



Improving Citrus Fruit Classification with X-ray Images Using Features Enhanced Vision Transformer Architecture

Syed Mudassir Raza¹ · Awais Raza² · Mohamed Ibrahim Abdallah Babeker¹ · Zia-Ul Haq³ · Muhammad Adnan Islam⁴ · Shanjun Li¹

Received: 23 May 2024 / Accepted: 15 July 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024, corrected publication 2024

Abstract

Quality assessment is a cornerstone of fruit production and distribution, particularly regarding storage conditions and duration. Citrus fruits, a staple in global consumption patterns, are the ultimate example. This study employs a nondestructive analytical technique, X-ray computed tomography (CT) scanning, to meticulously analyze a substantial sample of 300 citrus fruits, specifically satsuma, subjected to both ambient (20–22 °C, 50–60% humidity) and refrigeration conditions (6–8 °C, 65–75% humidity). The experiment was conducted through a methodologically rigorous approach, stratified dataset splitting, allocating 60% of the X-ray datasets for training, with 20% dedicated to validation and testing, respectively. The proposed research introduces a pioneering methodology termed features enhanced vision transformer (FEViT), meticulously designed to augment precision in citrus fruit classification and more precise freshness level prediction via X-ray image analysis. Our empirical findings unequivocally demonstrate the superior efficacy of FEViT models vis-a-vis conventional ViT counterparts across new X-ray citrus fruit datasets. Particularly noteworthy are the marked performance gains exhibited by FEViT-large variants, evidenced by notable increases in precision (5.08%), accuracy (5.47%), recall (4.55%), and F1 scores (5.28%) over original variants. This underscores the distinguishable enhanced discriminatory prowess of FEViT models in assessing citrus fruit quality in terms of freshness. Extensive validation through rigorous experimentation ratifies FEViT's supremacy over traditional deep learning architectures, affirming heightened accuracy (99.25%). The current study heralds the advent of FEViT architecture as a milestone in citrus fruit (satsuma) freshness prediction, promising augmented accuracy and robustness vis-a-vis extant methodologies. This research holds profound implications for the agricultural sector, especially in domains such as citrus and broader fruit classification, where nuanced image analysis is indispensable for quality attribute like freshness evaluation and informed decision-making.

Keywords Vision transformer · Non-destructive analysis · X-ray CT · Citrus fruits · Image classification

✉ Shanjun Li
shanjunlee@webmail.hzau.edu.cn

Syed Mudassir Raza
syedraza@webmail.hzau.edu.cn

Awais Raza
awaisraza@swjtu.edu.cn

Mohamed Ibrahim Abdallah Babeker
mohamedibrahim@webmail.hzau.edu.cn

Zia-Ul Haq
zia.ch@uair.edu.pk

Muhammad Adnan Islam
adnanislam632@zju.edu.cn

¹ College of Engineering, Huazhong Agricultural University, Wuhan Hubei 430070 NX, China

² School of Civil Engineering, South West Jiaotong University, Chengdu 610065, Sichuan, China

³ Faculty of Agricultural Engineering and Technology, PMAS-Arid Agricultural University, Rawalpindi 46300, Pakistan

⁴ College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, Zhejiang, China

Introduction

Citrus fruits play a vital role in the global food industry, with their consumption and demand increasing year by year (Iqbal et al. 2018). Accurate classification of citrus fruits is crucial for ensuring quality control and meeting market demands. Supplementary non-destructive techniques such as near-infrared spectroscopy (Hsiao et al. 2021), NMR (Cakmak 2019), X-ray, electronic nose, ultrasound, machine vision, and hyperspectral imaging (Wieme et al. 2022) have been used for fruit quality evaluation using deep learning models. These techniques' application has proven results in accurately evaluating fruit quality and their contribution to reducing the time, cost, and food losses associated with destructive methods of fruit evaluation (Pathmanaban et al. 2019). X-ray technology is superior in fruit imaging evaluation because it provides detailed internal information, enabling precise detection of defects and abnormalities (Semenov and Mitelman 2020). Using X-ray CT in fruit quality evaluation is a potential application of machine vision technology (Khan et al. 2022b). The image datasets generated after X-ray interaction with citrus fruit (satsuma) will help determine fruit classification based on its extracted feature by vision transformer modeling (Pathmanaban et al. 2019). X-ray images are reasonably priced and accessible; even in developing nations, sophisticated digital radiography equipment is within reach (Qin et al. 2018).

In recent years, object detection, image segmentation, and classification are just a few of the computer vision-related tasks that convolutional neural networks (CNNs) and deep learning approaches have proven to be quite successful at (Li et al. 2021). Different deep learning vision methods have become powerful tools in various industries, including agriculture and food processing (Sivaranjani et al. 2022). Numerous fields have substantially increased interest in computer vision, including medical imaging and agricultural applications (Zhuang et al. 2018). Advancements in computer vision have led to automated fruit classification using deep learning models like vision transformers (Zeeshan et al. 2020). Vision transformer (ViT), a more modern adaptation of the original transformer concept, shows its dominance over CNNs on various vision tasks (Yuan et al. 2021; Khan et al. 2022a). It allows for a more accurate analysis of fruit attributes and enhances the overall classification process (Siddiqi 2020; Tripathi and Maktedar 2020). With the help of vision transformers, citrus fruits can be classified more consistently and objectively, eliminating human error and subjective judgments. Applying self-attention in ViT architecture to capture global and local dependencies within fruit images allows for a better understanding of complex patterns and variations (Dosovitskiy et al. 2020).

The research paper addresses challenges related to citrus fruit classification by leveraging the vision transformer model for superior accuracy and efficiency (Bhargava and Bansal 2021). Automating the process with computer vision transformers technology can analyze fruits' quality features, enabling faster and more accurate classification than traditional methods. Advancements in computer vision have led to automated fruit classification using deep learning models like vision transformer (Dosovitskiy et al. 2020). Dharmasiri and Jayalal (2019) presented a method for detecting fruit disease in agriculture using feature extraction and image processing techniques. It used support vector machines, decision trees, and artificial neural networks as machine learning techniques. Bhargava and Bansal (2021) examined techniques in computer vision for grading and classifying fruits and vegetables according to appearance, highlighting the need for improved performance and suggesting exploring color spaces and different image directions. Using computer vision and machine learning, Habib et al. (2021) described recent advancements in fruit and vegetable disease detection, compared performance measures to pinpoint cutting-edge methods, and recommended future lines of inquiry. Koç and Vatandaş (2021) developed an image processing algorithm for classifying fruits based on size and color characteristics, with training success rates of 93.6% for KNN, 90.3% for DT, 88.3% for Naive Bayes, 92.6% for MLP, and 94.3% for RF. Si et al. (2019) examined how fruit-picking robots use computer vision technology, examining construction, workflow, imaging equipment, preprocessing algorithms, localization methods, and future research directions. Joseph et al. (2021) proposed a deep learning model for the classification of fruits, achieving an accuracy of 94.35% in 131 fruit and vegetable classes, trained on the Fruits 360 dataset and using TensorFlow backend and 50 epochs. Mandal et al. (2022) showed a deep learning framework for fruit classification, comparing two architectures and achieving an average classification performance of 0.9688 on 1200 images. Al-Shawwa and Abu-Naser (2020) suggested that fruit classification was crucial for import and export in agriculture. The Gaussian filter removes image noise, and Convolutional Neural Network, AlexNet, and MobileNetV2 are used for type classification and quality checking. MobileNetV2 achieves 100% accuracy for orange and apple defects.

Mimma et al (2022) presented automated fruit classification and detection systems using deep learning algorithms, achieving high accuracy using ResNet50 and VGG16 models and implementing an Android smartphone application for real-time detection. Nugraha et al. (2019) studied how fruit and vegetable pores affect gas transport by using X-ray computed tomography, which led to reduced internal O₂ and CO₂ concentrations. Fermentation in apples could cause hypoxic or anoxic stress, cell death, and internal disorders. Kakani et al. (2020) analyzed the emerging technologies

like computer vision and AI that were revolutionizing the food industry by leveraging big data, improving operational efficiency, and promoting sustainable food production (Al-Shawwa and Abu-Naser 2020). Apple, a fruit tree with numerous health benefits, was identified using a machine learning-based approach, achieving 96% accuracy in identifying types using a dataset of 8554 images. Mavani et al. (2022) suggested that artificial intelligence (AI) has become increasingly important in the food industry due to rising food demands and population growth. It could be applied to prediction, food classification, control tools, and quality assessment. Bashir et al. (2020) reported that external and internal features like shape, color, ripeness, sugar content, and biochemical composition determined fruit quality. Monitoring could be done manually or using non-destructive methods like optical, magnetic, acoustic, and dynamic techniques. Using depth cameras, Bortolotti et al. (2022) proposed that a computer vision system was tested to measure and weigh peaches at harvest time to enhance fruit selection and boost growers' income. The Intel RealSense D435i camera achieved the best results, with future studies focusing on sizing and color estimation.

The non-destructive assessment of various attributes has become increasingly paramount in fruit quality evaluation, with freshness being a critical determinant of consumer acceptability and shelf life (Bhargava and Bansal 2021). It is highly dependent on the fruit's storage conditions. Reliance on traditional visual inspection or manual testing often fails to accurately capture the intricate cellular structural patterns of fruit. While RGB and NIR imaging techniques offer valuable insights into external appearance, they are inherently limited to surface properties and may overlook internal changes with the passage of storage period proceedings, which is crucial for assessing fruit freshness. X-ray (CT) imaging has emerged as a promising solution to address these limitations (Sivaranjani et al. 2022). Offering high-resolution imaging of cellular structures, X-ray CT outperforms conventional conventional RGB or NIR imaging in various aspects (Cakmak 2019; He et al. 2022). By delving deeper into the fruit's interior, X-ray CT detects internal physiological cellular structural changes or anomalies indicative of ripening stages, defects, or spoilage—details often concealed from the naked eye or surface-based imaging (Arendse et al. 2018). It provides critical insights into fruit physiology, allowing for the identification of internal cellular structural variations within citrus fruits. By capturing these variations as grayscale levels later learned by the proposed features enhanced vision transformer (FEViT) architecture, X-ray radiography reveals the internal anatomy and cellular structural variations associated with fruit freshness levels and deterioration. X-ray imaging is used alongside the FEViT model architecture to classify fruit freshness accurately. This approach integrates advanced vision transformer

techniques (Priya et al. 2020) with X-ray radiography, enabling a comprehensive assessment of fruit freshness levels under varying storage conditions.

