IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5643~5656
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5643-5656

Fruit sorting and quality assessment using convolutional neural networks and vision transforms

Divyashree Sreerangaiah[1,2], Krishna Alabujanahalli Neelegowda[1], Anitha Arekattedoddi Chikkalingaiah[3]
[1]Department of Computer Science and Engineering, SJB Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, India
[2]Department of Computer Science and Engineering (Cyber Security), RNS Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, India
[3]Department of Computer Science and Engineering, Government Sri Krishnarajendra Silver Jubilee Technological Institute, Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jul 30, 2025
Revised Jul 31, 2026
Accepted Aug 31, 2026

Keywords:

Fruit classification
Image classification
Quality assessment
Transfer learning
Vision transformer

ABSTRACT

Fruits play a significant role in diets across the world, and inadequate categorization and quality evaluation might have detrimental effects on consumer health and the economy. This study especially demonstrates the performance of vision transformers (ViTs) and convolutional neural networks (CNNs) in multi-label fruit type and quality classification in order to assess automated fruit categorization systems employing deep learning. Finding the most precise and efficient design is necessary to guarantee food safety and boost market value. The ViT model outperforms the CNN model using a dataset of 13,714 photos divided into six categories. Although the CNN scores 98.07% for quality and 96.14% for fruit type, it achieves 99.31% accuracy for quality rating and 99.71% accuracy for fruit type classification. Deep learning models work similarly to support vector machines (SVM) and other conventional classifiers. The paper outlines the benefits and drawbacks of both quality inspection schemes and real-time classification.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Divyashree Sreerangaiah
Department of Computer Science and Engineering, SJB Institute of Technology
Affiliated to Visvesvaraya Technological University
Belagavi 590018, Karnataka, India
Email: divyashrees2008@gmail.com

1. INTRODUCTION

Fruits must be carefully sorted and inspected in order to maintain consistent quality. If physical work is employed, this method may be labor-intensive and prone to mistakes. Deep learning and computer vision techniques are being used more often to improve fruit categorization procedures as agriculture develops and the need for automated and accurate solutions grows. These techniques have demonstrated good performance in picture recognition tests [1]–[3], providing a more effective substitute for conventional techniques. They can automatically identify hierarchical characteristics in raw data. Convolutional neural networks (CNN) are a prime example of this [4], [5]. However, a different strategy that avoids the inductive biases of convolutional processes is provided by the recent development of vision transformers (ViTs), which are used by self-attention mechanisms to capture global dependencies in pictures [6].

A lot of research has been conducted on automated fruit classification and quality evaluation with deep learning methods. Xiao *et al.* [6] found swin transformers excelled in ripeness classification. Ukwuoma *et al.* [7] discussed the transition to deep learning in fruit detection with improved dataset

performances. On inexpensive hardware, Ismail and Malik [8] created a useful real-time fruit rating system. Hayat *et al.* [9] developed FruitVision, a deep learning grading system with 99.5% accuracy across fruits. According to this paper, a two-step image analysis technique for detecting and counting pepper fruits in cluttered greenhouse environments is presented. Song *et al.* [10] approach achieved a 74.2% correlation with manual counts using over 28,000 images. Amatya *et al.* [11] built a Bayesian machine vision system achieving 89.2% accuracy in detecting cherry branches to automate harvesting. Sarigu *et al.* [12] supported taxonomic and biodiversity investigations by classifying plum cultivars using image analysis and linear discriminant analysis (LDA). Peña *et al.* [13] improved fruit-tree crop classification in Chile using Landsat-8 spectral data along with penalized LDA. Hameed *et al.* [14] examined computer vision methods to aid in the categorization of fruits and vegetables, highlighting difficulties and developments. gave a thorough analysis of fruit and vegetable classification methods, emphasizing issues like high-dimensional characteristics and interclass similarity. For practical applications, they evaluated many machines learning methods, including support vector machines (SVM), k-nearest neighbors (KNN), artificial neural network (ANN), and CNN.

Abbas *et al.* [15] proposed automated image processing to overcome inefficiencies in traditional fruit sorting. Using EfficientNet and MixNet, Duong *et al.* [16] created quick fruit identification models that are appropriate for low-resource devices. Torres *et al.* [17] studied CNN applications in fruit image processing, who noted benefits from transfer learning. For evaluating apple quality, Li *et al.* [18] created a CNN that outperformed Inceptionv3 and conventional techniques. Mimma *et al.* [19] used you only look once (YOLO)v3 with ResNet50 and VGG16 for fruit categorization and identification with great accuracy. In their evaluation of Indian fruit grading technology, Ratha *et al.* [20] emphasized automation to cut down on waste and mistakes. Knott *et al.* [21] demonstrated ViTs efficacy with limited training data in defect detection. Dhiman *et al.* [22] systematically reviewed methods for product classification. Chen *et al.* [23] enhanced MobileNetV2 for better agricultural product classification.

For real-time pear identification, Zhao *et al.* [24] suggested high-level deformation-perception network with multi-object search non-maximum suppression (HDMNet), which outperformed YOLOv8. In order to help with agricultural planning, to determine the mass and volume of baobab fruit, Dono *et al.* [25] used computer vision. [26], [27] presented a multi-task CNN that improves fruit freshness and type recognition, outperformed manual examination by using X-ray imaging and random forest (RF) classifiers to identify internal blackheart disease in pomegranates with 93.3% accuracy. In order to decrease food wastage and enhance supply chain quality, Rajini and Voola [28] combined internet of things (IoT) and machine learning utilizing categorical boosting (CatBoost) for precise banana ripeness detection. A CNN-based multi-label fruit classification method was proposed by Biswas *et al.* [29], and it achieved approximately 98% accuracy on apple, lemon, tomato, and plum. In agro-based applications, the method performed better than SVM and traditional classifiers. Espinoza *et al.* [30] reviewed CNN, region-based convolutional neural network (R-CNN), and YOLO models are among the deep learning methods for fruit analysis. Focusing on how models like CNN, R-CNN, and YOLO identify characteristics like the color, brightness, and sheen of ripe fruit using deep learning techniques for fruit analysis.

This study compares and contrasts two deep learning methods, each used for classification and evaluation of fruits based on images: i) an end-to-end designed CNN model (deep learning approach) for fruit quality evaluation and image classification and ii) the ViT model is trained in advance and functions as a feature extractor with machine learning techniques [31]–[33]. To meet useful multi-output needs in real-world applications, this study evaluates both models' capacities for fruit type classification as well as their quality levels. In order to provide useful information for the creation of scalable, automated fruit grading systems, it attempts to draw attention to the trade-offs between accuracy, computing efficiency, and flexibility of both approaches [7], [34].

The following structure is provided by the subsequent sections of this work. Section 2 explains a detailed study of CNN and ViT architecture. The outcomes are elaborated on in section 3. The conclusion is focused on in section 4.

2. METHOD

This section compares the two architectures, CNN and ViT. It explains the sequential processes in the proposed CNN-based architecture. These processes include dataset preparation, feature extraction, feature encoding, and machine learning-based classification.

