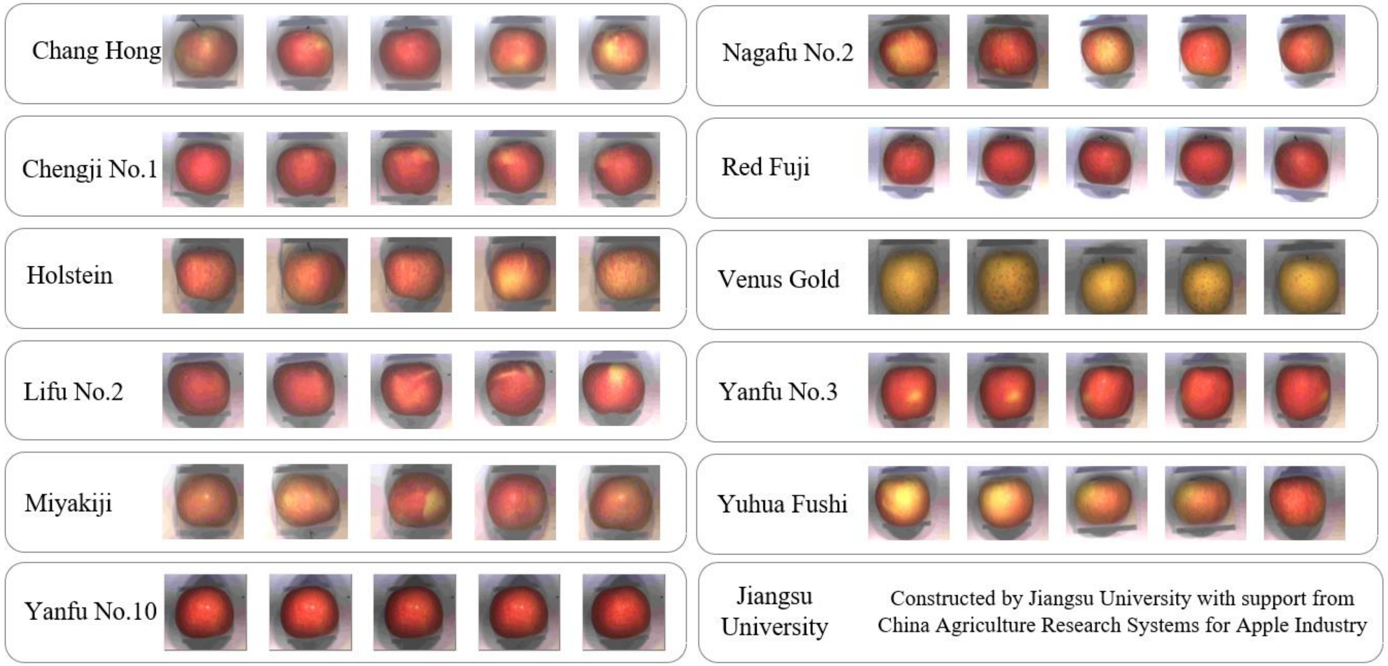


Fig. 2. Images of the eleven major apple cultivars in the detection dataset created by Jiangsu University and supported by China Agriculture Research Systems for the Apple Industry.