The X-ray CT datasets comprising 300 citrus fruit samples subjected to ambient and refrigeration conditions also provide diverse and representative training and evaluation data for the proposed FEViT architecture. Few studies have been reported on the non-destructive study of citrus fruits and their quality analysis attributes. The study was designed with the following objectives:

1. To create a novel dataset for the satsuma citrus fruit variety using X-ray CT scans of two distinct storage types for 10 days.
2. To develop a new architecture, features enhancement vision transformer (FEViT), for a more accurate classification of citrus fruits (satsuma) samples.
3. To study how efficiently the proposed model classifies the citrus fruits (satsuma) into different freshness levels due to varied storage conditions.

The remaining portion of the paper was structured like this: The materials and procedures considering the proposed study were described in the “[Materials and Methods](#)” section. The “[Results and Discussion](#)” section acts for the results and discussion of the datasets. By and large, the “[Conclusion](#)” section contained conclusions.

Materials and Methods

This study presents an architecture for classifying X-ray CT images called the FEViT. On over-image datasets containing images connected to citrus fruit storage type and storage day, the suggested architecture was compared to the original ViT architecture. It was tested on images to confirm our proposed method's generalization and resilience compared to the original ViT architecture.

Dataset Acquisition and Scanner Specifications

The study utilized the NanoVoxel 2000 high-resolution X-ray CT imaging system, comprising two software components: NanoVoxel Scan for scanning and VoxelStudio Recon for image reconstruction. This system seamlessly managed the entire imaging process, from data acquisition to post-processing. The interface of NanoVoxel Scan and Voxel Studio Recon (Gou et al. 2021) is depicted in Figs. 1 and 2, respectively. Scanning was conducted at Huazhong Agricultural University, Wuhan, China, with the fruit samples positioned on a rotation platform for axial radiographs. During scanning, X-rays interacted with the fruit samples, producing distinct attenuation patterns captured by the

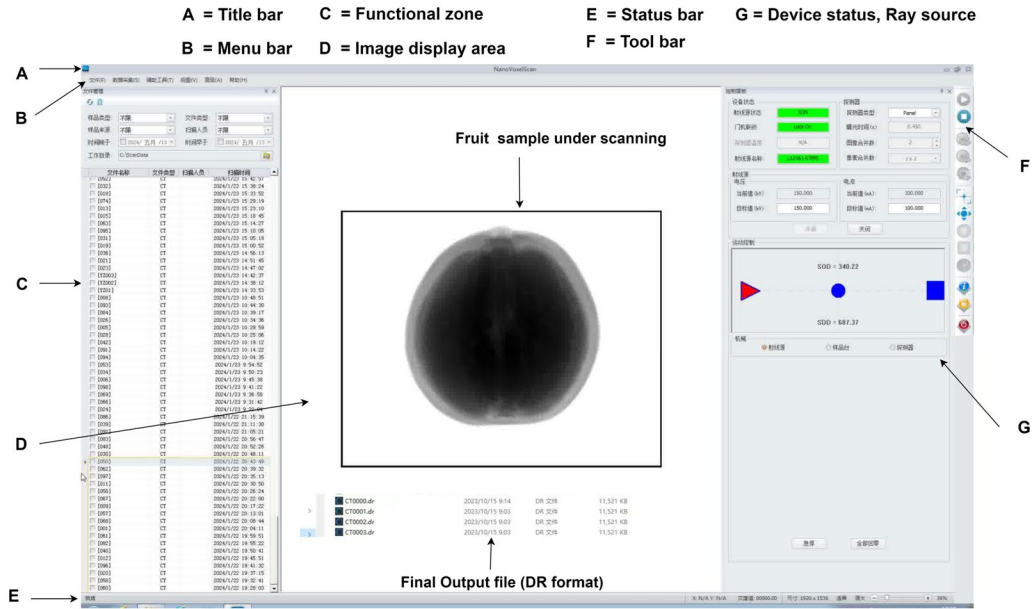


Fig. 1 User interface for NanoVoxel scan system

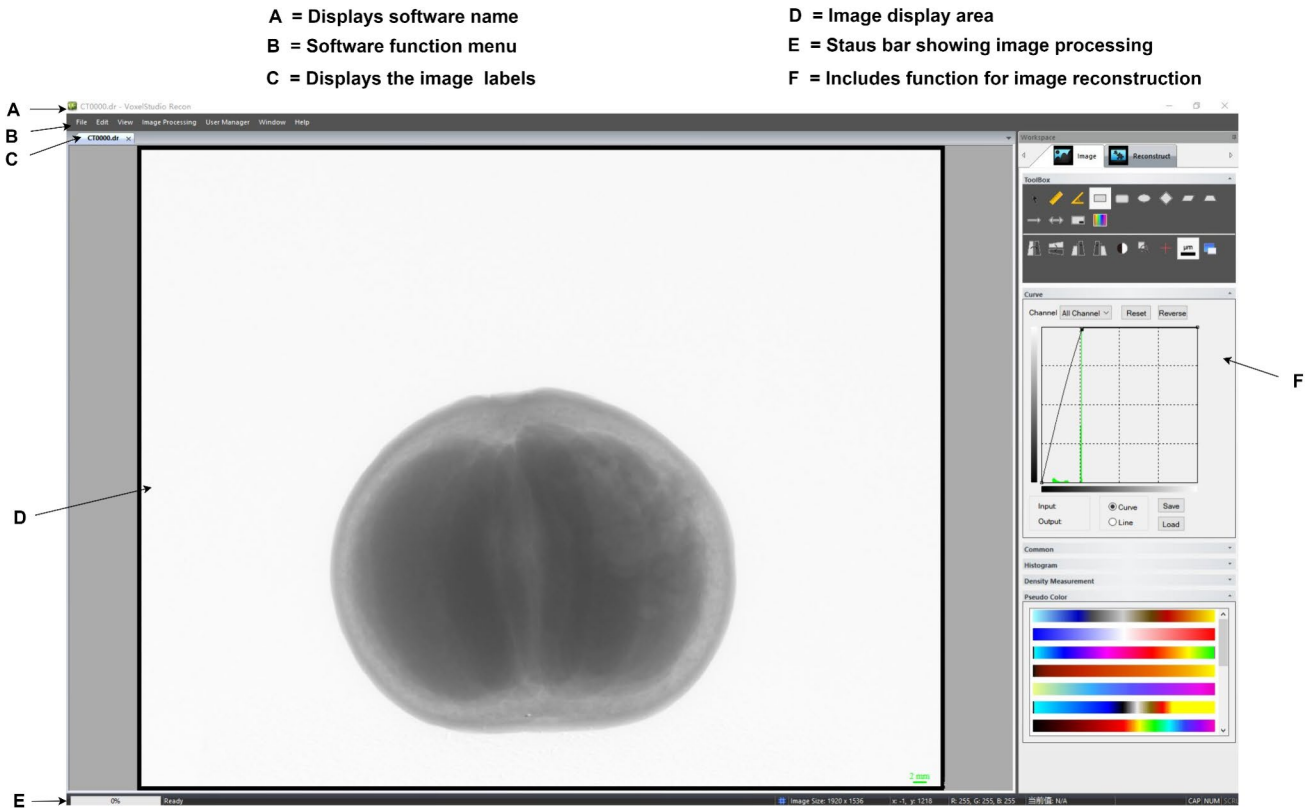


Fig. 2 Voxel Studio Rencon: front-end user interaction

detector. The resulting radiographs provided comprehensive insight into the internal structure of the fruits, enabling further analysis and classification. The X-ray CT system allowed adjustable resolution, rotation degree, and image count per scan, generating digital radiograph (DR) formats (Table 1). These DR formats were converted into jpg format using VoxelStudio Recon software, ensuring consistency and facilitating precise analysis (Gou et al. 2021). High-end computing hardware is crucial for computer vision applications. Model training and validation were performed using GPUs, i-7 processor, 64 GB RAM, Professional Pycharm, TensorFlow, Keras DL frameworks, and Python programming (Ngugi et al. 2021). The lab team analyzed and annotated fruit sample datasets for 2 weeks, identifying various citrus photos and partitioning them according to the proposed architecture's operational requirements.

Satsuma fruits were harvested from Huazhong Agricultural University's orchard, cleaned, dried, and stored. Three hundred samples were sorted and stored under ambient

(20–22 °C, 50–60%) and refrigeration (6–8 °C, 65–75%) treatments for 10 days.

After scanning all fruit samples, four data augmentation methods were employed on original datasets to generate more training images. These methods included operations like rotation with a range of 20°, width shift with a range of 0.2, height shift with a range of 0.2, and horizontal flipping enabled, applied using the “ImageDataGenerator” to the original dataset to create diverse and augmented images for training ViT and FEViT models. Each method increased the dataset's size, diversity, and quality, enhancing the model's performance and generalization capabilities. The data augmentation application on datasets generates five augmented images for each original image as shown in Table 2.

