scientific reports

Check for updates

OPEN

High precision banana variety identification using vision transformer based feature extraction and support vector machine

Ebru Ergün

Bananas, renowned for their delightful flavor, exceptional nutritional value, and digestibility, are among the most widely consumed fruits globally. The advent of advanced image processing, computer vision, and deep learning (DL) techniques has revolutionized agricultural diagnostics, offering innovative and automated solutions for detecting and classifying fruit varieties. Despite significant progress in DL, the accurate classification of banana varieties remains challenging, particularly due to the difficulty in identifying subtle features at early developmental stages. To address these challenges, this study presents a novel hybrid framework that integrates the Vision Transformer (ViT) model for global semantic feature representation with the robust classification capabilities of Support Vector Machines. The proposed framework was rigorously evaluated on two datasets: the four-class BananaImageBD and the six-class BananaSet. To mitigate data imbalance issues, a robust evaluation strategy was employed, resulting in a remarkable classification accuracy rate (CAR) of 99.86% ± 0.099 for BananaSet and 99.70% ± 0.17 for BananaImageBD, surpassing traditional methods by a margin of 1.77%. The ViT model, leveraging self-supervised and semi-supervised learning mechanisms, demonstrated exceptional promise in extracting nuanced features critical for agricultural applications. By combining ViT features with cutting-edge machine learning classifiers, the proposed system establishes a new benchmark in precision and reliability for the automated detection and classification of banana varieties. These findings underscore the potential of hybrid DL frameworks in advancing agricultural diagnostics and pave the way for future innovations in the domain.

Keywords Agricultural diagnostics, Banana classification, Deep learning, Machine learning, Precision agriculture, Vision transformer

Fruit has always played a central role in global agriculture, serving as an essential component of human nutrition and economic livelihoods[1,2]. Among these, bananas stand out as one of the most important fruits worldwide, not only because of their widespread consumption, but also because of their exceptional nutritional profile, digestibility and cultural importance. As a staple food for millions of people and a major export commodity for many tropical and subtropical regions, bananas make a significant contribution to food security, economic development and sustainable agricultural practices. Their versatility in different forms, from fresh consumption to processed products, underscores their value in agricultural markets and household diets. In modern agricultural practices, the emergence of advanced technologies such as image processing, computer vision and deep learning (DL) has opened new avenues for increasing productivity and ensuring quality control[2,3]. The application of DL to fruit image analysis has been transformative, enabling accurate, automated identification and classification tasks that were previously labor-intensive and error-prone. Fruits, including bananas, have subtle visual characteristics that vary between varieties and stages of development. Accurately capturing and analyzing these characteristics using advanced imaging techniques has become an essential aspect of agricultural diagnostics, ensuring consistent quality and facilitating effective supply chain management.

Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Recep Tayyip Erdogan University, Rize, Turkey. email: ebru.yavuz@erdogan.edu.tr

Fruit variety classification, particularly for bananas, plays a critical role in addressing several agricultural challenges[3]. Correctly identifying fruit types can optimize harvesting processes, improve the accuracy of yield predictions and streamline sorting and packaging operations. It is also essential for breeding programmers to develop disease-resistant, high-yielding varieties and to ensure that consumers receive produce that matches their preferences. Beyond the practical implications, such classification contributes to the scientific understanding of fruit development and variety differentiation, making it a cornerstone of precision agriculture.