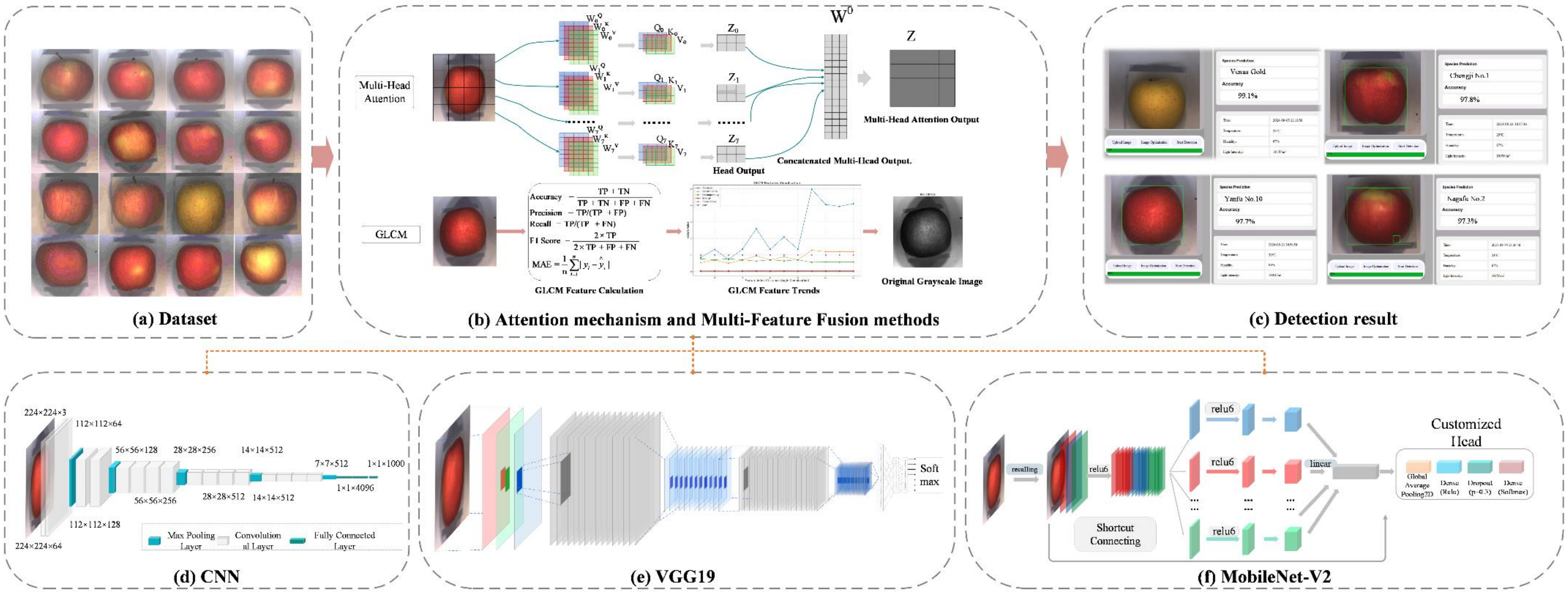


Fig. 3. Apple classification algorithm based on deep learning, including (a) Apple dataset, (b) Multi-Head Attention mechanism, (c) Detection result, (d) CNN algorithm, (e) VGG19 algorithm, and (f) MobileNet-V2 algorithm.

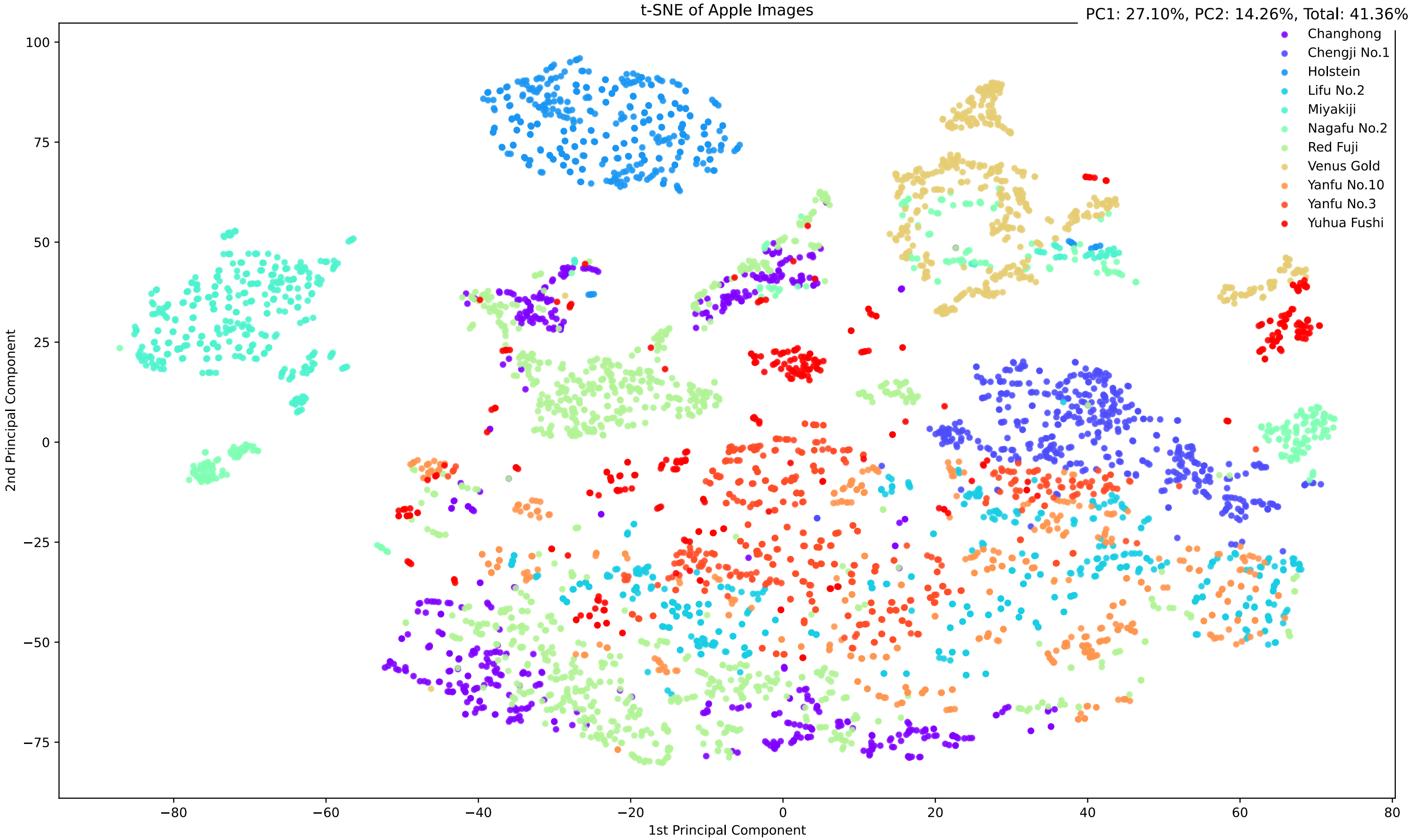


Fig. 4. t-SNE dimensionality reduction distribution of apple samples.