The augmented dataset, enriched with diverse images, was used for model training and analysis, potentially improving the model's performance and learning robust features. The dataset was divided into three subsets: 60% for training, 20% for validation, and 20% for testing, with each subset

Table 1 Technical X-ray CT scanner variables

Main technical indicators	Parameters value or description
Spot size	5 mm
Spatial resolution	3 μm
Detectability	1 μm
Multiple scanning imaging modes	2D perspective detection, circular trajectory scanning, spiral scanning
Closed tube radiation source	Maximum tube voltage of 150 kV; maximum current of 500 μA ; maximum power 30 W
Flat panel detector	2496 \times 3008 pixels, 16 bit, 65,536 grayscale imaging field of view 249.6 mm \times 300.8 mm, pixel size 100 μm
Sample stage	$\pm 360^\circ$
Wavelength range	100.01 nm
X-ray source and fruit sample mutual distance	31.2 cm
Fruit sample and detector mutual distance	38.7 cm

Table 2 Distribution of image data across storage periods

Storage category	Total samples/day (Amb + Ref)	Number of images/sample	Total images/day	Augmented images/sample
Day 1	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 2	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 3	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 4	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 5	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 6	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 7	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 8	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 9	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400
Day 10	15 + 15 = 30	36 (1920 \times 1536)px ¹	1080	5400

¹px stands for pixels:1920 pixels width and 1536 pixels height

²Rotation angle = 10°

maintained in two folders: ambient and refrigeration. This stratified approach ensures an accurate evaluation of the model's performance. Each image was labeled following a specific pattern day condition labeled following a specific pattern as illustrated in Figure 3 to assist efficient mapping to its relevant metadata.

$$\langle \text{condition} \rangle \langle \text{day} \rangle \langle \text{sample} \rangle \quad (1)$$

The Proposed Feature Enhancement Vision Transformer (FEViT) for Citrus Fruit Classification

The vision transformer (Wu et al. 2020) is an architecture for computer vision tasks that applies the principles of transformers, developed initially for natural language processing (Dosovitskiy et al 2020; Kameswari et al. 2023), to image data. With this deep learning architecture, computer vision tasks like segmentation, object identification, and image classification are completed at the cutting edge. The vision transformers function by first converting an image into a series of patches, which are subsequently processed by self-attention mechanisms. The ViT can efficiently extract local and global information from picture collections thanks to this sequence-to-sequence method (Turner 2023). The number of attention heads, the dimensions of the patch embeddings, the number of layers in the self-attention mechanism, the size of the feedforward network, and the regularization dropout rate are among the parameters that make up the hyperparameter configuration for the ViT. Generally, the transformer has viewed every patch as a unique token. Consequently, from an image of y size $C \times H \times W$, where H is the height, W is the breadth, and C is the number of channels, patches with dimensions of $C \times P \times P$ were recovered.

A series of N -length patches was created (y_1, y_2, \dots, y_n), where $N = HW/P^2$. Semantic segmentation and fruit picture classification were two computer vision tasks that employed the high-level representation of ViT models.

The proposed FEViT (feature enhanced vision transformer) architecture builds upon the standard ViT model by incorporating three CNN blocks parallel to the transformer encoder. These CNN blocks were designed to extract additional features from the input image, complementing the feature extraction capabilities of the transformer encoder. Each of the three CNN blocks starts with a different number of filters, ranging from 64 to 256, to capture features at multiple scales. The CNN blocks consist of convolutional layers followed by max-pooling layers, allowing for hierarchical feature extraction and increasing the receptive field of the network. The output of each CNN block was then concatenated with the corresponding output of the transformer encoder, ensuring that the network has access to both the transformer-based features and the CNN-based features. By incorporating these CNN blocks into the ViT architecture, FEViT significantly increased the number of features available for image classification (Dosovitskiy et al. 2020).

The transformer model processed input data in two steps. First, it divided the image into patches and converted each patch into a vector. Then, it combined these vectors with positional information. This combined data was projected into a lower-dimensional space, which helped the model focus on the most important features while retaining spatial information. Figure 4 shows the suggested model's entire end-to-end architecture sections. Patches were processed using a multi-head self-attention mechanism, generating attention maps indicating patch importance. These maps were used to weigh the importance of each patch for further processing. In the suggested study, patch embeddings

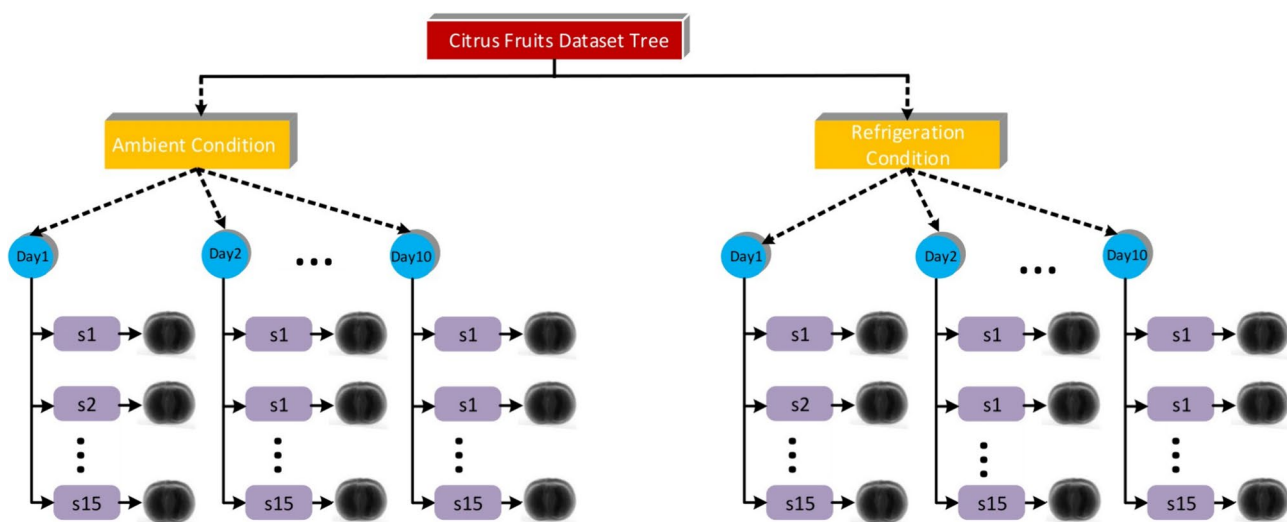


Fig. 3 Schematic representation of citrus fruit based on freshness levels

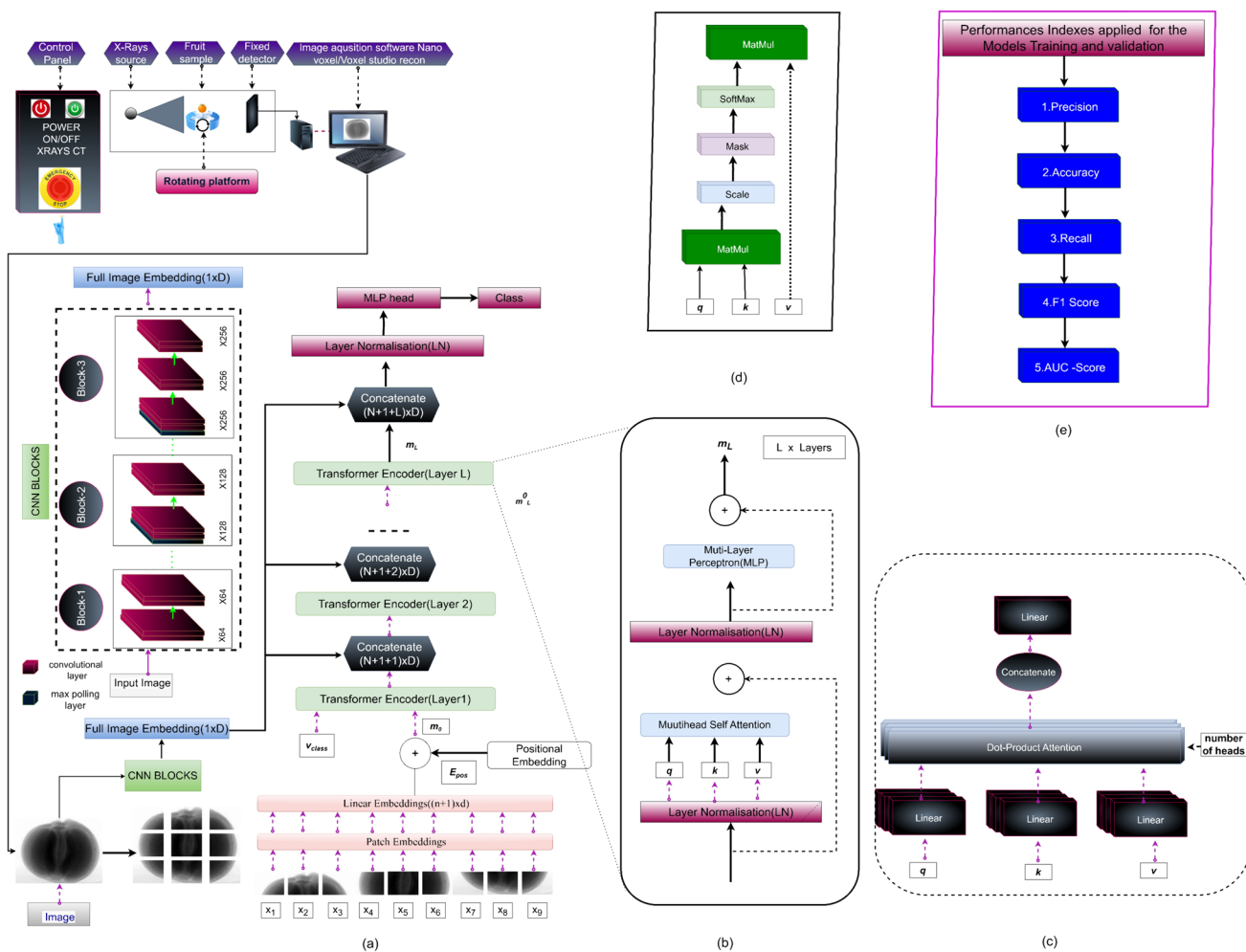


Fig. 4 The proposed FEViT model architecture. **(a)** Main model architecture. **(b)** The encoder module for Transformers **(c)** represents the multi-scale-self attention (MSA) head. **(d)** The self-attention (SA) head. **(e)** Evaluation metrics used

served as input tokens for the transformer encoder, incorporating position information for ViTs. Positional embeddings are added for spatial integration during training. The transformer encoder used patch embeddings with positional information to process images, capturing spatial dependencies and classifying citrus fruit images. Patch embedding involved processing patches using self-attention. They were then linearly projected onto a model dimension vector using a learned embedding matrix and fed into the encoder. The transformer classified images using a learnable token and concatenates embedding representations. It viewed patches as a group, transforming them into a lower-dimensional feature space for easier analysis (Turner 2023). The positional data E_{pos} was encoded and attached to patch representations to maintain patch spatial arrangement, as shown in Eq. 2.