2.1. Dataset used

The publicly available FruitNet dataset, introduced by Meshram and Patil [35], is used in this study which is also available in the Mendeley data repository [36]. That dataset contains 13,714 RGB images of fruits, arranged for multi-label classification of type and quality. Each image has two labels: one is a class

name that includes apple-banana-guava-lime-orange and pomegranate, as well as its respective quality, good or bad. The folder structure mirrors this labelling system with subfolders labelled Apple_Good, Banana_Bad, and others, grouped under parent folders Good_quality and Bad_quality. A custom dataset class in PyTorch was created to load the images and grab the appropriate labels from the folder names for faster processing. The dataset was designed for training both CNN and ViT models to classify the fruit type and its quality jointly. It contains 6 fruit classes in 2 qualities, spanning a total of 10,971 training images, and 2,743 test images. Additionally, Figure 1 displays the sample images.

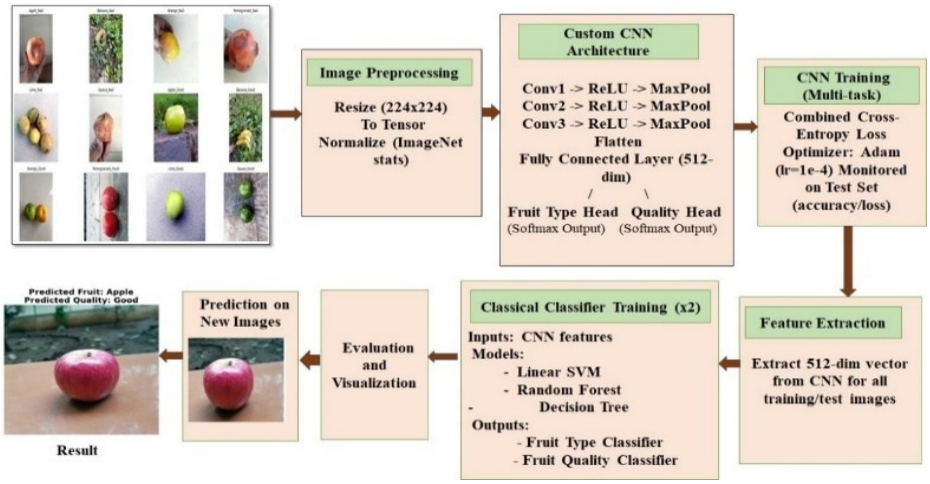


Figure 1. Flowchart of the architecture based on CNN

2.2. Convolutional neural network-based architecture

The CNN-based model developed to learn classify fruit type and quality using shared features. The architecture uses a multitask learning framework to efficiently extract and use hierarchical image characteristics, as seen in Figure 1. Every image was resized to 224 by 224 pixels, converted to three-channel RGB format, and normalized using ImageNet mean and standard deviation values, as specified in (1), before to being input into the model.

$$\mathrm{I}_{\mathrm{normalized}} = \frac{\mathrm{I}_{\mathrm{tensor}} - \mu}{\sigma} \quad (1)$$

Here $\mathrm{I}_{\mathrm{tensor}}$ is the input image tensor, $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ show the mean and standard-deviation of ImageNet, whereas inormalized is the standardized image tensor. The image preprocessing results are shown in Figure 2, where Figure 2(a) presents the original image and Figure 2(b) shows the preprocessed image, demonstrating the effect of this step. To enable supervised classification, categorical fruit labels are transformed into numerical values using label encoding. This transformation is defined in (2).

$$\mathrm{LabelEncoder}(l_i) = y_i \in \{0,1, \ldots \ldots, n-1\} \quad (2)$$

Here, the original label for the i-th image is represented by l_i, the integer-encoded label is represented by y_i, and the total number of fruit classes is represented by n. The use of label encoding allows the CNN to utilize information about both the fruit quality and the fruit class to make predictions.



(a) (b)

Figure 2. Image preprocessing of fruit images: (a) original image and (b) preprocessed image

Every convolutional layer has a rectified linear unit (ReLU) activation mechanism along with a maximum pooling process to enable hierarchical extraction of features, a layer that was fully linked to the end of the convolutional layers to produce a 512-dimensional feature space that would represent the output of the convolutional layers as stated in (3).

$$z_i = f_{cnn}(x_i) \tag{3}$$

Where x_i stands for the normalized input tensor, f_{cnn} is the CNN feature extraction function, and z_i comes out as a 512-dimensional feature vector—like a neat stack of numbers capturing what the network sees. The CNN feature extraction process is illustrated in Figure 3, where Figure 3(a) shows the feature maps generated by the third convolutional layer, displaying the first 32 channels, and Figure 3(b) presents the resulting 512-dimensional feature vector. For multi-task classification, two task-specific softmax heads are employed to predict fruit type and quality simultaneously, using the shared feature vector z_i. This prediction is computed as in (4).

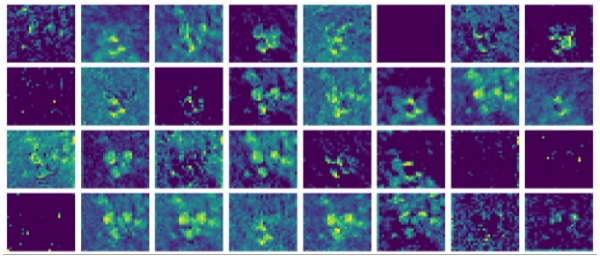
$$\hat{y}_i^{type} = softmax\left(W_{type} z_i + b_{type}\right), \hat{y}_i^{quality} = softmax\left(W_{quality} z_i + b_{quality}\right) \tag{4}$$

Here W_{type}, $W_{quality}$ are the classifier weight matrices, b_{type}, $b_{quality}$ are biases, and softmax converts logits to probability distributions. This configuration allows the model to learn both tasks efficiently from the shared representation. The network learns through a loss function for many tasks that blends the distinct losses from both tasks, as shown in (5). In this equation, $\mathcal{L}_{type}$ and $\mathcal{L}_{quality}$ represent the classification losses for fruit type and quality, respectively, while λ_1 and λ_2 are weighting coefficients to balance each task. This approach ensures both tasks contribute to the learning process. After training, the CNN is fine-tuned as a feature extractor, where the fully connected layer produces the final 512-dimensional feature vector for activities that come after like clustering or visualization. This process is defined in (6).

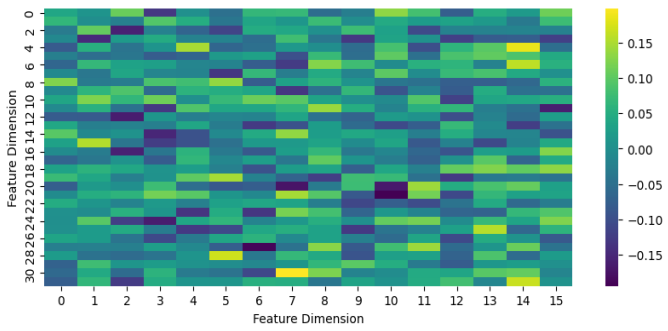
$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{type} + \lambda_2 \mathcal{L}_{quality} \tag{5}$$

$$z_i = f_{fc_feat}(x_i) \tag{6}$$

Here f_{fc_feat} represents the fully connected feature extraction function and x_i is the normalized input tensor. The extracted feature vector provides a compact numerical representation of each fruit image. The example of a feature vector shown in Figure 3(b) was generated from an input image.