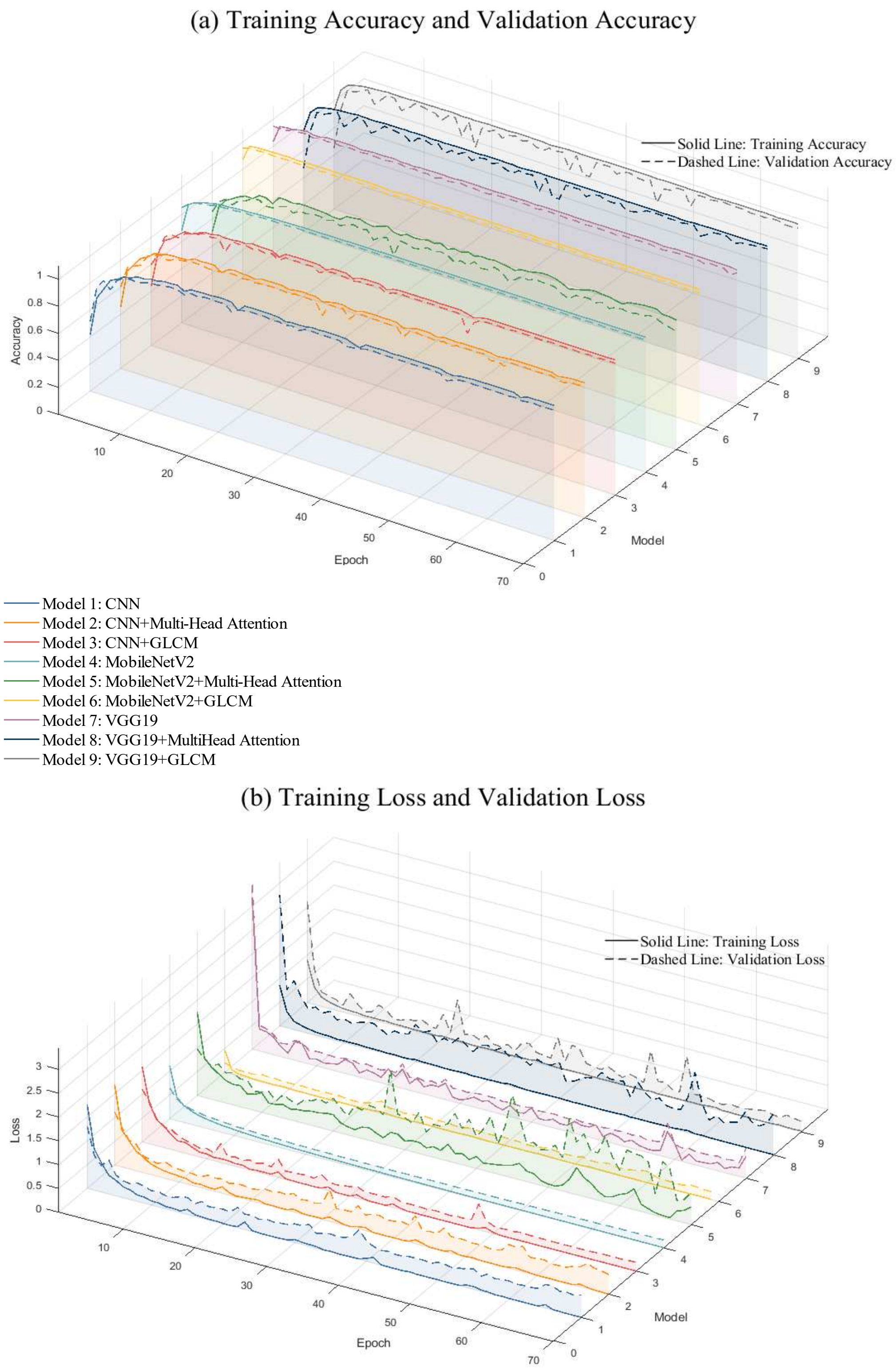
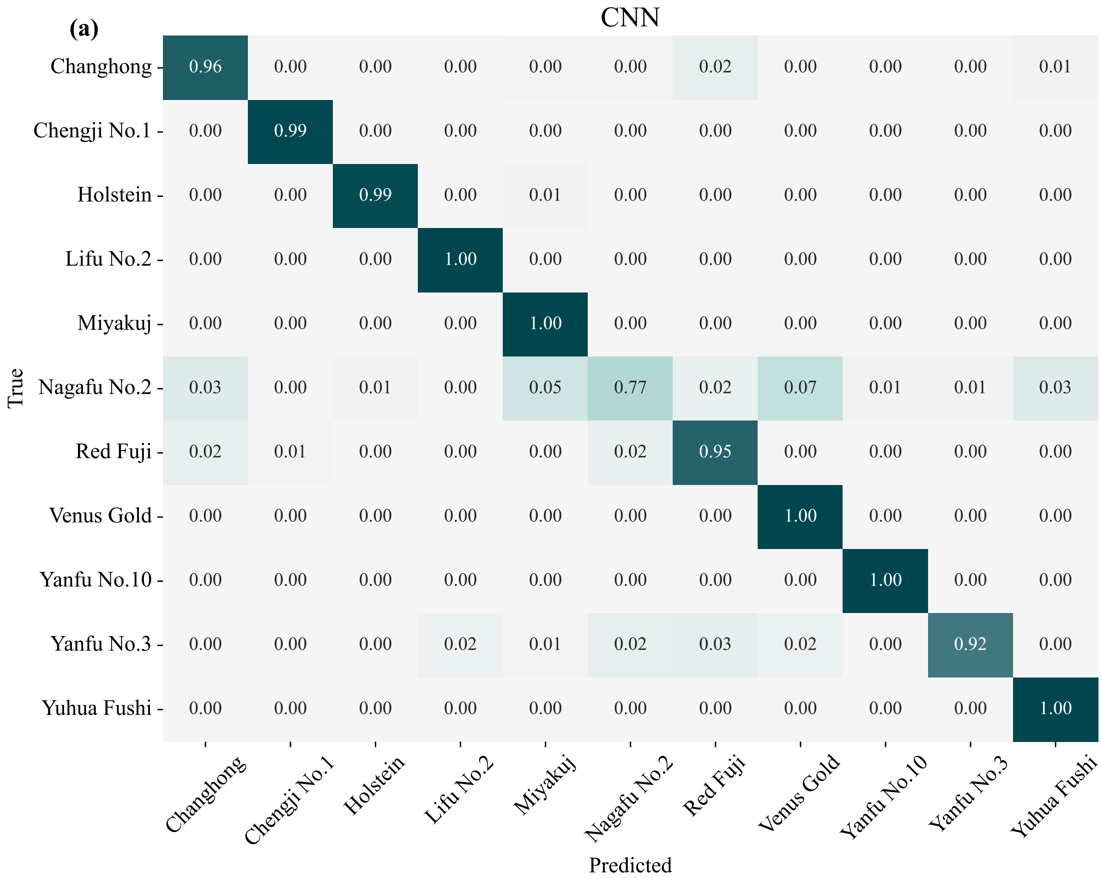