$$m_0 = [v_{class}; y_1 E; y_2 E; \dots; y_n E] + E_{pos}, E \in \mathbb{R}^{(P^2 C) \times d}, E_{pos} \in \mathbb{R}^{(N+1) \times d} \quad (2)$$

After completion of linear embedding, the series of embedded patches m_0 was generated and sent to the transformer encoder. The vision transformers comprised several transformer encoder layers (L). The proposed model (FEViT) comprised feedforward dense blocks, GeLU activation, and multi-head self-attention blocks, with sub-components using residual skip connections and normalization layers as shown in Eq. 4.

$$m_{\ell'} = \text{MSA}(\text{LayerNormalization}(m_{\ell} - 1)) + m_{\ell} - 1, \ell = 1 \dots L \quad (3)$$

$$m_{\ell} = \text{MLP}(\text{LayerNormalization}(m_{\ell'})) + m_{\ell'}, \ell = 1 \dots L \quad (4)$$

The feedforward MLP used a transformer encoder to capture local and global details and a multi-head self-attention mechanism to attend to different parts of the input image simultaneously. The initial element in the sequence (m_L^0) was

transmitted to an external head classifier at the encoder's final layer to anticipate the specific class label (Han et al. 2022).

$$y = \text{LayerNormalization} \cdot m_L^0 \quad (5)$$

The MSA block, consisting of four layers, linear, self-attention, concatenation, and linear, was evaluated by each patch embedding in a series near the transformer's center as shown by Fig. 4c. The attention weight, a sophisticated attention representation, was calculated by summing the weights of all values in a sequence using the self-attention head. The specifics of the calculation that occurs in the MSA block are displayed in Fig. 4b. As seen in Eq. 6, three values q (query), k (key), and v (value) were retrieved for each element in the input sequence. through multiplying every element by the three U_{qkv} learnt matrices. As demonstrated in Fig. 4c, the dot product between an element's "q" vector and the "k" vectors of other items in the sequence was computed to determine the element's relevance to different elements. The relevance of an element to other elements was determined by computing the dot product between its "q" vector and the "k" vectors of other items in the sequence. The study evaluated patch importance in a sequence using dot-product outputs scaled and fed into a softmax classifier, considering the essential dimension of D_K as shown in Equation 7. Next, using Equation 7, the value of each patch embedding's vector was multiplied by the softmax output. To determine which patch received the most attention points, these equations give the complete operation:

$$[q, k, v] = mU_{qkv}, U_{qkv} \in \mathbb{R}^{d \times 3D^k} \quad (6)$$

$$A = \text{softmax}\left(\frac{qk^T}{\sqrt{D_K}}\right), A \in \mathbb{R}^{n \times n} \quad (7)$$

$$SA(m) = A \cdot V \quad (8)$$

Using the preceding technique, the MSA block independently computed the scaled dot-product attention for each of the h heads. However, instead of a single value, many values were utilized for the query, key, and value. The outputs from each attention head were concatenated and projected to the target dimension after a feed-forward layer with learnable weights W . This equation could be used to express this operation.

The MSA block computed scaled dot-product attention for h heads using multiple values for query, key, and value, concatenated and projected to the target dimension after a feed-forward layer with learnable weights. This equation was used to express this operation:

$$MSA(m) = \text{Concat}(SA_1(m); SA_2(m); \dots SA_h(m))W, W \in \mathbb{R}^{h \cdot D_k \times D} \quad (9)$$

The transformer encoder produces feature vectors for each patch, with a classification head estimating image class. The vision transformer captures visual patterns and relationships, training models on 16×16 and 32×32 patches. The recommended models were initialized using pre-trained weights on ImageNet (Nayeem et al. 2022; Russakovsky et al 2015), with other components randomly initialized and fine-tuned through end-to-end training and classifier selection. The proposed architecture, trained using supervised learning on large datasets, was fine-tuned for citrus fruit classification, unlike CNN-based models, which detect information locally and globally. The vision transformer function, used in classifying citrus fruits into different freshness levels, determines relevant features for categorization with the cross-entropy loss. Table 3 describes the best parameters applied during the training and validation process.

Evaluation Metrics for Proposed Algorithm

The results of model performance were illustrated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The following metrics assessed each model's performance (training and validation) in classification: accuracy, recall, precision, AUC, and F1 scores.

Accuracy Accuracy is the percentage of accurately predicted observations in a given dataset. It is among the most accessible measures for classification. A decision to classify is not wise if there is an imbalance in the classes. It is also important to mention that all input photographs were resized to 224×224 .

Precision The precision measures the proportion of accurately forecasted observations as positive to all anticipated positives. High precision means the false-positive rate is low (Powers 2020).

$$\text{PRECISION} = \frac{TP}{TP + FN} \times 100 \quad (10)$$

Recall The percentage of accurately predicted positive observations to all real positive considerations is known as recall. High recall indicates a model that can find all the positive samples (Powers 2020).

$$\text{RECALL} = \frac{TP}{TP \times FN} \times 100 \quad (11)$$

F1 Score It is also known as the F score, denoted by F1. It represents the model's accuracy on any dataset (Powers 2020). The equation is given below:

$$F \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \quad (12)$$

Table 3 Hyper-parameter configuration for best performance

Operation	Description	Strength/value
Epoch	Represented one complete pass through the training data; determined the number of times the entire dataset was shown to the model during training	200
Batch size	The number of samples or examples propagated through the neural network at one time during training; larger batch sizes generally lead to more stable convergence and faster training	32
Patch size	Allowed separating input images or frames into smaller, manageable parts, which was beneficial for computational efficiency and handling large-scale data	16 / 32
Optimizer	The optimizer was responsible for adjusting the parameters of the model (such as weights and biases) iteratively during the training phase to minimize the error or loss function	Adam
Learning rate	Controls the step size or magnitude of weight updates during training; a higher learning rate may lead to faster convergence, but too high can cause divergence	0.0001
Weight-decay	A regularization technique that introduced a penalty term for large weights, helping to prevent overfitting by encouraging smaller weight values	0.00001
Patience	A hyperparameter was used in early stopping, which stopped the training process if the model's performance on a validation set did not improve after a specified number of epochs	10
random state	Setting a specific seed value that reproduced the same results every time running the code was essential for testing, debugging, and verifying the model's performance	42
Learning rate drop factor	A tuning parameter in an optimization algorithm that determined the step size at each iteration towards a minimum of a loss function	0.2
Verbose	Applied to initialize the random number generator, ensuring the reproducibility of results by controlling the randomness in operations like weight initialization and data shuffling	1

Table 4 Specifications of the ViT variants and proposed modifications in the FEViT variants

Models	Patch size	Layers	Hidden size	MLP size	Heads	Parms
ViT-base-16	16 × 16	12	768	3072	12	86 M
ViT-base-32	32 × 32	12	768	3072	12	87.8 M
ViT-large-16	16 × 16	24	1024	4096	16	304.5 M
ViT-large-32	32 × 32	24	1024	4096	16	305.5 M
FEViT-base-16	16 × 16	12	768	3072	12	91 M
FEViT-base-32	32 × 32	12	768	3072	12	93 M
FEViT-large-16	16 × 16	24	1024	4096	16	310 M
FEViT-large-32	32 × 32	24	1024	4096	16	313.55 M

ROC Curve AUC-ROC, the area under the ROC curves, is a performance indicator for classification issues. It demonstrates how well the model can distinguish between classes. The AUC (Maxwell et al. 2017) for single-comparing is proportional to how well the model predicts 0s as 0s and 1s as 1s.

Results and Discussion

The proposed FEViT architecture and original variants were used to precisely predict and classify fruit samples' (satsuma) freshness using X-ray datasets. Base-16, base-32, large-16, and large-32 variants were applied to datasets during the training and validation (Table 4).

Table 5 Classification performances of original ViT variants and proposed FEViT variants on our dataset

Models	Accuracy	Precision	Recall	F1 score	AUC score
ViT-base-16	0.92533	0.91471	0.92645	0.92054	0.8934
ViT-base-32	0.91502	0.92479	0.92501	0.92491	0.9176
ViT-large-16	0.93241	0.92632	0.93045	0.92838	0.9148
ViT-large-32	0.94334	0.93444	0.93865	0.93654	0.9196
FEViT-base-16	0.97635	0.97471	0.96645	0.97056	0.9501
FEViT-base-32	0.98026	0.97479	0.98501	0.97987	0.9400
FEViT-large-16	0.98333	0.98586	0.98321	0.98453	0.9858
FEViT-large-32	0.99256	0.98378	0.98469	0.98423	0.9954

Model Performance Evaluation

Table 5 reports the models' performance metrics of the original ViT variant and the investigated FEViT variant on the observed datasets. It showed that the FEViT-large-32 model had the most outstanding performance for the dataset, with an F1 score of 98.42%, an accuracy of 99.25%, recall of 98.37%, and precision of 98.46%. The "FEViT" models consistently outperform the "ViT" models in terms of AUC scores, with the "FEViT-large-32" model achieving the highest AUC score of 0.9951. Additionally, within the "ViT" models, the "ViT-large-32" model has the highest AUC score of 0.9196.