(a)



(b)

Figure 3. Feature extraction process of the CNN model: (a) convolutional feature maps and (b) final 512-dimensional feature vector

2.3. Vision transformer -based architecture

This work uses the ViT architecture to produce a large number of features from an input image which represent fine-grained variations such as color changes to identify the fruit's type and quality in a single pass through the network. The difference between how a CNN and a ViT handle an input image are depicted in Figure 4 — a CNN views an image as a single object while a ViT divides the image into many smaller patches, converts each of these patches into a feature vector representation and then uses the collection of all feature vectors to understand the overall content of the image. Prior to processing by the ViT model. The image preprocessing process is illustrated in Figure 5, which shows the original image as in Figure 5(a) and its transformed (normalized) version as in Figure 5(b).

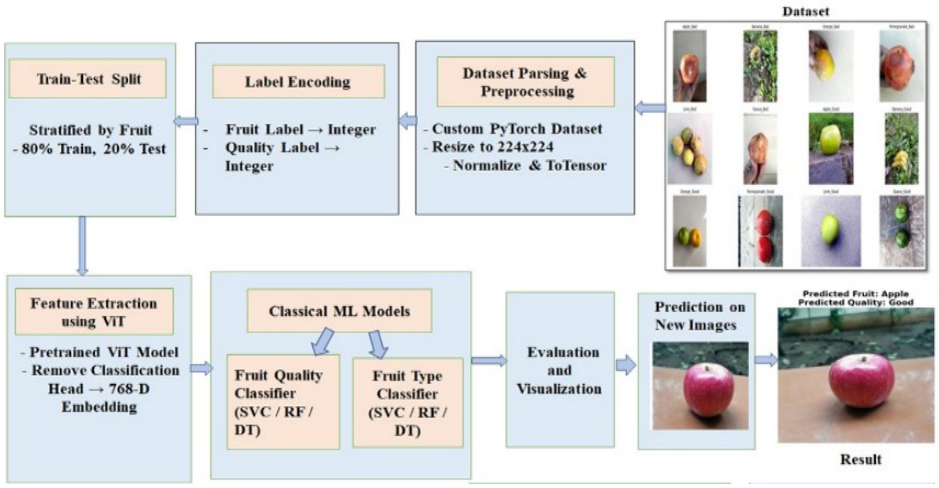


Figure 4. ViT - based architecture



(a) (b)

Figure 5. Image preprocessing: (a) original image and (b) transformed (normalized) image

Each input image $Ii \in \mathbb{R}^{H\times W\times 3}$ is subjected to multiple preprocessing steps so that the images are consistent and compatible with ViT model. The resizing process is mathematically expressed in (7).

$$I_i' = Resize(I_i, 224 \times 224) \quad (7)$$

Where I_i is the original image in this case, with height H and width W, and 3 RGB channels, while I_i' is the resized image. This confirms that very inputs have a consistent shape for batch processing, as visualized in Figure 5(b). Next, normalization adjusts pixel values using ImageNet statistics to match the pretraining distribution, enhancing model convergence and stability, as defined in (8).

$$\mu = [0.485,0.456,0.406] \; and \; \sigma = [0.229,0.224,0.225] \quad (8)$$

In this step, μ and σ represent the mean and standard-deviation of each RGB channel, and centering and scaling the intensity of pixels means that the model will learn from the same scales as the original images to avoid any type of bias. After normalizing the image, it is then changed to a PyTorch tensor, with the first dimension being the channels, by utilizing the ToTensor method shown in (9).

$$x_i = ToTensor(I_i') \Longrightarrow x_i \in \mathbb{R}^{3\times224\times224} \quad (9)$$

The final input tensor to be processed by ViT has a representation of "channel first" which is suitable to process the ViT model. The conversion also enables the processing of the input data on a GPU with a batch. For use in supervised classification, label encoding transforms the categorical fruit sort and quality labels into numerical representations, as given in (10).

$$y_i^{fruit} \in \{0,1,\ldots\ldots,5\}, y_i^{quality} \in \{0,1\} \quad (10)$$

In this case, y_i^{fruit} stands for the encoded label for the i-th fruit-type — for example, apple, banana, and others, and $y_i^{quality}$ represents its quality label. The numeric version of the original labels is shown in Figure 6. This encoding is a key part of the process that allows us to train the model — since it helps keep the categorical data ordered and processed appropriately. To be able to assess the model accurately, the dataset was then divided into separate groups — one is training and testing sets used for a stratified sampling approach, where the proportion of each class was preserved across both splits, as explained in (11).

$$P(y^{fruit} = k) \approx constant \quad (11)$$

This ensures that the probability of encountering each fruit class k remains consistent throughout the training and test sets. Figure 7 visualizes the distribution of classes and confirms the numerical stratification used for model evaluation. After preprocessing, the normalized input tensor is passed through a pretrained ViT model (from the timm library) for feature extraction. The classification head is removed, leaving just the embedding output described in (12) and (13), the part that holds each feature.

$$f_{vit}: \mathbb{R}^{3\times224\times224} \rightarrow \mathbb{R}^{768} \quad (12)$$

$$z_i = f_{vit}(x_i) \in \mathbb{R}^{768} \quad (13)$$

Here f_{vit} represents the features from the ViT mode, and z_i is the resulting 768-dimensional embedding for each image. Figure 8 shows a sample embedding vector, highlighting the first 50 dimensions for interpretability.

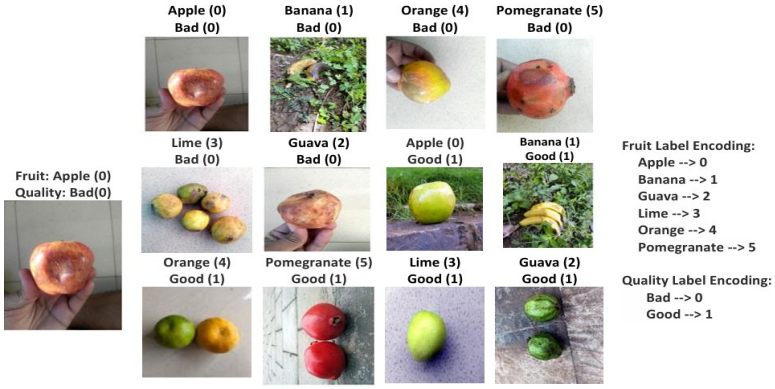


Figure 6. Fruit and quality label encoding

```
Total images found: 13714
Training samples: 10971
Testing samples: 2743

Sample training images:
/content/drive/MyDrive/25062025/Good_quality/Pomegranate_Good/20190820_144620_23316.jpg
/content/drive/MyDrive/25062025/Bad_quality/Orange_Bad/IMG_20190824_174822.jpg
/content/drive/MyDrive/25062025/Bad_quality/Apple_Bad/IMG_20190910_171631.jpg
/content/drive/MyDrive/25062025/Bad_quality/Guava_Bad/IMG_20190822_081507_1.jpg
/content/drive/MyDrive/25062025/Good_quality/Pomegranate_Good/IMG20200728125504_27009.jpg

Sample testing images:
/content/drive/MyDrive/25062025/Bad_quality/Orange_Bad/IMG_20190824_174852_1.jpg
/content/drive/MyDrive/25062025/Bad_quality/Orange_Bad/IMG_20190824_174931.jpg
/content/drive/MyDrive/25062025/Bad_quality/Pomegranate_Bad/IMG20200728182822.jpg
/content/drive/MyDrive/25062025/Good_quality/Lime_Good/IMG_20190902_104343.jpg
/content/drive/MyDrive/25062025/Bad_quality/Pomegranate_Bad/IMG20200728182938.jpg
```