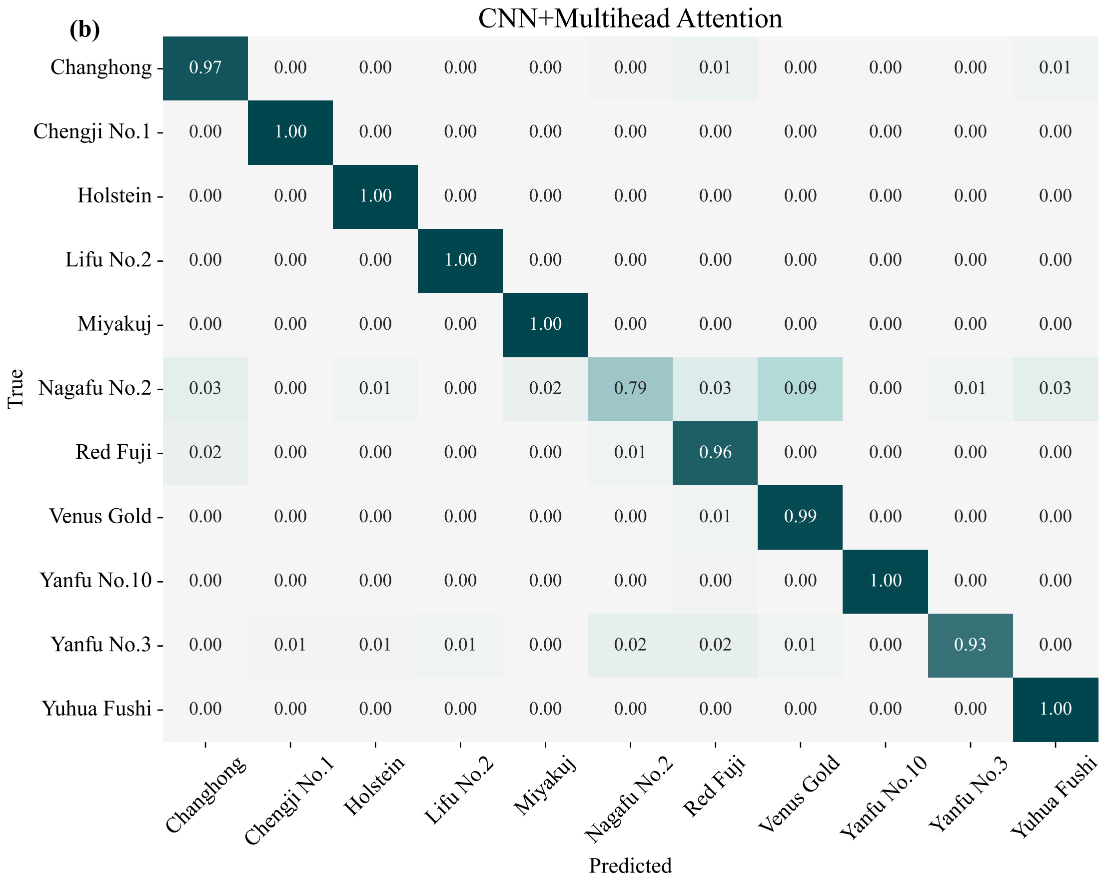