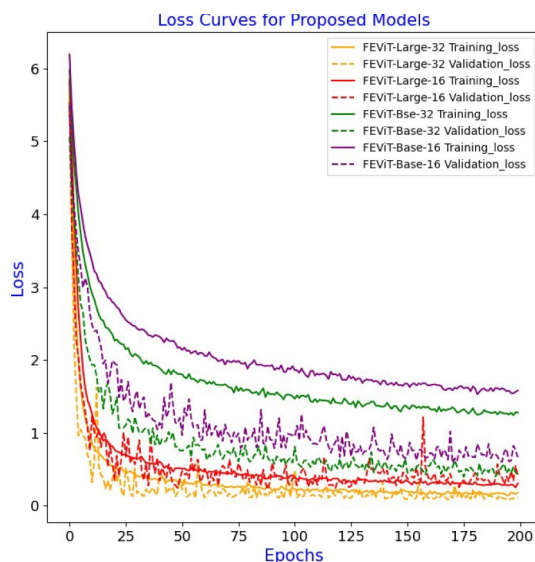
As can be seen from Table 6, the performance differences for preset evaluation metrics are shown. Every variant of the proposed FEViT model outperformed its matching ViT model variant for every examined measure. The graphical representation of the differences between the evaluation metrics values, precision, accuracy, recall, F1 score, and AUC score, is shown in Fig. 10. With an average improvement of +4.88% across all variations, the realized improvement in

terms of F1 score varies from +3.81% for the base-16 version to +6.52% for the large-32 variant.

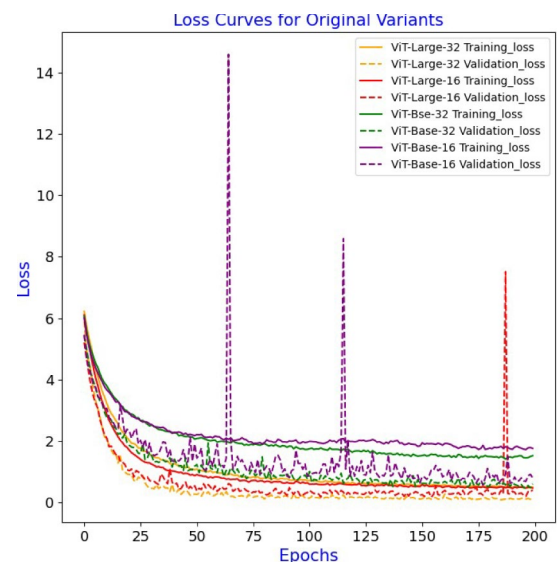
The findings showed stability for all models studied, with an average F1 score for the original ViT variations falling between 92.54 and 93.86% and for the variants of the proposed FEViT model falling between 97.02 and 98.46% (Figs. 5 and 6). When evaluating a model's performance in classifying different class labels, an analysis tool called a confusion matrix shows how well the model did. The confusion matrices for the ViT and suggested FEViT models are displayed in Figs. 7 and 8. It was evident by the confusion matrix plots that the efficiency of fruit samples classified into different freshness levels was much better for the FEViT model than other variants. In this investigation, the epoch number was restricted to 200. The loss curve's oscillations can be further reduced by increasing the number of epochs. However, the duration of the tests would be greatly extended if the number of iterations increased. The training and validation loss curves for original ViT and FEViT models are displayed in Fig. 5. The proposed FEViT architecture demonstrated remarkable

Table 6 Original ViT and proposed FEViT evaluation metrics: comparative analysis

Models	ViT		FEViT- δ		ViT		FEViT- δ		ViT		FEViT- δ	
	B/16	B/16	B/16	B/16	B/32	B/32	L/16	L/16	L/32	L/32	L/32	L/32
Precision	0.9253	0.9764	3.81	0.9150	0.9803	6.52	0.9324	0.9833	5.09	0.9433	0.9926	4.91
Accuracy	0.9145	0.9747	6.02	0.9248	0.9748	5.01	0.9263	0.9859	5.95	0.9344	0.9838	4.93
Recall	0.9265	0.9665	4.01	0.9250	0.9850	6.01	0.9305	0.9832	5.27	0.9387	0.9847	4.60
F1 score	0.9205	0.9706	5.01	0.9249	0.9799	5.76	0.9284	0.9845	5.62	0.9365	0.9842	4.76
AUC score	0.9286	0.9686	4.01	0.9222	0.9757	5.34	0.9386	0.9758	3.71	0.9347	0.9864	5.16



(a)



(b)

Fig. 5 **a** Act for the loss curve for the proposed FEViT models. **b** Act for the loss curve for the original ViT models.

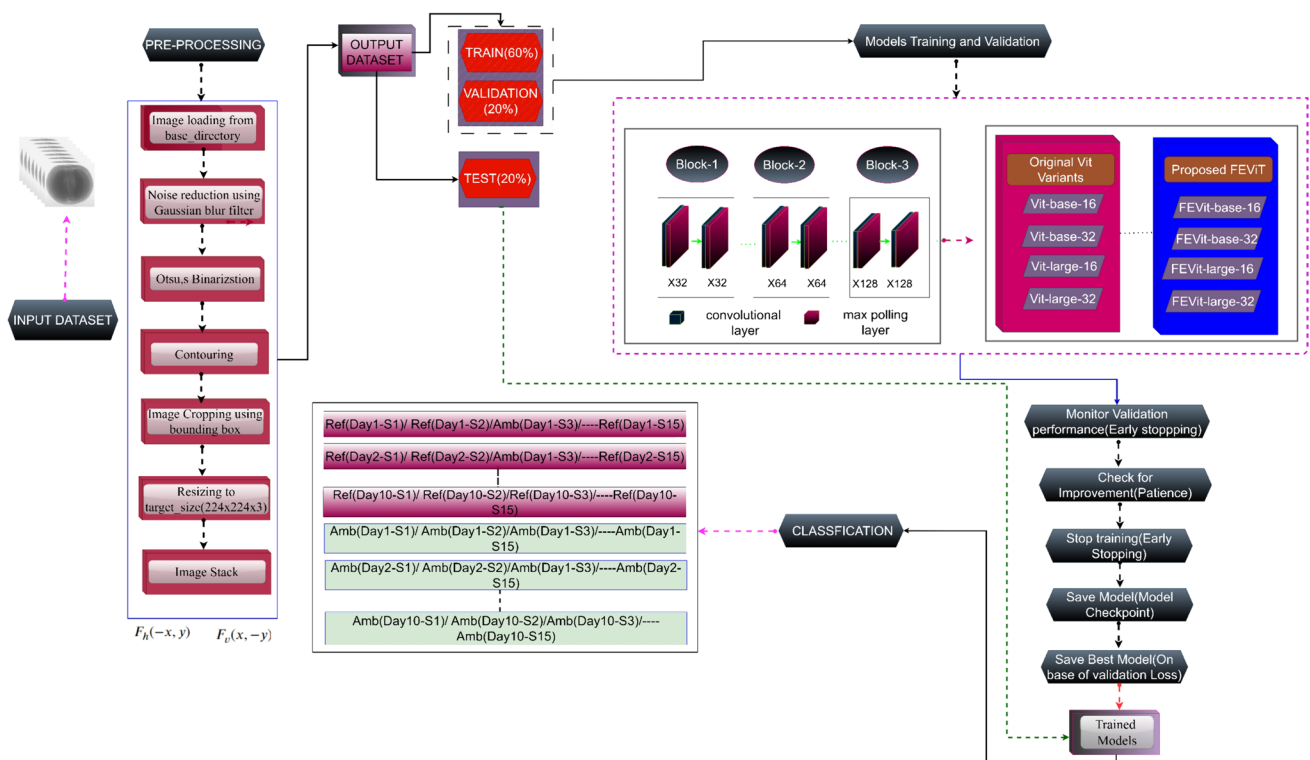


Fig. 6 Datasets processing techniques and algorithm working flowchart

performance, outperforming the original ViT variants across various model configurations. The validation loss curves revealed a significant reduction in loss values, indicating superior generalization and convergence capabilities. Specifically, the FEViT-base-16 achieved a validation loss of 0.0243, an impressive 82.2% improvement over the ViT-base-16’s 0.1363 loss. Similarly, the FEViT-base-32 loss of 0.0421 represented an 84.9% decrease compared to the ViT-base-32’s 0.2788 loss. FEViT-large-16 and FEViT-large-32 experienced a sharp decline in loss values, followed by a gradual fall as the training iterations grew steadily. Approximately 100 epochs through, the loss curve of the FEViT-large-16 and FEViT-large-32 started to avoid oscillations. But, few oscillations were observed during the loss curves of FEViT-base-16 and FEViT-base-32.

The larger model variants exhibit even more substantial gains. The FEViT-large-16 achieved a validation loss of 0.0071, a staggering 94.2% reduction from the ViT-large-16’s 0.1232 loss. Remarkably, the FEViT-large-32 attains the lowest validation loss of 0.0042, an astonishing 98.3% improvement over the ViT-large-32’s 0.2437 loss. These results underscore the efficacy of the proposed FEViT architecture, demonstrating its ability to learn more robust and generalizable representations, ultimately leading to superior performance on the validation set.

Methodological Framework of the Ablation Study

The obtained experimental results showed that upon adding the CNN block to the ViT architecture, each transformer encoder layer’s output was concatenated with its output. For every assessed dataset and variant, FEViT consistently performed better in categorization than ViT. The three CNN blocks’ performance was assessed by building a model that included just the MLP head and the CNN block to confirm the block’s additional contribution to the performance gain and investigate whether using ViT in conjunction with the CNN blocks named FEViT on its own. Classification results for other deep learning models are provided in Table 7.