Figure 7. Train-test data splitting

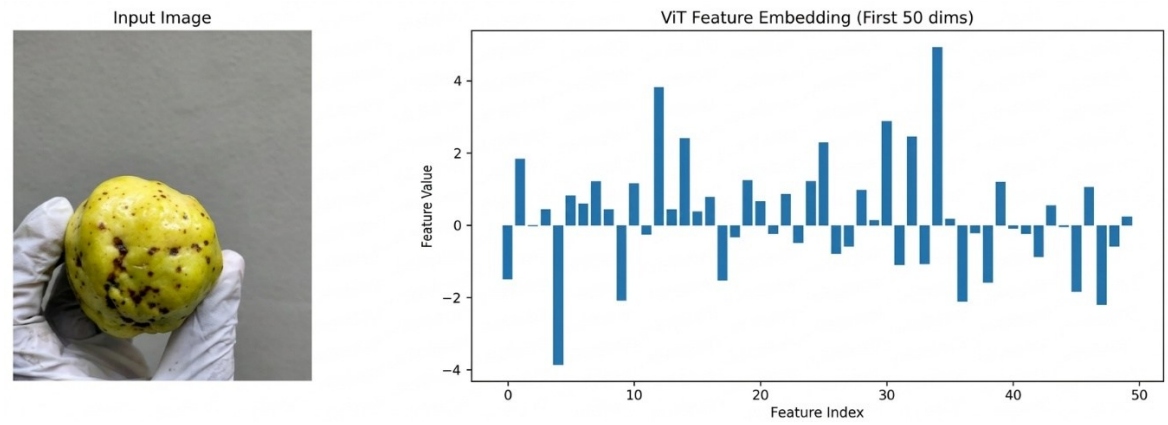


Figure 8. Input image with embedded ViT feature extraction

2.4. Training classical machine learning classifiers for CNN and ViT

After obtaining feature vectors from CNN and ViT, these embeddings are used to train classical machine-learning classifiers for two tasks. The first task is the fruit type classification: $C_{fruit}: R^{768} \rightarrow \{0,1,2,3,4,5\}$, $\hat{y}_i^{type} = C_{fruit}(x_i)$. The next task is dealing with the quality of fruit - each fruit image has a vector dimension of 768 and is processed by model: $C_{quality}: R^{768} \rightarrow \{0,1\}$, $\hat{y}_i^{quality} = C_{quality}(x_i)$. The input variable xi in this instance is represented by the feature vector for the i-th image. Therefore, three traditional classifiers were tested for both tasks: RF, decision tree (DT), and linear support vector classifier (LinearSVC). Table 1 illustrations the quantitative analysis of CNN and ViT models overall. The detailed CNN-based classification outcomes are summarized in Table 2, and the classification outcomes using all three classifiers based on ViT are summarized in Tables 3 and 4. Lastly, Table 5 presents a comparison between the proposed and current models from the literature.

3. RESULTS AND DISCUSSION

Fruit type classification and fruit quality prediction were the two tasks utilized to evaluate CNN and ViT models. While the ViT model uses attention processes to capture global relationships, the CNN model uses convolutional filters to extract local characteristics (Figure 4). Feature extraction strategies have a considerable influence on results, according to a comparison of classifier performance approaches in fine-grained fruit analysis [14], [20], [30]. According to earlier studies, ViT's attention mechanism effectively captures global dependencies, which enhances fine-grained categorization [6], [30].

3.1. Performance metrics

Standard criteria were used to statistically evaluate the models. Accuracy, which measures the total correctness of predictions, is defined in (14). Where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), respectively.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{14}$$

In (15) and (16) illustrates the computation of class-specific precision and recall.

$$\text{Precision}_k = \frac{\text{TP}_\text{k}}{\text{TP}_\text{k}+\text{FP}_\text{k}} \tag{15}$$

$$\text{Recall}_k = \frac{\text{TP}_\text{k}}{\text{TP}_\text{k}+\text{FN}_\text{k}} \tag{16}$$

Where TP_k, FP_k, FN_k corresponding to class k. The F1-score is derived from the harmonic mean of precision and recall (17).

$$F1_\text{k} = 2 \times \frac{\text{Precision}_\text{k} \times \text{Recall}_\text{k}}{\text{Precision}_\text{k}+\text{Recall}_\text{k}} \tag{17}$$

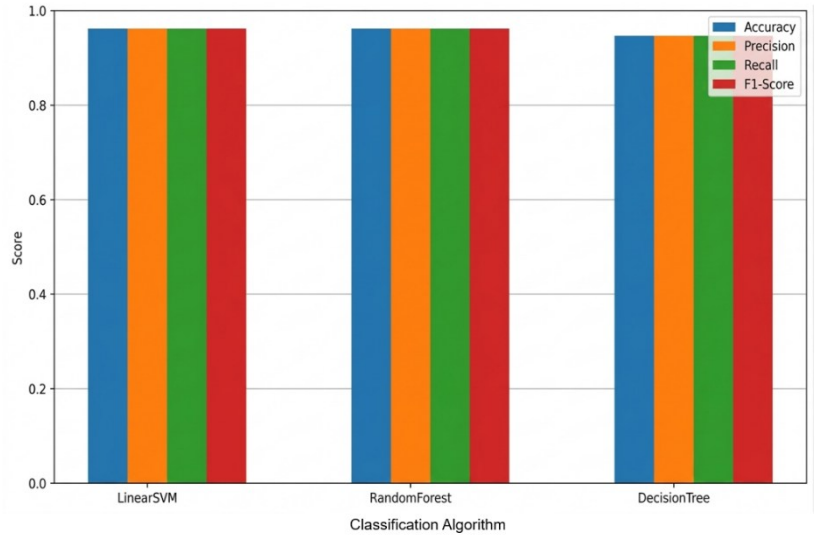
Weighted averages account for class imbalance. Log loss, which measures prediction confidence, is defined in (18).

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\ \sum_{k=1}^{K} y_{i,k}\log(p_{i,k}) \quad (18)$$

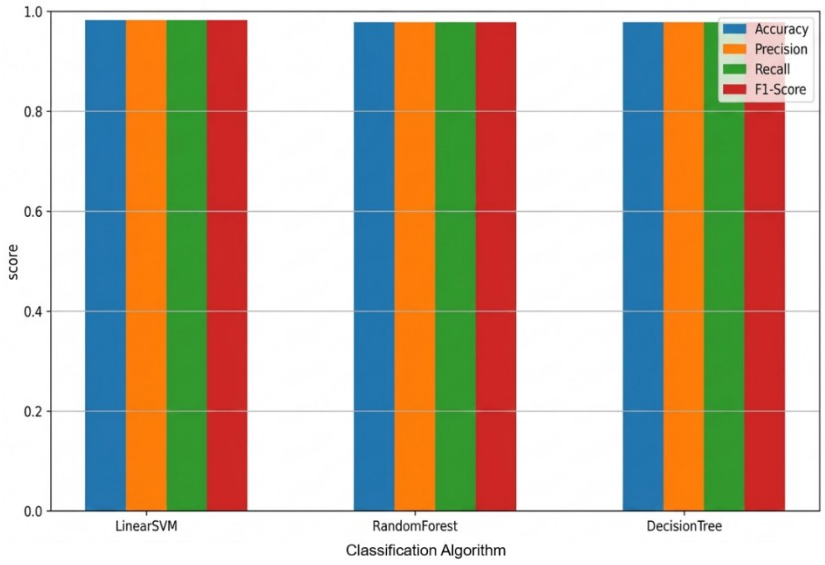
Where $y_{i,k}$ indicates true class membership and $p_{i,k}$ is the predicted probability. The evaluation metrics are summarized in Tables 1 to 5 and the curves for accuracy and loss are shown.