Fig. 5. Comparison of Training Accuracy and Validation Accuracy of the eight proposed models.


(a)
CNN
True
Changhong 0.96 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.01
Chengji No.1 0.00 0.99 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Holstein 0.00 0.00 0.99 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00
Lifu No.2 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Miyakuj 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00
Nagafu No.2 0.03 0.00 0.01 0.00 0.05 0.77 0.02 0.07 0.01 0.01 0.03
Red Fuji 0.02 0.01 0.00 0.00 0.00 0.02 0.95 0.00 0.00 0.00 0.00
Venus Gold 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
Yanfu No.10 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00
Yanfu No.3 0.00 0.00 0.00 0.02 0.01 0.02 0.03 0.02 0.00 0.92 0.00
Yuhua Fushi 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
Changhong
Chengji No.1
Holstein
Lifu No.2
Miyakuj
Nagafu No.2
Red Fuji
Venus Gold
Yanfu No.10
Yanfu No.3
Yuhua Fushi
Predicted


(b)
CNN+Multihead Attention
True
Changhong 0.97 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.01
Chengji No.1 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Holstein 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Lifu No.2 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Miyakuj 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00
Nagafu No.2 0.03 0.00 0.01 0.00 0.02 0.79 0.03 0.09 0.00 0.01 0.03
Red Fuji 0.02 0.00 0.00 0.00 0.00 0.01 0.96 0.00 0.00 0.00 0.00
Venus Gold 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.99 0.00 0.00 0.00
Yanfu No.10 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00
Yanfu No.3 0.00 0.01 0.01 0.01 0.00 0.02 0.02 0.01 0.00 0.93 0.00
Yuhua Fushi 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
Changhong
Chengji No.1
Holstein
Lifu No.2
Miyakuj
Nagafu No.2
Red Fuji
Venus Gold
Yanfu No.10
Yanfu No.3
Yuhua Fushi
Predicted