These findings clearly showed that the performance of the deep learning models was not superior to that of the vision transformer variations. This study further ascertains the applicability of the proposed model; the performance of seven popular CNN models was evaluated using the newly developed citrus fruit dataset. All models were initialized using weights pretrained on the ImageNet dataset, end-to-end training, and the same training settings (including data augmentation) for the ViT and FEViT trials. It should be noted that the default Keras implementations of the researched CNN models were used for the experimental evaluation.

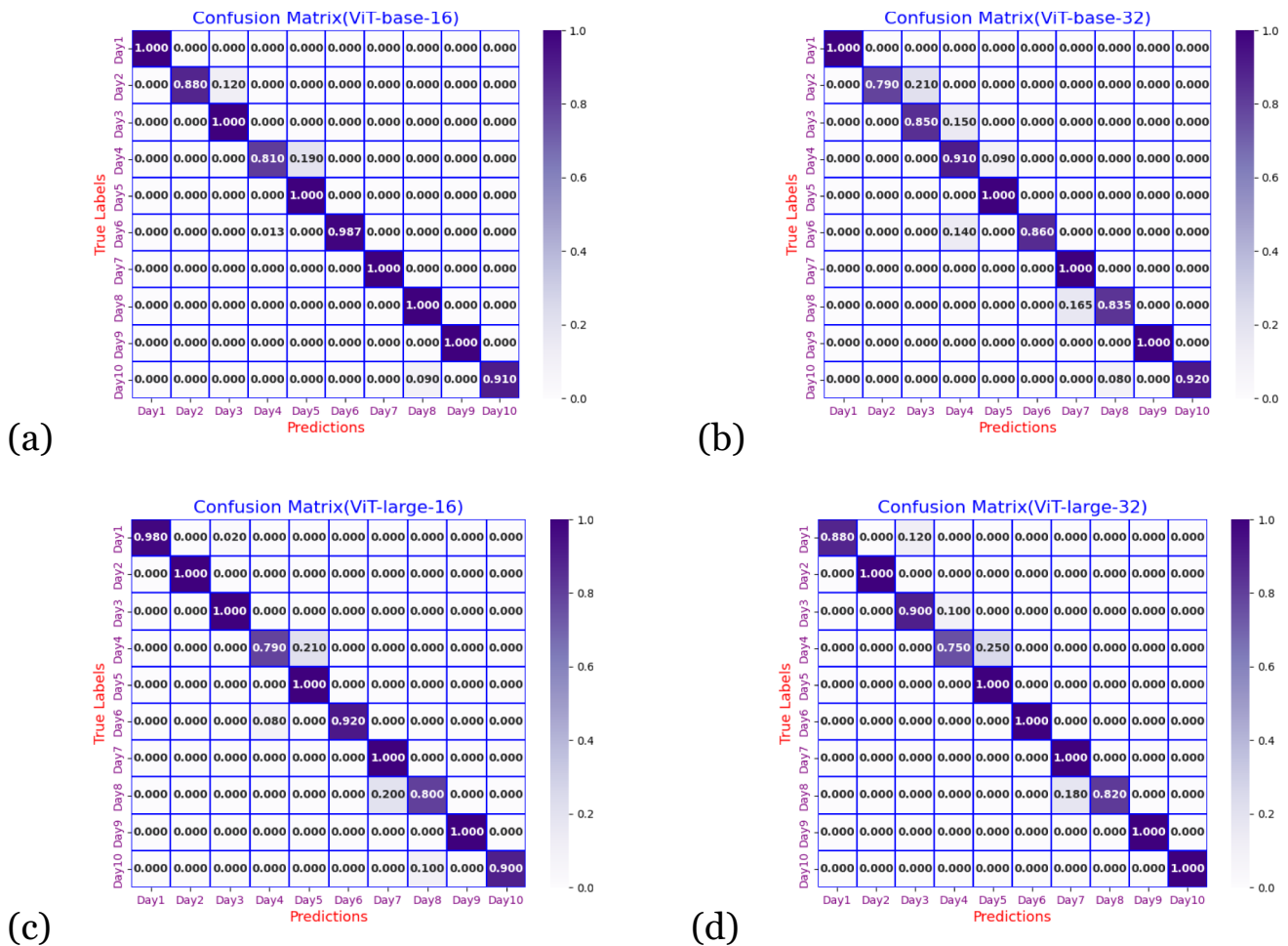


Fig. 7 **a** Act for the confusion matrix of ViT-base-16. **b** Act for the confusion matrix of ViT-base-32. **c** Act for the confusion matrix of ViT-large-16. **d** Act for the confusion matrix of ViT-large-32.

Comparative Analysis with Existing Deep Learning Models

Understanding how different deep learning models perform on diverse tasks requires evaluating them. In this research, the performance of several state-of-the-art models was assessed, including VGG19, DenseNet, EfficientNetB6, ResNetRS420, Reg-NetY320, NASNetMobile, and Xception, alongside the proposed feature enhancement vision transformer (FEViT) models: FEViT-base-16, FEViT-base-32, FEViT-large-16, and FEViT-large-32 and original variants of ViT. VGG19 and DenseNet demonstrated competitive performance across multiple metrics among traditional deep learning architectures. VGG19 achieved an accuracy of 0.96586 and a precision of 0.97296, while DenseNet exhibited an accuracy of 0.97417 and a precision of 0.97646. These deep learning models showcased high precision and accuracy, indicating their effectiveness in classification tasks. In contrast, EfficientNetB6, ResNetRS420,

and RegNetY320 displayed relatively lower performance metrics, with accuracies ranging from 0.06583 to 0.33341. These models exhibited lower precision, recall, and F1 scores than VGG19 and DenseNet, suggesting potential limitations in their generalization capabilities or model complexity as shown in Fig. 9b.

The evaluation of FEViT models unveiled promising results, showcasing competitive performance compared to traditional and modern deep learning architectures. FEViT-base-16, FEViT-base-32, FEViT-large-16, and FEViT-large-32 achieved accuracies ranging from 0.97635 to 0.99256, with precision values exceeding 0.98378 in most cases. Interestingly, FEViT models performed better regarding F1 scores and recall, indicating their robustness in capturing both positive and negative instances within the dataset (Fig. 9a).

Furthermore, FEViT models exhibited notable AUC scores, highlighting their ability to discriminate between classes effectively. For instance, FEViT-large-32 attained an

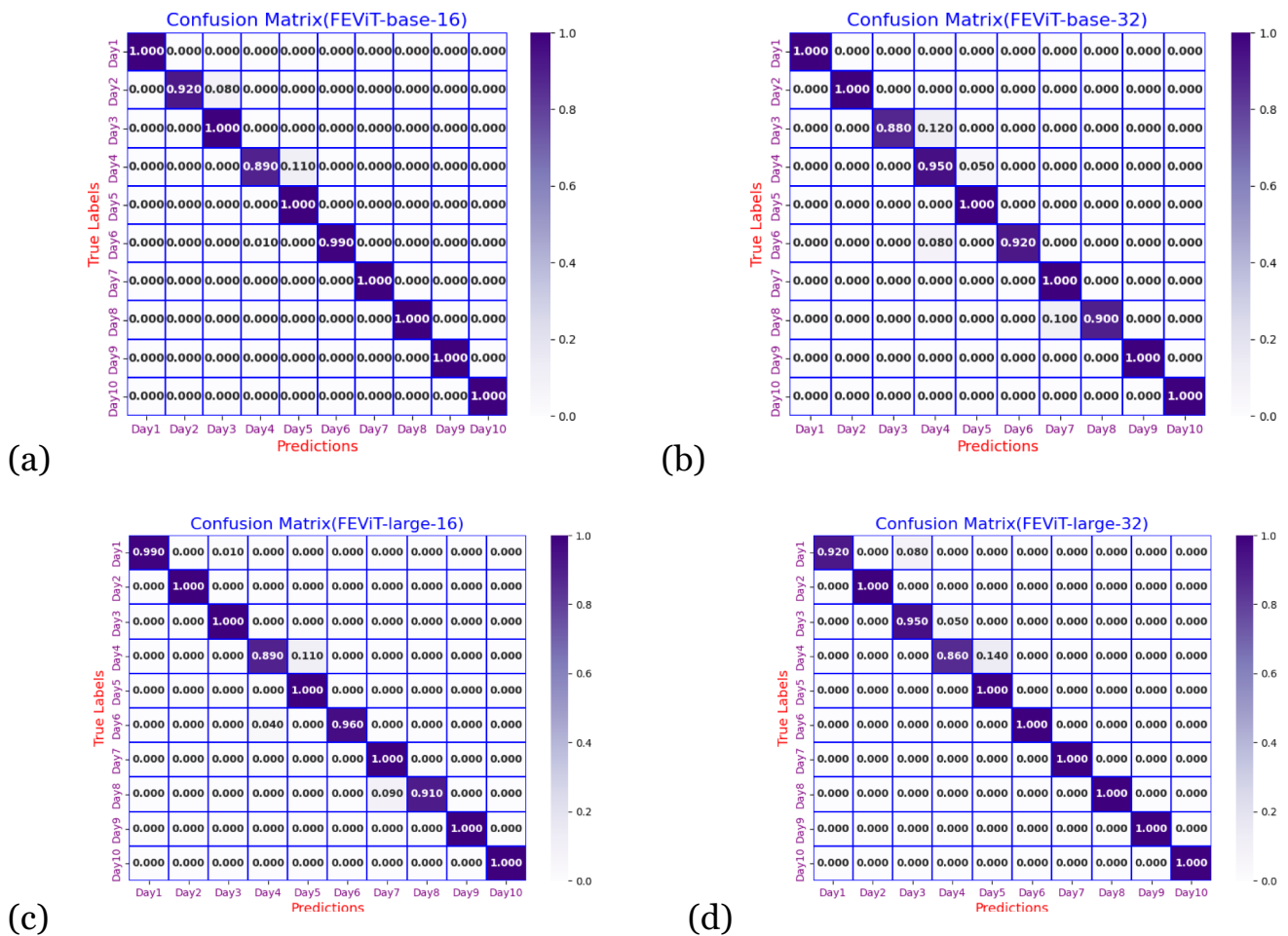


Fig. 8 **a** Act for the confusion matrix of FEViT-base-16. **b** Act for the confusion matrix of FEViT-base-32. **c** Act for the confusion matrix of FEViT-large-16. **d** Act for the confusion matrix of FEViT-large-32.