3.2. Performance evaluation of CNN and ViT-model based results with different classifiers

Figures 9 display the CNN model's performance for fruit classification and quality prediction. Figure 9(a) shows the results of fruit type classification, while Figure 9(b) presents the results of fruit quality. CNN models have shown to be successful in fruit classification task [6], [26]. The training and validation for the SVM, RF, and DT classifiers are displayed in Figure 10. Figure 10(a) shows the accuracy curves, while Figure 10(b) presents the loss curves.

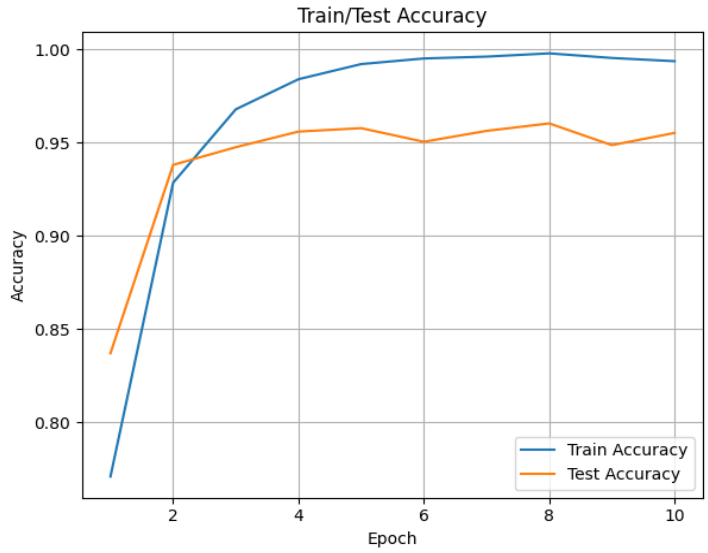


(a)

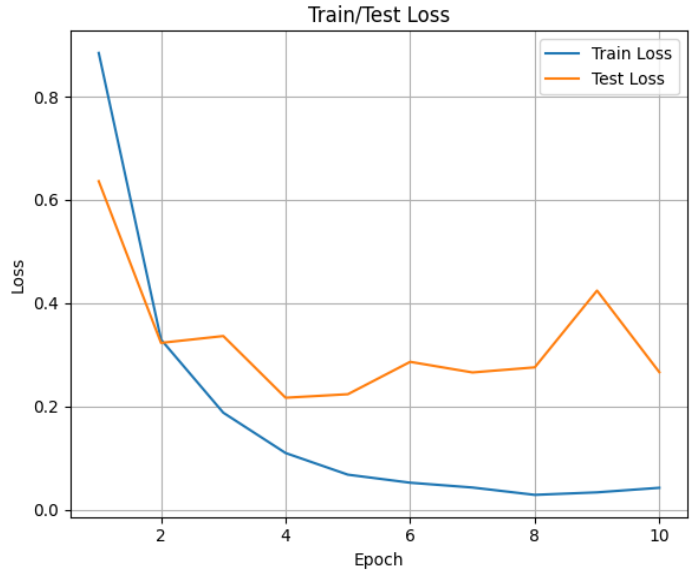


(b)

Figure 9. Performance evaluation using CNN-based features with SVM, RF, and DT classifiers: (a) fruit type classifiers and (b) fruit quality classifiers



(a)



(b)

Figure 10. Training and validation of the CNN model: (a) accuracy curves and (b) loss curves

Figure 11 presents visual predictions for fruit classification and quality prediction using ViT-based features with SVM, RF, and DT classifiers. Figure 11(a) shows the results of fruit type classification, while Figure 11(b) presents the results of fruit quality classification. Figure 12 presents the accuracy and loss curves of the CNN model. Figure 12(a) shows the accuracy and loss curves for fruit type classification, while Figure 12(b) presents the accuracy and loss curves for fruit quality classification. Global image features can be captured using ViT, which can enhance the classification performance of fine-grained images [6], [9]. Table 1 compares the CNN and ViT classifiers. CNN outperforms ViT with LinearSVM [6], [9] but not as much as ViT does. This is a better representation of features.

The detailed performance evaluation of the CNN and ViT models for fruit type and fruit quality classification is provided in Tables 2 and 3, accordingly. The comparative results of both models are summarized in Table 4. The outcomes confirm the superiority of the ViT-based features in terms of classification performance, compared with CNN features [6], [9]. The proposed models are compared with ongoing studies in Table 5. As observed from the comparison, the proposed ViT model has competitive results on the fruit classification and quality prediction tasks [6], [9], [17], [26], [29].

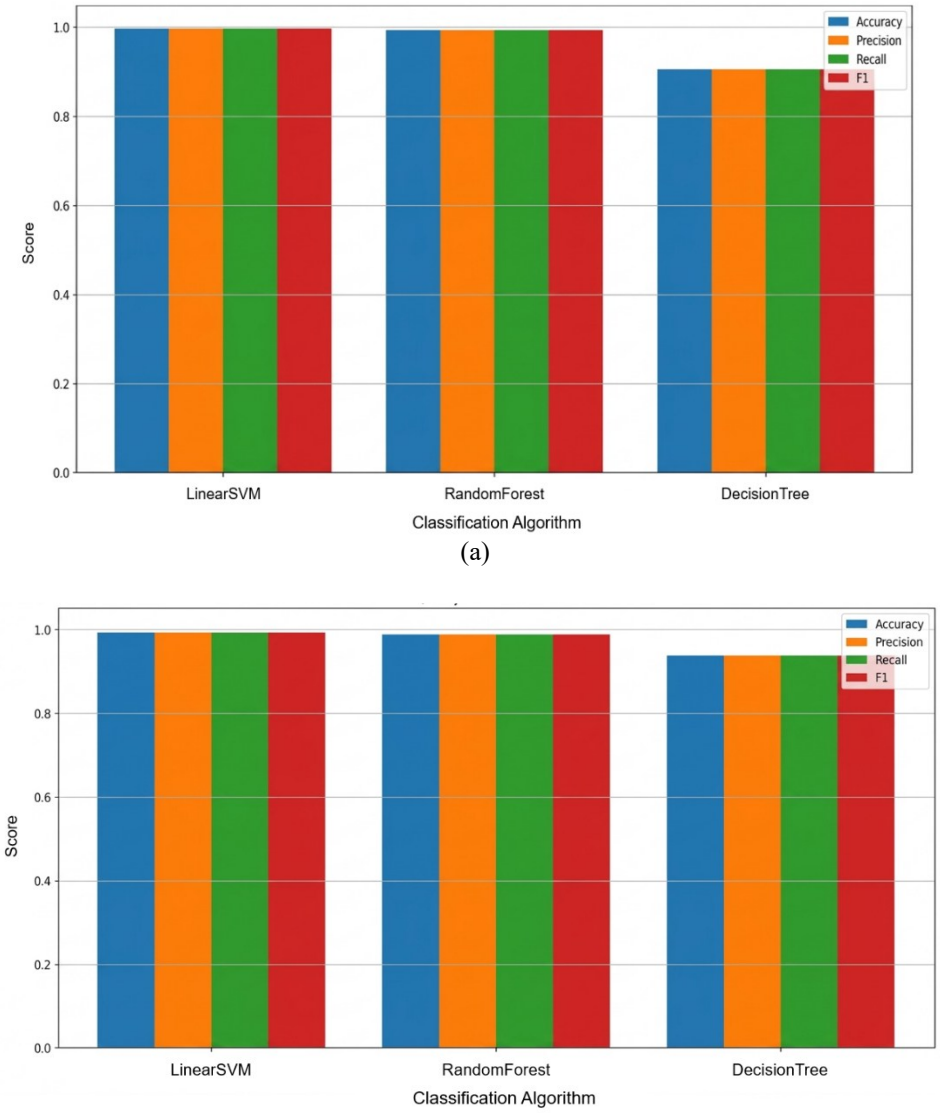


Figure 11. Performance evaluation using ViT-based features with SVM, RF, and DT classifiers: (a) fruit type classifiers and (b) fruit quality classifiers