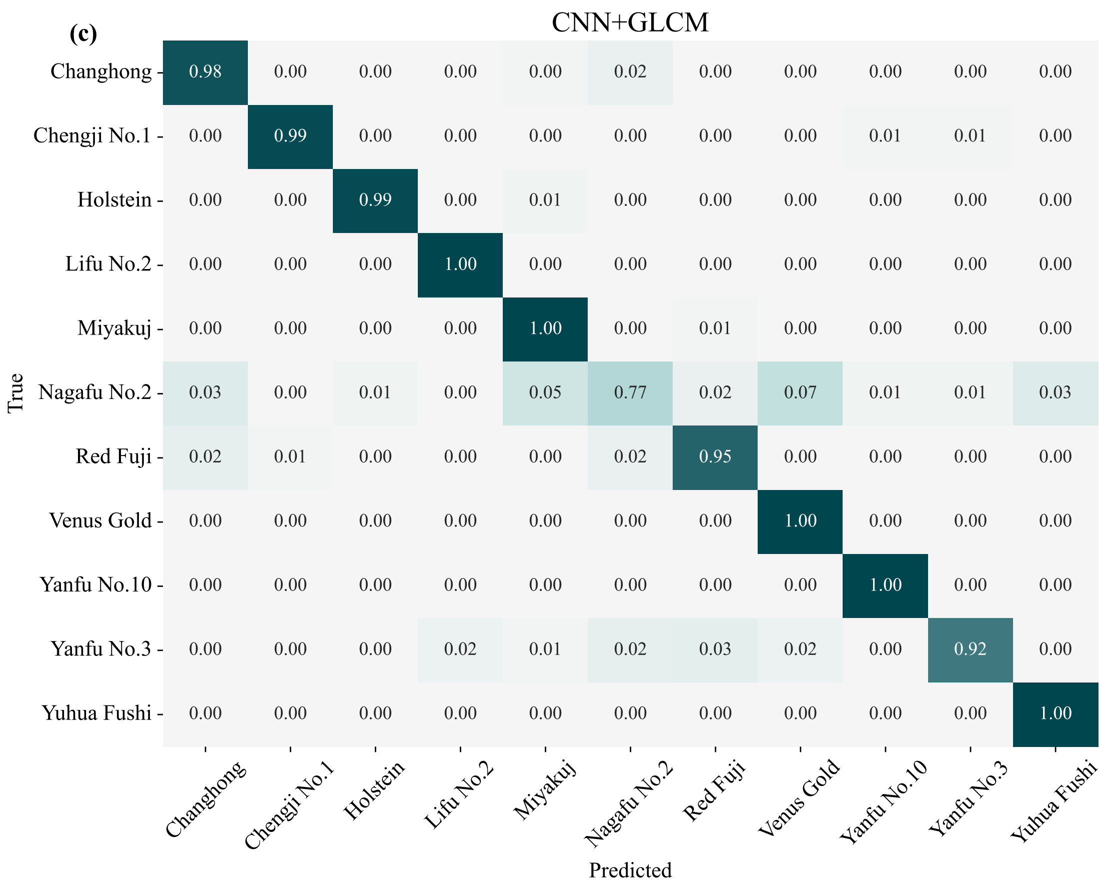
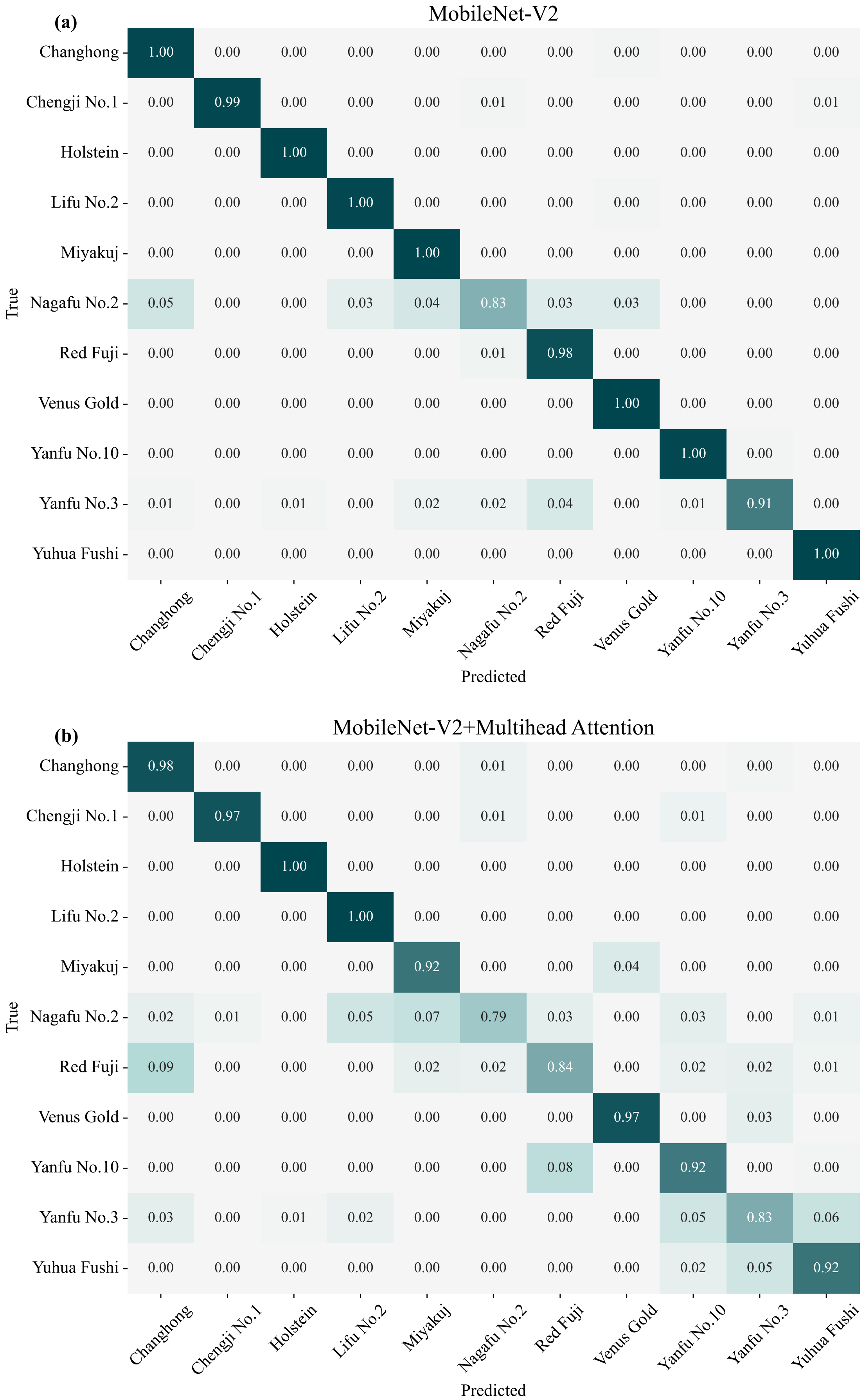