Table 7 Classification performances for deep learning models

Models	Accuracy	Precision	Recall	F1 score	AUC score
VGG19	0.96586	0.97296	0.96583	0.96339	0.9999
DensNet121	0.97417	0.97646	0.97421	0.97208	1.0000
Efficient-netB6	0.33341	0.11245	0.32654	0.36982	0.49301
ResNetRS420	0.06583	0.04141	0.06583	0.04098	0.64302
RegNetY320	0.07250	0.0336	0.07240	0.03493	0.63301
NASNetMobile	0.92250	0.92989	0.92253	0.91525	0.90401
Xception	0.91001	0.92237	0.91010	0.90154	0.93001

AUC score of 0.9954, outperforming several traditional deep learning models in this aspect. The evaluation results suggest that FEViT models, particularly FEViT-large variants, offer a compelling alternative to conventional deep learning architectures. Their superior performance across various evaluation metrics underscores their potential for diverse

computer vision applications, including image classification and semantic segmentation.

The suggested FEViT improvement increases the computational complexity of the ViT architecture. Since the input to the transformer encoder layers is enlarged and convolution layers are added, the recommended models have more trainable parameters than the original ViT models. The FEViT”base” variants have over 5.5 million more trainable parameters on average than the equivalent ViT”base” variants, according to Table 4. At the same time, the FEViT”large” versions have an average of 7.5 million more. However, as Fig. 10 illustrates, higher performance resulted from its complexity (number of parameters) rise in every case, with some showing significant increases (up to 4 + 0.62% in the F1 score for the large-32 variation). Numerous precautions were made to prevent overfitting to guarantee the suggested models’ capacity for generalization. The models were trained and tested on distinct datasets for each dataset. The final test set was entirely unseen during the training process.

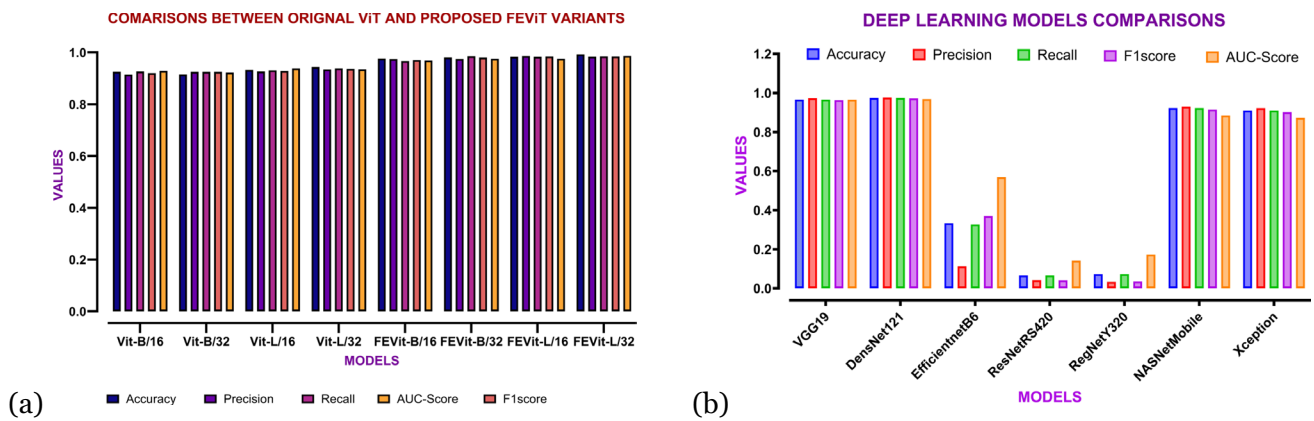
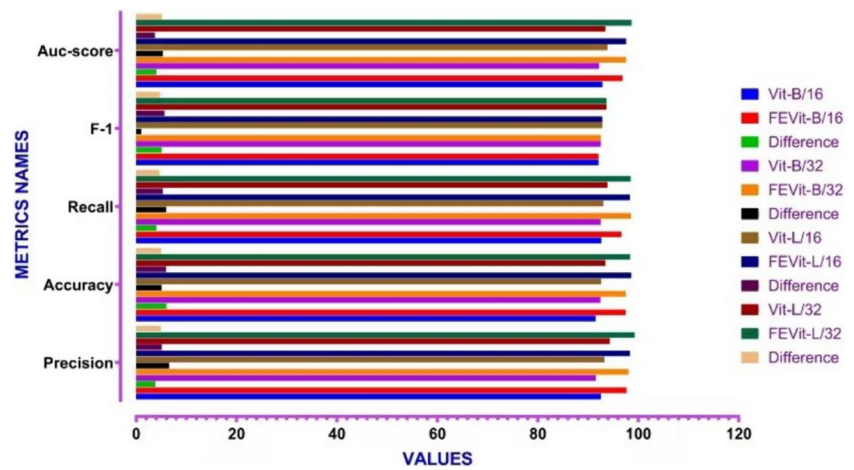


Fig. 9 a Visualizing performance differences via ViT and FEViT bar charts. b Analyzing deep learning techniques through bar graph visualization

Fig. 10 Representation for mutual differences in metrics for ViT and FEViT architectures



Because data augmentation added variance to the training set and had been demonstrated to enhance deep learning models' capacity for generalization while lowering overfitting Kumar et al (2023), its usage throughout the training process assisted in addressing overfitting even further. It was also demonstrated that applying the utilized label smoothing strategy reduced overfitting and improves generalization (Barbedo 2022). Datasets were used to train and evaluate the original ViT variations and the suggested FEViT variants. Both models performed consistently across datasets, and FEViT regularly outperformed ViT in every dataset. The proposed FEViT model improved upon the state-of-the-art original ViT model for image classification, consistently outperforming ViT in categorizing all images in datasets through transfer learning with weights pre-trained on ImageNet. Transfer learning simplifies FEViT models for various image classification tasks, eliminating the requirement to train the model on massive datasets. This is possible since the original ViT model has pretrained weights available. With this in mind, as well as the published experimental

findings, it is clear that the suggested FEViT model represents a potent solution for categorizing citrus fruit X-ray images based on different freshness levels, consistently outperforming original ViT model variants.

Comparison with the Literature

We collected recent literature employing dedicated models and examining a single fruit or general models applied to various fruit representations. Table 8 compares the general model of this study and models suggested by related works. It is evident that the proposed study, focusing on citrus fruit (satsuma), achieved a remarkable accuracy of 99%. This accuracy surpassed the performance of various existing methods in the literature. The deep CNN models referenced in the literature, such as those by Rodríguez et al. (2018), Mithun et al. (2018), Kumar et al. (2021), and Aherwadi et al. (2022), demonstrated accuracies ranging from 81.75 to 98.% across different fruit types and objectives, including plum ripeness calculation, banana ripeness detection, quality assessment of

Table 8 Related work comparisons

Method	Fruit	Objective	Accuracy
Deep CNN, Rodríguez et al. (2018)	Plum	Calculating plum ripeness from images	91–97%
Deep CNN, Mithun et al (2018)	Banana	distinguishing between naturally ripened bananas	98.74%
Deep CNN, Kumar et al (2021)	Multiple fruits	Quality assessment	95%
Deep CNN, Aherwadi et al (2022)	Banana	Prediction of banana fruit maturity, quality, and its life	81.75–98.25%
E-Alexnet, Ni et al (2021)	Strawberry	Quality evaluation	95.75%
CNNs, Naik and Desai (2022)	Mango	Grading analysis	91.43%
Faster-RCNN, Wan and Goudos (2020)	Mango, apple, orange	Fruit detection and quality estimation	91%
CNN, Wan and Goudos (2020)	Apple	Online detection of defective apples	92%
ResNet50, Fan et al (2020)	Tomatoes	Detection of external defects on tomatoes	94.6%
Proposed study	Citrus fruit (satsuma)	Discern between freshness levels of satsumas under different storage	99%

multiple fruits, and prediction of banana fruit maturity and quality. Other models, such as E-Alexnet (Ni et al. 2021; Naik and Desai 2022), Faster-RCNN (Wan and Goudos 2020), and ResNet50 (Fan et al. 2020), had shown accuracies ranging from 91 to 95.75%, focusing on quality evaluation, grading analysis, fruit detection, and defect detection on various fruits like strawberries, mangoes, apples, oranges, and tomatoes. In comparison, the proposed study achieved higher accuracy and addressed the specific context of discerning between citrus fruit samples (satsuma) belonging to different freshness levels, which adds significant value in practical applications in the food industry, especially in the context of fruit (external and internal disorders) quality and supply chain management. These results underscore the efficacy of the proposed approach in the context of Citrus fruit analysis, potentially offering enhanced accuracy and applicability compared to existing methods in the literature. Our model effectively classifies citrus fruit freshness levels, demonstrating the potential for automated food processing industries.

Conclusion

In this study, the vision transformer (ViT) architecture was employed to assess and predict the freshness levels of satsuma citrus fruit using X-ray imaging under varying storage conditions. Leveraging the innovative features enhanced vision transformer (FEViT) architecture, it accurately interpreted internal cellular structural change patterns or features associated with fruit freshness, demonstrating its precise analysis and classification capability. The FEViT architecture effectively learned and captured these feature changes inside X-ray images that happened during various storage periods and conditions. Experimental investigation highlights the profound impact of storage conditions on fruit quality in terms of freshness over time. Notably, the FEViT-large-32 variant consistently

outperformed traditional deep learning architectures and standard ViT models in citrus fruit quality nondestructive assessment and freshness prediction. Despite their increased computational demands, FEViT models exhibited significant performance improvements ranging from +3.81 to +6.52%, underscoring their effectiveness and efficiency.