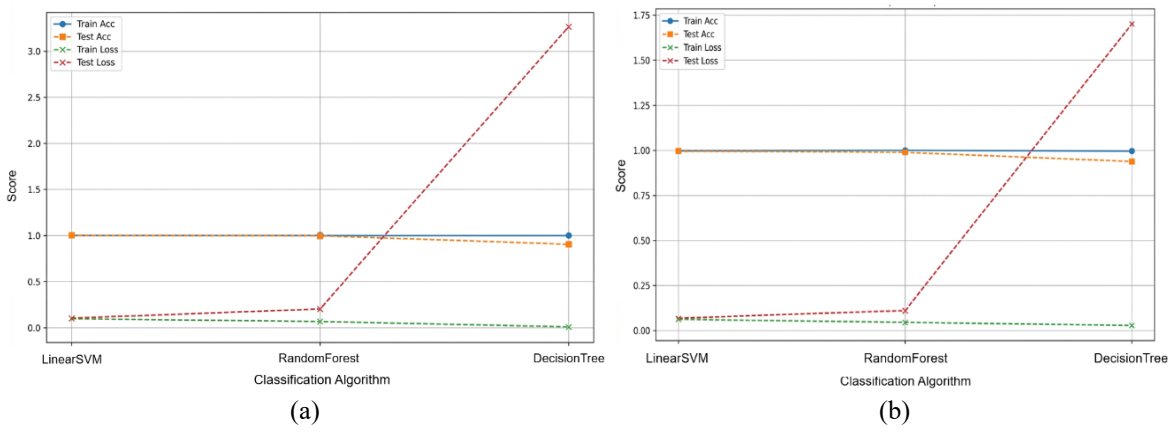


Figure 12. Accuracy and loss curves of the CNN model: (a) fruit type and (b) fruit quality

Table 1. Quantitative comparison for CNN and ViT

Model type	Classifier	Accuracy (fruit type) (%)	Macro F1-score	Notable observations
CNN	LinearSVM	96	0.96	Occasional misclassifications (e.g., apple vs guava)
	RF	96	0.96	High precision, minor recall drop
	DT	95	0.95	More misclassification in apple/guava
ViT	LinearSVM	99.71	99.71%	Perfect class separation
	RF	99	0.99	Small confusion (e.g., orange)
	DT	90	0.90	Underperforms with high-dimensional ViT features

Table 2. Performance evaluation of fruit type classifiers and quality for CNN model

Classifiers	Types of fruits	Precision	Recall	F1-score	Support
		Fruit type			
Linear SVM	Apple	0.94	0.92	0.93	455
	Banana	0.96	0.99	0.98	440
	Guava	0.94	0.93	0.94	456
	Lime	0.97	0.97	0.97	436
	Orange	0.98	0.97	0.97	475
	Pomegranate	0.98	0.98	0.98	481
	Accuracy	-	-	0.96	2743
	Macro avg	0.96	0.96	0.96	2743
	Weighted avg	0.96	0.96	0.96	2743
RF	Apple	0.94	0.91	0.93	455
	Banana	0.96	1.00	0.98	440
	Guava	0.95	0.93	0.94	456
	Lime	0.97	0.97	0.97	436
	Orange	0.96	0.97	0.97	475
	Pomegranate	0.98	0.99	0.98	481
	Accuracy	-	-	0.96	2743
	Macro avg	0.96	0.96	0.96	2743
	Weighted avg	0.96	0.96	0.96	2743
DT	Apple	0.91	0.90	0.90	455
	Banana	0.96	0.98	0.97	440
	Guava	0.94	0.90	0.92	456
	Lime	0.94	0.97	0.95	436
	Orange	0.94	0.94	0.94	475
	Pomegranate	0.98	0.99	0.98	481
	Accuracy	-	-	0.95	2743
	Macro avg	0.95	0.95	0.95	2743
	Weighted avg	0.95	0.95	0.95	2743
		Fruit quality			
Linear SVM	Bad	0.98	0.98	0.98	1396
	Good	0.98	0.98	0.98	1347
	Accuracy	-	-	0.98	2743
	Macro avg	0.98	0.98	0.98	2743
	Weighted avg	0.98	0.98	0.98	2743
RF	Bad	0.98	0.98	0.98	1396
	Good	0.98	0.98	0.98	1347
	Accuracy	-	-	0.98	2743
	Macro avg	0.98	0.98	0.98	2743
	Weighted avg	0.98	0.98	0.98	2743
DT	Bad	0.98	0.98	0.98	1396
	Good	0.97	0.98	0.98	1347
	Accuracy	-	-	0.98	2743
	Macro avg	0.98	0.98	0.98	2743
	Weighted avg	0.98	0.98	0.98	2743

Table 3. Performance evaluation of fruit type classifiers and quality for ViT model

Classifiers	Types of fruits	Precision	Recall	F1-score	Support
		Fruit type			
Linear SVM	Apple	1.00	0.99	1.00	455
	Banana	1.00	1.00	1.00	440
	Guava	1.00	1.00	1.00	456
	Lime	0.99	0.99	0.99	436
	Orange	1.00	1.00	1.00	475
	Pomegranate	1.00	1.00	1.00	481
	Accuracy	-	-	1.00	2743
	Macro avg	1.00	1.00	1.00	2743
	Weighted avg	1.00	1.00	1.00	2743
RF	Apple	0.99	0.99	0.99	455
	Banana	1.00	1.00	1.00	440
	Guava	0.99	0.99	0.99	456
	Lime	0.99	0.98	0.99	436
	Orange	0.98	1.00	0.99	475
	Pomegranate	1.00	1.00	1.00	481
	Accuracy	-	-	0.99	2743
	Macro avg	0.99	0.99	0.99	2743
	Weighted avg	0.99	0.99	0.99	2743
DT	Apple	0.86	0.84	0.85	455
	Banana	0.98	0.98	0.98	440
	Guava	0.89	0.90	0.89	456
	Lime	0.89	0.89	0.89	436
	Orange	0.87	0.91	0.89	475
	Pomegranate	0.94	0.91	0.93	481
	Accuracy	-	-	0.90	2743
	Macro avg	0.90	0.90	0.90	2743
	Weighted avg	0.90	0.90	0.90	2743
		Fruit quality			
Linear SVM	Bad	0.99	0.99	0.99	1335
	Good	1.00	0.99	0.99	1408
	Accuracy	-	-	0.99	2743
	Macro avg	0.99	0.99	0.99	2743
	Weighted avg	0.99	0.99	0.99	2743
RF	Bad	0.99	0.99	0.99	1335
	Good	0.99	0.99	0.99	1408
	Accuracy	-	-	0.99	2743
	Macro avg	0.99	0.99	0.99	2743
	Weighted avg	0.99	0.99	0.99	2743
DT	Bad	0.93	0.98	0.98	1335
	Good	0.94	0.94	0.94	1408
	Accuracy	-	-	0.98	2743
	Macro avg	0.94	0.94	0.94	2743
	Weighted avg	0.94	0.94	0.94	2743