Fig. 6. Prediction of eleven apple cultivars using three CNN-based models, including (a) CNN model, (b) CNN+Multi-Head Attention model, and (c) CNN+GLCM mode.


(a)
MobileNet-V2
True
Predicted
Changhong
Chengji No.1
Holstein
Lifu No.2
Miyakuj
Nagafu No.2
Red Fuji
Venus Gold
Yanfu No.10
Yanfu No.3
Yuhua Fushi
(b)
MobileNet-V2+Multihead Attention
True
Predicted

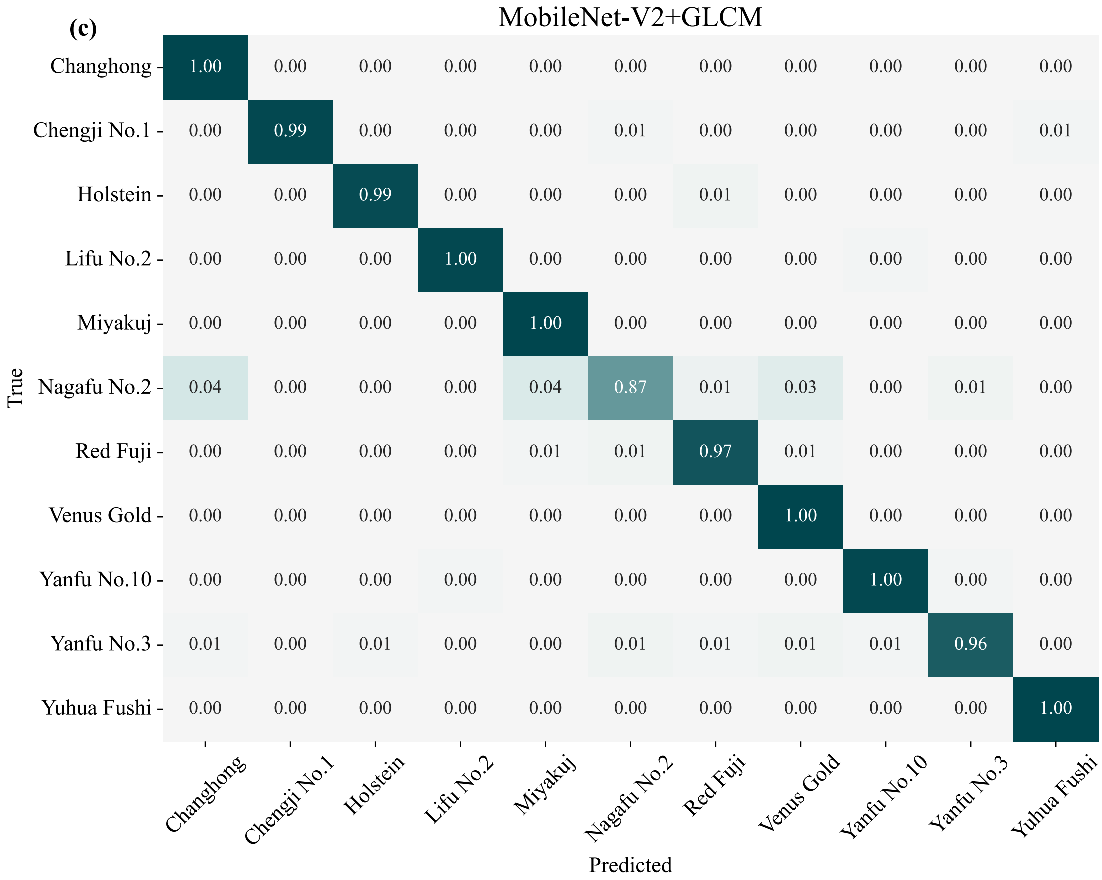
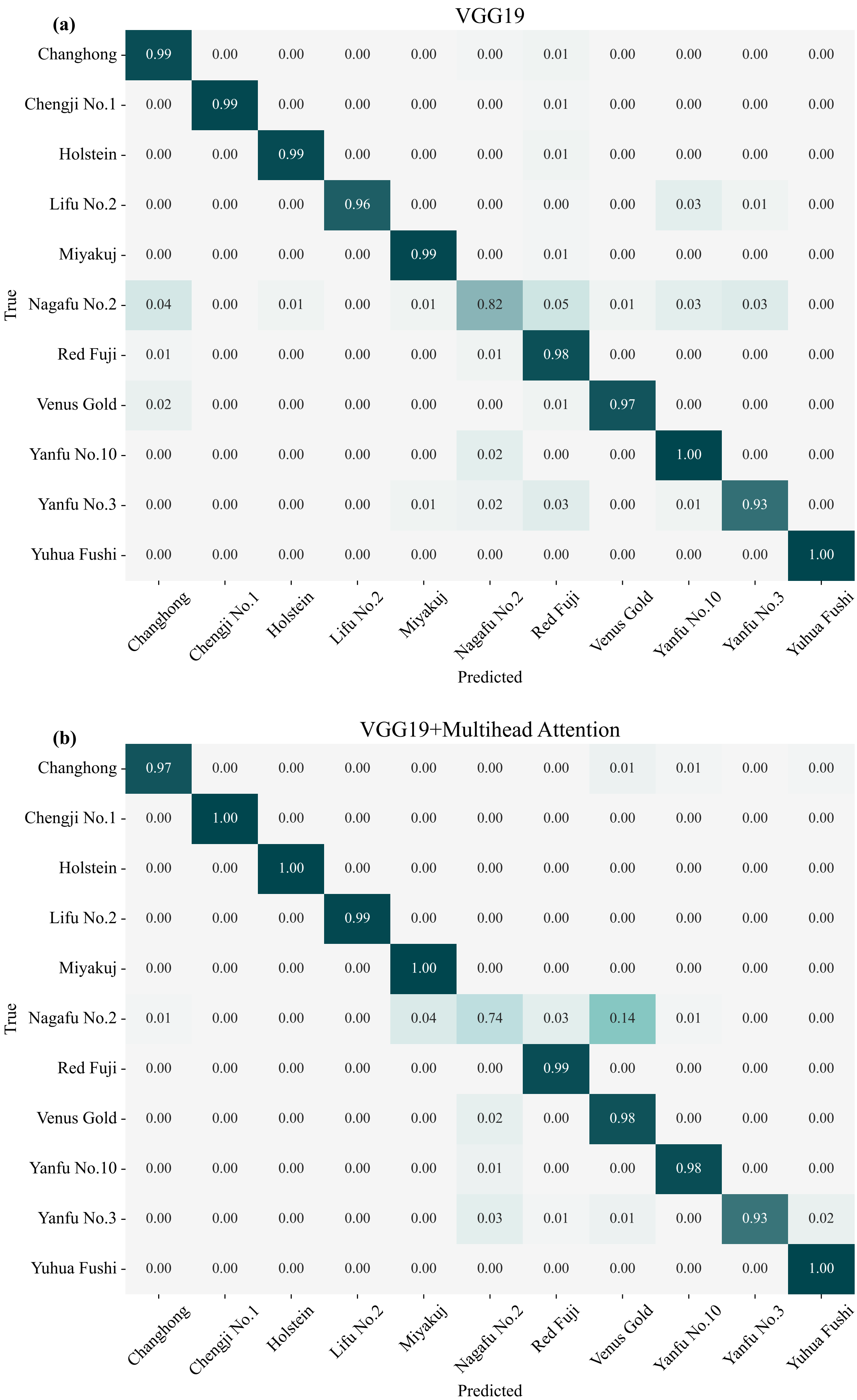


Fig. 7. Prediction of eleven apple cultivars using three MobileNet-V2-based models, including (a) MobileNet-V2 model, (b) MobileNet-V2+Multi-Head Attention model, and (c) MobileNet-V2+GLCM model.



True \ Predicted	Changhong	Chengji No.1	Holstein	Lifu No.2	Miyakuj	Nagafu No.2	Red Fuji	Venus Gold	Yanfu No.10	Yanfu No.3	Yuhua Fushi
Changhong	0.99	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Chengji No.1	0.00	0.99	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Holstein	0.00	0.00	0.99	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Lifu No.2	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.03	0.01	0.00
Miyakuj	0.00	0.00	0.00	0.00	0.99	0.00	0.01	0.00	0.00	0.00	0.00
Nagafu No.2	0.04	0.00	0.01	0.00	0.01	0.82	0.05	0.01	0.03	0.03	0.00
Red Fuji	0.01	0.00	0.00	0.00	0.00	0.01	0.98	0.00	0.00	0.00	0.00
Venus Gold	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.00	0.00
Yanfu No.10	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	1.00	0.00	0.00
Yanfu No.3	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.01	0.93	0.00
Yuhua Fushi	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

True \ Predicted	Changhong	Chengji No.1	Holstein	Lifu No.2	Miyakuj	Nagafu No.2	Red Fuji	Venus Gold	Yanfu No.10	Yanfu No.3	Yuhua Fushi
Changhong	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
Chengji No.1	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Holstein	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lifu No.2	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Miyakuj	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Nagafu No.2	0.01	0.00	0.00	0.00	0.04	0.74	0.03	0.14	0.01	0.00	0.00
Red Fuji	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00
Venus Gold	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.98	0.00	0.00	0.00
Yanfu No.10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.98	0.00	0.00
Yanfu No.3	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.93	0.02
Yuhua Fushi	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