In conclusion, the FEViT model emerges as a robust and effective solution for citrus fruit freshness prediction and evaluation through image classification. Its superiority over traditional deep learning methods and ViT models reaffirms its potential in various computer vision applications, where accurate classification is crucial for informed decision-making. Future research directions could explore the extension of FEViT models to other agricultural sectors and investigate optimization techniques to enhance further their efficiency and scalability in classifying industrial fruits for quality assessment and freshness prediction.

Author Contribution Conceptualization, S.M.R. and S.L.; methodology, S.M.R.; software, A.R.; validation, M.A.I; formal analysis, M.I.A.B; investigation, S.M.R.; resources, S.L; data curation, Z.U.H.; writing—original draft preparation, S.M.R.; writing—review and editing, S.M.R. and S.L.; visualization, A.R; supervision, S.L; project administration, S.L; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding Special funds supported this work from the Construction of Industrial Technology System of Modern Agriculture (Citrus) (CARS-26) and the Fundamental Research Funds for the Central Universities (2662022YJ005).

Data Availability Datasets will be made available from corresponding author upon reasonable request.

Declarations

Ethical Approval Not applicable.

Conflict of Interest The authors declare no competing interests.

References

- Aherwadi N, Mittal U, Singla J et al (2022) Prediction of fruit maturity, quality, and its life using deep learning algorithms. *Electronics* 11(24):4100
- Al-Shawwa MO, Abu-Naser SS (2020) Classification of apple fruits by deep learning. *Int J Acad Eng Res (IJAER)* 3(12)
- Arendse E, Fawole OA, Magwaza LS et al (2018) Non-destructive prediction of internal and external quality attributes of fruit with thick rind: a review. *J Food Eng* 217:11–23
- Barbedo JGA (2022) Data fusion in agriculture: resolving ambiguities and closing data gaps. *Sensors* 22(6):2285
- Bashir S, Jabeen A, Makroo H et al (2020) Application of computer vision system in fruit quality monitoring. In: *Sensor-based quality assessment systems for fruits and vegetables*. Apple Academic Press, p 267–290
- Bhargava A, Bansal A (2021) Fruits and vegetables quality evaluation using computer vision: a review. *J King Saud Univ-Comput Inf Sci* 33(3):243–257
- Bortolotti G, Mengoli D, Piani M et al (2022) A computer vision system for in-field quality evaluation: preliminary results on peach fruit. In: *2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, IEEE, pp 180–185 <https://doi.org/10.1109/MetroAgriFor55389.2022.9965022>
- Cakmak H (2019) Assessment of fresh fruit and vegetable quality with non-destructive methods. In: *Food quality and shelf life*. Elsevier, p 303–331 <https://doi.org/10.1016/B978-0-12-817190-5.00010-0>
- Dharmasiri S, Jayalal S (2019) Passion fruit disease detection using image processing. In: *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, IEEE, pp 126–133 <https://doi.org/10.23919/SCSE.2019.8842799>
- Dosovitskiy A, Beyer L, Kolesnikov A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:201011929*
- Fan S, Li J, Zhang Y et al (2020) On line detection of defective apples using computer vision system combined with deep learning methods. *J Food Eng* 286:110102
- Gou L, Li S, Yin J et al (2021) Morphological and physico-mechanical properties of mycelium biocomposites with natural reinforcement particles. *Constr Build Mater* 304:124656
- Habib MT, Arif MAI, Shorif SB et al (2021) Machine vision-based fruit and vegetable disease recognition: a review. *Comput Vis Mac Learn Agric* 143–157
- Han K, Wang Y, Chen H et al (2022) A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell* 45(1):87–110
- He Y, Xiao Q, Bai X et al (2022) Recent progress of nondestructive techniques for fruits damage inspection: a review. *Crit Rev Food Sci Nutr* 62(20):5476–5494
- Hsiao WT, Kuo WC, Lin HH et al (2021) Assessment and feasibility study of lemon ripening using X-ray image of information visualization. *Appl Sci* 11(7):3261
- Iqbal Z, Khan MA, Sharif M et al (2018) An automated detection and classification of citrus plant diseases using image processing techniques: a review. *Comput Electron Agric* 153:12–32
- Joseph JL, Kumar VA, Mathew SP (2021) Fruit classification using deep learning. In: *Innovations in electrical and electronic engineering: Proceedings of ICEEE 2021*, Springer, pp 807–817
- Kakani V, Nguyen VH, Kumar BP et al (2020) A critical review on computer vision and artificial intelligence in food industry. *J Agric Food Res* 2:100033
- Kameswari CS, J K, Reddy TS et al (2023) An overview of vision transformers for image processing: a survey. *Int J Adv Comput Sci Appl* 14(8). <https://doi.org/10.14569/IJACSA.2023.0140830>
- Khan S, Naseer M, Hayat M et al (2022a) Transformers in vision: a survey. *ACM Comput Surv (CSUR)* 54(10s):1–41
- Khan S, Xiaobo Z, Irfan M et al (2022b) A review of X-ray for detection of insect infestation in fruits and vegetables. *Ann Roman Soc Cell Biol* 26(01):164–173
- Koç DG, Vatandaş M (2021) Classification of some fruits using image processing and machine learning. *Turk J Agric-Food Sci and Technol* 9(12):2189–2196
- Kumar A, Joshi RC, Dutta MK et al (2021) Fruit-cnn: an efficient deep learning-based fruit classification and quality assessment for precision agriculture. In: *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, IEEE, pp 60–65
- Kumar T, Turab M, Raj K et al (2023) Advanced data augmentation approaches: a comprehensive survey and future directions. *arXiv preprint arXiv:230102830*
- Li Z, Liu F, Yang W et al (2021) A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*
- Mandal OS, Dey A, Nath S et al (2022) Fruit-net: fruits recognition system using convolutional neural network. In: *International Conference on Advanced Communication and Intelligent Systems*, Springer, pp 120–133
- Mavani NR, Ali JM, Othman S et al (2022) Application of artificial intelligence in food industry—a guideline. *Food Eng Rev* 14(1):134–175
- Maxwell A, Li R, Yang B et al (2017) Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinform* 18:121–131
- Mimma NE, Ahmed S, Rahman T, et al (2022) Fruits classification and detection application using deep learning. *Scientific Programming* 2022
- Mithun B, Shinde S, Bhavsar K, et al (2018) Non-destructive method to detect artificially ripened banana using hyperspectral sensing and rgb imaging. In: *Sensing for agriculture and food quality and safety X*, SPIE, pp 122–130
- Naik S, Desai P (2022) Mango (*Mangifera indica* L.) classification using convolutional neural network and linear classifiers. In: *Proceedings of Third International Conference on Sustainable Computing: SUSCOM 2021*, Springer, pp 163–172
- Nayeem TA, Motaharuzzaman S, Hoque AT, et al (2022) Computer vision based object detection and recognition system for image searching. In: *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*, IEEE, pp 148–151
- Ngugi LC, Abelwahab M, Abo-Zahhad M (2021) Recent advances in image processing techniques for automated leaf pest and disease recognition—a review. *Inf Proc Agric* 8(1):27–51
- Ni J, Gao J, Li J et al (2021) E-alexnet: quality evaluation of strawberry based on machine learning. *J Food Meas Charact* 15(5):4530–4541
- Nugraha B, Verboven P, Janssen S et al (2019) Non-destructive porosity mapping of fruit and vegetables using X-ray CT. *Postharvest Biol Technol* 150:80–88
- Pathmanaban P, Gnanavel B, Anandan SS (2019) Recent application of imaging techniques for fruit quality assessment. *Trends Food Sci Technol* 94:32–42
- Powers DM (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:201016061*
- Priya PS, Jyoshna N, Amaraneni S et al (2020) Real time fruits quality detection with the help of artificial intelligence. *Mater Today: Proc* 33:4900–4906
- Qin C, Yao D, Shi Y et al (2018) Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomed Eng Online* 17(1):1–23
- Rodríguez FJ, García A, Pardo PJ et al (2018) Study and classification of plum varieties using image analysis and deep learning techniques. *Prog Artif Intell* 7:119–127

- Russakovsky O, Deng J, Su H et al (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vision* 115:211–252
- Semenov V, Mitelman Y (2020) Non-destructive fruit quality control using radioelectronics: a review. In: 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), IEEE, pp 0281–0284
- Si H, Lv J, Lin K et al (2019) A review of application of computer vision in fruit picking robot. *Recent Dev Intell Comput Commun Dev: Proc ICCD* 5:346–355
- Siddiqi R (2020) Comparative performance of various deep learning based models in fruit image classification. In: Proceedings of the 11th International Conference on Advances in Information Technology, pp 1–9
- Sivaranjani A, Senthilrani S, Ashok Kumar B et al (2022) An overview of various computer vision-based grading system for various agricultural products. *J Hortic Sci Biotechnol* 97(2):137–159
- Tripathi MK, Maktedar DD (2020) A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: a survey. *Inform Proc Agric* 7(2):183–203
- Turner RE (2023) An introduction to transformers. arXiv preprint arXiv:230410557
- Wan S, Goudos S (2020) Faster r-cnn for multi-class fruit detection using a robotic vision system. *Comput Netw* 168:107036
- Wieme J, Mollazade K, Malounas I et al (2022) Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: a review. *Biosyst Eng* 222:156–176
- Wu B, Xu C, Dai X et al (2020) Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:200603677
- Yuan L, Chen Y, Wang T et al (2021) Tokens-to-token ViT: training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 558–567
- Zeeshan M, Prabhu A, Arun C, et al (2020) Fruit classification system using multiclass support vector machine classifier. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, pp 289–294
- Zhuang J, Luo S, Hou C et al (2018) Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. *Comput Electron Agric* 152:64–73

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.