Table 4. Performance comparison of CNN and ViT models for fruit type and quality classification

Model	Classification task (best classifier)	Accuracy (%)	F1-score (%)	Log loss
CNN	Fruit (Best: LinearSVM))	~96.14	~96.13	~0.12
	Quality (Best: LinearSVM)	~98.07	~98.07	~0.15
ViT	Fruit (Best: SVM)	~99.71	~99.71	~0.10
	Quality (Best: SVM)	~99.31	~99.31	~0.06

Table 5. CNN and ViT model performance comparison with existing studies

Author(s)	Model	Dataset	Task	Results	Observations
Xiao *et al.* [6]	Swin transformer	Apples and pears	Ripeness detection	Mean average precision (mAP) ≈ 87%	Real-time; ViT struggles on small data
Hayat *et al.* [9]	FruitVision (MobileNetV3)	Multi-fruit	Fruit grading	98–99%	Pretraining + preprocessing help
Torres *et al.* [17]	CNNs	Various fruits	Classification and detection	90–100%	Transfer learning helps
Zhang *et al.* [26]	Multi-task CNN	Multi-fruit	Freshness + type	Freshness 93%, Type 89%	Multi-task > single-task
Biswas *et al.* [29]	Deep CNN	4 fruits, ~1200 images	Multi-label classification	≈ 98%	CNN > SVM
This paper	CNN and ViT	13,714 images	Type & quality	Type: CNN 96.14% and ViT 99.71% Quality: CNN 98.07% and ViT 99.31%	Large dataset; ViT slightly better; accurate and practical

4. CONCLUSION

This research shows that, in classifying the varietal and grade quality of fruits, both CNN and ViT model architectures are effective. The ViT models obtained the best results, with accuracy from 99.71% (fruit type) to 99.31% (fruit quality) and the confusion matrix presents a very clear distinction between classes. VGG based models showed relatively low accuracy with 96.14% for fruit type to 98.03% for fruit quality, and confused among similar fruits like apples and guavas, as well. Modifying ViT models by training them on large and adding more different datasets can improve their effectiveness and resiliency for future needs. The application of CNN models on mobile or edge devices will enable real-time fruit classification in practical environments. Exploring hybrid architectures that integrate the advantages of both ViT and CNN could provide a promising tradeoff between operational efficiency and precision, thereby increasing the system's use in practical scenarios.

FUNDING INFORMATION

Authors state there is no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Divyashree Sreerangaiah	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Krishna Alabujanahalli Neelegowda	✓	✓				✓		✓	✓	✓	✓	✓	✓	
Anitha Arekattedoddi Chikkalingaiah	✓				✓					✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly available through the FruitNet dataset repository hosted on Mendeley Data at https://data.mendeley.com/datasets/b6fftwbr2v/1. The same dataset is also available through the Kaggle repository at https://www.kaggle.com/datasets/shashwatwork/fruitnet-indian-fruits-dataset-with-quality/data.

REFERENCES

[1] H. Khalid, "Modern techniques in detecting, identifying and classifying fruits according to the developed machine learning algorithm," *Journal of Applied Research and Technology*, vol. 22, no. 2, pp. 219–229, Apr. 2024, doi: 10.22201/icat.24486736e.2024.22.2.2269.

[2] I. D. Apostolopoulos, M. Tzani, and S. I. Aznaouridis, "A general machine learning model for assessing fruit quality using deep image features," *AI*, vol. 4, no. 4, pp. 812–830, Sep. 2023, doi: 10.3390/ai4040041.

[3] M. Rizzo, M. Marcuzzo, A. Zangari, A. Gasparetto, and A. Albarelli, "Fruit ripeness classification: a survey," *Artificial Intelligence in Agriculture*, vol. 7, pp. 44–57, Mar. 2023, doi: 10.1016/j.aiia.2023.02.004.

[4] X. Gong and Q. Wu, "Fruit detection methods based on deep learning in agricultural planting: a systematic literature review," *IEEE Access*, vol. 13, pp. 96092–96110, 2025, doi: 10.1109/ACCESS.2025.3573364.

[5] B. Xiao, M. Nguyen, and W. Q. Yan, "Fruit ripeness identification using YOLOv8 model," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 28039–28056, Aug. 2023, doi: 10.1007/s11042-023-16570-9.

[6] B. Xiao, M. Nguyen, and W. Q. Yan, "Fruit ripeness identification using transformers," *Applied Intelligence*, vol. 53, no. 19, pp. 22488–22499, Oct. 2023, doi: 10.1007/s10489-023-04799-8.

[7] C. C. Ukwuoma, Q. Zhiguang, M. B. B. Heyat, L. Ali, Z. Almaspoor, and H. N. Monday, "Recent advancements in fruit detection and classification using deep learning techniques," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–29, Jan. 2022, doi: 10.1155/2022/9210947.

[8] N. Ismail and O. A. Malik, "Real-time visual inspection system for grading fruits using computer vision and deep learning techniques," *Information Processing in Agriculture*, vol. 9, no. 1, pp. 24–37, Mar. 2022, doi: 10.1016/j.inpa.2021.01.005.

[9] A. Hayat, F. M. -Dias, T. Choudhury, T. P. Singh, and K. Kotecha, "FruitVision: a deep learning based automatic fruit grading system," *Open Agriculture*, vol. 9, no. 1, May 2024, doi: 10.1515/opag-2022-0276.

[10] Y. Song, C. A. Glasbey, G. W. Horgan, G. Polder, J. A. Dieleman, and G. W. A. M. V. D. Heijden, "Automatic fruit recognition and counting from multiple images," *Biosystems Engineering*, vol. 118, pp. 203–215, Feb. 2014, doi: 10.1016/j.biosystemseng.2013.12.008.

[11] S. Amatya, M. Karkee, A. Gongal, Q. Zhang, and M. D. Whiting, "Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting," *Biosystems Engineering*, vol. 146, pp. 3–15, Jun. 2016, doi: 10.1016/j.biosystemseng.2015.10.003.

[12] M. Sarigu *et al.*, "Phenotypic identification of plum varieties (*Prunus domestica* L.) by endocarps morpho-colorimetric and textural descriptors," *Computers and Electronics in Agriculture*, vol. 136, pp. 25–30, Apr. 2017, doi: 10.1016/j.compag.2017.02.009.

[13] M. A. Peña, R. Liao, and A. Brenning, "Using spectrotemporal indices to improve the fruit-tree crop classification accuracy," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 128, pp. 158–169, Jun. 2017, doi: 10.1016/j.isprsjprs.2017.03.019.

[14] K. Hameed, D. Chai, and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques," *Image and Vision Computing*, vol. 80, pp. 24–44, Dec. 2018, doi: 10.1016/j.imavis.2018.09.016.