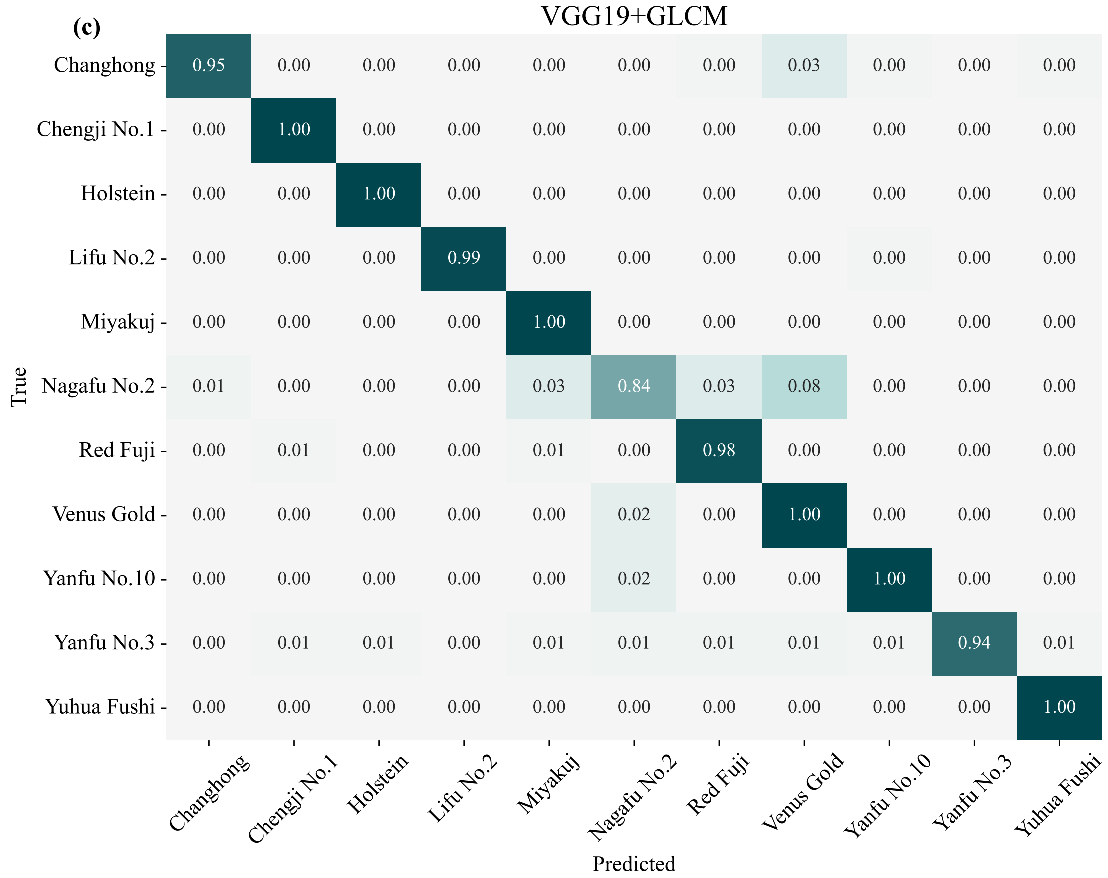


Fig. 8. Prediction of eleven apple cultivars using three VGG19-based models, including (a) VGG19 model, (b) VGG19+Multi-Head Attention model and (c) VGG19+GLCM model.

Table Captions

Table 1 Apple cultivar recognition dataset, featuring initial images of eleven apple samples and their augmented versions.

Table 2 Performance of deep learning models combining Multi-Head Attention mechanism and Gray-Level Co-occurrence Matrix with CNN, MobileNet-V2, and VGG19.

Table 1 Apple cultivar recognition dataset, featuring original images of eleven apple cultivar samples and their augmented versions.

Cultivar Name	Apple Origin	Original Images	Darkened Images	Varied Images	Gaussian Filtered Images	Gaussian Noise Images	Total Images
Changhong	Shaanxi Province	294	50	50	50	50	494
Chengji No.1	Shanxi Province	294	50	50	50	50	494
Holstein	Shanxi Province	312	50	50	50	50	512
Lifu No.2	Henan Province	306	50	50	50	50	506
Miyakji	Hebei Province	312	50	50	50	50	512
Nagafu No.2	Hebei Province	312	50	50	50	50	512
Red fuji	Hebei Province	288	50	50	50	50	488
Venus Gold	Liaoning Province	288	50	50	50	50	488
Yanfu No.3	Shandong Province	312	50	50	50	50	512
Yanfu No.10	Shandong Province	324	50	50	50	50	524
Yuhua Fushi	Shandong Province	300	50	50	50	50	500

Table 2 Model performance of CNN, MobileNet-V2, and VGG19.

Morphological method	Model optimization content	Accuracy	Precision	Recall	F-score	MAE
CNN	/	0.9646	0.9648	0.9646	0.9641	0.1311
CNN	Multi-Head Attention	0.9708	0.971	0.9708	0.9704	0.1090
CNN	GLCM	0.9792	0.9791	0.9792	0.9786	0.0980
MobileNet-V2	/	0.9778	0.9778	0.9778	0.9774	0.0696
MobileNet-V2	Multi-Head Attention	0.8733	0.8834	0.8733	0.8704	0.5490
MobileNet-V2	GLCM	0.9825	0.9829	0.9825	0.9820	0.0571
VGG19	/	0.9725	0.9725	0.9725	0.9722	0.0992
VGG19	Multi-Head Attention	0.9701	0.9703	0.9700	0.9698	0.0975
VGG19	GLCM	0.9792	0.9791	0.9772	0.9766	0.0921