[15] H. M. T. Abbas, U. Shakoor, M. J. Khan, M. Ahmed, and K. Khurshid, "Automated sorting and grading of agricultural products based on image processing," in *2019 8th International Conference on Information and Communication Technologies (ICICT)*, Nov. 2019, pp. 78–81, doi: 10.1109/ICICT47744.2019.9001971.

[16] L. T. Duong, P. T. Nguyen, C. D. Sipio, and D. D. Ruscio, "Automated fruit recognition using EfficientNet and MixNet," *Computers and Electronics in Agriculture*, vol. 171, Apr. 2020, doi: 10.1016/j.compag.2020.105326.

[17] J. N. -Torres, M. Mora, R. H. -García, R. J. Barrientos, C. Fredes, and A. Valenzuela, "A review of convolutional neural network applied to fruit image processing," *Applied Sciences*, vol. 10, no. 10, May 2020, doi: 10.3390/app10103443.

[18] Y. Li, X. Feng, Y. Liu, and X. Han, "Apple quality identification and classification by image processing based on convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, Aug. 2021, doi: 10.1038/s41598-021-96103-2.

[19] N.-E.-A. Mimma, S. Ahmed, T. Rahman, and R. Khan, "Fruits classification and detection application using deep learning," *Scientific Programming*, vol. 2022, pp. 1–16, Nov. 2022, doi: 10.1155/2022/4194874.

[20] A. K. Ratha, N. K. Barpanda, P. K. Sethy, G. Sharada, and S. K. Behera, "Computer intelligence-based fruit grading: a review," *Revue d'Intelligence Artificielle*, vol. 37, no. 2, pp. 465–474, Apr. 2023, doi: 10.18280/ria.370223.

[21] M. Knott, F. P. -Cruz, and T. Defraeye, "Facilitated machine learning for image-based fruit quality assessment," *Journal of Food Engineering*, vol. 345, May 2023, doi: 10.1016/j.jfoodeng.2022.111401.

[22] P. Dhiman, A. Kaur, V. R. Balasaraswathi, Y. Gulzar, A. A. Alwan, and Y. Hamid, "Image acquisition, preprocessing and classification of citrus fruit diseases: a systematic literature review," *Sustainability*, vol. 15, no. 12, Jun. 2023, doi: 10.3390/su15129643.

[23] H. Chen, G. Zhou, W. He, X. Duan, and H. Jiang, "Classification and identification of agricultural products based on improved MobileNetV2," *Scientific Reports*, vol. 14, no. 1, Feb. 2024, doi: 10.1038/s41598-024-53349-w.

[24] P. Zhao, W. Zhou, and L. Na, "High-precision object detection network for automate pear picking," *Scientific Reports*, vol. 14, no. 1, Jun. 2024, doi: 10.1038/s41598-024-65750-6.

[25] F. X. Dono, B. N. Baatuuwie, F. K. Abagale, and P. B. Sørensen, "Application of computer vision and machine learning in morphological characterization of Adansonia digitata fruits," *Smart Agricultural Technology*, vol. 9, Dec. 2024, doi: 10.1016/j.atech.2024.100528.

[26] Y. Zhang *et al.*, "Fruit freshness detection based on multi-task convolutional neural network," *Current Research in Food Science*, vol. 8, 2024, doi: 10.1016/j.crfs.2024.100733.

[27] S. Munera, A. R. -Ortega, S. Cubero, N. Aleixos, and J. Blasco, "Automatic detection of pomegranate fruit affected by blackheart disease using X-ray imaging," *LWT*, vol. 215, Jan. 2025, doi: 10.1016/j.lwt.2024.117248.

[28] M. Rajini and P. Voola, "Developing an IoT and ML-driven platform for fruit ripeness evaluation and spoilage detection: a case study on bananas," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 11, Mar. 2025, doi: 10.1016/j.prime.2025.100896.

[29] B. Biswas, S. K. Ghosh, and A. Ghosh, "A robust multi-label fruit classification based on deep convolution neural network," in *Computational Intelligence in Pattern Recognition*, Singapore: Springer, 2020, pp. 105–115, doi: 10.1007/978-981-13-9042-5_10.

[30] S. Espinoza, C. Aguilera, L. Rojas, and P. G. Campos, "Analysis of fruit images with deep learning: a systematic literature review and future directions," *IEEE Access*, vol. 12, pp. 3837–3859, 2024, doi: 10.1109/ACCESS.2023.3345789.

[31] P. Sahu, A. P. Singh, A. Chug, and D. Singh, "A systematic literature review of machine learning techniques deployed in agriculture: a case study of banana crop," *IEEE Access*, vol. 10, pp. 87333–87360, 2022, doi: 10.1109/ACCESS.2022.3199926.

[32] M. Mukhiddinov, A. Muminov, and J. Cho, "Improved classification approach for fruits and vegetables freshness based on deep learning," *Sensors*, vol. 22, no. 21, Oct. 2022, doi: 10.3390/s22218192.

[33] L. Colaco and P. Kamat, "Artificial intelligence advances for cashew fruit maturity and quality detection: a systematic review on models, sensors, and farming applications," *Journal of Big Data*, vol. 12, no. 1, Nov. 2025, doi: 10.1186/s40537-025-01296-2.
[34] Y. Safari, J. N. -Nabende, R. Nakasi, and R. Nakibuule, "A review on automated detection and assessment of fruit damage using machine learning," *IEEE Access*, vol. 12, pp. 21358–21381, 2024, doi: 10.1109/ACCESS.2024.3362230.
[35] V. Meshram and K. Patil, "FruitNet: Indian fruits image dataset with quality for machine learning applications," *Data in Brief*, vol. 40, Feb. 2022, doi: 10.1016/j.dib.2021.107686.
[36] V. Meshram and K. Patil, "FruitNet: Indian fruits dataset with quality (good, bad & mixed quality)," *Mendeley Data, V1*, 2021, doi: 10.17632/b6fftwbr2v.1.

BIOGRAPHIES OF AUTHORS



Divyashree Sreerangaiah is pursuing her Ph.D. at Department of Computer Science and Engineering, SJB Institute of Technology, Bengaluru, and is an assistant professor in the Department of Computer Science and Engineering (Cyber Security) at RNS Institute of Technology, Bengaluru. She holds M.Tech. and B.E. degrees from Visvesvaraya Technological University, Belagavi. Her research interests include machine learning, deep learning, and image processing. She can be contacted at email: divyashrees2008@gmail.com.



Krishna Alabujanahalli Neelegowda is a professor and head of the Department of Computer Science and Engineering at SJB Institute of Technology, Bengaluru, with extensive teaching and research experience. He holds B.E. and M.E. degrees from University Visvesvaraya College of Engineering, Bengaluru, and Ph.D. from Visvesvaraya Technological University, Belagavi. His research interests include image processing, computer vision, deep learning, and artificial intelligence. He can be contacted at email: drkrishnaan@gmail.com.



Anitha Arekattedoddi Chikkalingaiah is an assistant professor at the Department of Computer Science and Engineering, Government Sri Krishnarajendra Silver Jubilee Technological Institute Bengaluru, India. She holds an M.Tech. and B.E. from Visvesvaraya Technological University, Belagavi, and conducts research in machine learning and image processing. She can be contacted at email: anithakrishna2008@gmail.com.