Artificial intelligence (AI), particularly in the area of DL, has emerged as a revolutionary tool for solving complex agricultural problems[4]. By integrating AI with traditional farming practices, it is now possible to achieve unprecedented accuracy in tasks such as disease detection, pest monitoring and yield estimation. In the context of fruit classification, AI-based systems equipped with sophisticated algorithms can process vast amounts of image data to extract nuanced features that are often imperceptible to the human eye[5]. These advances not only reduce the reliance on manual labor, but also enable real-time decision making, which is critical to maintaining the competitiveness and sustainability of agricultural businesses[6]. As a highly diverse crop with numerous varieties, bananas present unique challenges for automated classification. Early developmental stages often exhibit subtle differences that require advanced computational techniques to distinguish. This study addresses these challenges by employing a hybrid framework that integrates the Vision Transformer (ViT) model for global feature extraction with Support Vector Machines (SVM) for robust classification. Such an approach leverages the strengths of DL and machine learning (ML), setting a new standard for accuracy and reliability in agricultural diagnostics. The main contributions of this paper are as follows:

- A hybrid framework is proposed, combining the ViT for global semantic feature extraction with SVM for robust classification.
- Comprehensive evaluations are conducted on two datasets, BananaImageBD and BananaSet, addressing the challenges of data imbalance and achieving state-of-the-art classification accuracy rate (CAR).
- Comparative performance analyses are performed by integrating Convolutional Neural Networks (CNN), Multi-Layer Perceptron (MLP) and Gradient Boosting Machines (GBM), highlighting the superiority of the proposed method.
- The framework demonstrates exceptional capability in distinguishing subtle features of banana varieties, especially at early developmental stages, which is critical for agricultural diagnostics.
- The integration of self-supervised and semi-supervised learning within the ViT model enhances feature extraction, paving the way for future advancements in smart agriculture.
- The study establishes a new benchmark for precision and reliability in automated fruit classification, showcasing the potential of hybrid DL frameworks in addressing agricultural challenges.

Related works

In computer vision, image classification remains a central topic, especially in the context of agricultural applications[7,8]. Traditional ML approaches often require manual feature extraction prior to training classifiers, which can be labor intensive and prone to limitations in generalizability. The advent of DL, driven by advances in computational power and big data algorithms, has revolutionized the field, particularly for fruit and vegetable classification tasks[9]. CNNs have been particularly transformative, offering automated feature extraction and classification capabilities that have led to remarkable results in diverse applications, including grading, sorting, variety identification, and disease detection[10]. Among the notable studies in this area, Kunduracioglu et al. made significant contributions by employing various CNNs, including ResNet50, InceptionV4, Xception, DenseNet121, EfficientNetV2_m, and VGG13, to diagnose diseases in apple leaves. Their findings demonstrated that all models achieved high CAR, with EfficientNetV2_m outperforming others by attaining CAR and F1 score of 100.00%[11]. Similarly, in another study, Kunduracioglu et al. explored DL models for tomato leaf disease detection, where Res2Next50 and Res2Net50d exhibited superior performance. Notably, Res2Next50 achieved a CAR of 99.85%, surpassing VGG16 and DenseNet121[12]. Borra et al. introduced a stacked Bi-LSTM-based classification approach aimed at improving the accuracy of fruit classification. Their methodology combined CNNs for feature extraction with Recurrent Neural Networks (RNNs) for feature selection, and finally applied the Stacked Bi- Long Short-Term Memory (LSTM) model for classification. The proposed method outperformed previous algorithms, achieving a precision of 85.40%, a recall (RCL) of 88.30% and an F1 score of 83.90%, demonstrating its robustness in improving classification metrics[13]. Similarly, Ghazal et al. explored the use of hand-crafted visual features, including hue, color Scale-Invariant Feature Transform (SIFT), discrete wavelet transforms, and Haralick features, to address the challenges of multi-class fruit sorting. Evaluated on the Fruits 360 dataset, their approach achieved near-perfect CAR of 99.00–100.00% using Back Propagation Neural Networks, SVM and k-Nearest Neighbors (k-NN) classifiers[7]. Other contributions in this area include the work of Gill et al. who investigated the performance of CNNs, RNNs and LSTM networks in extracting and optimizing image features for fruit classification. Their results were compared with traditional classifiers such as SVM, Artificial Neuro-Fuzzy Inference Systems and Feed-Forward Neural Networks, and showed significant improvements in CAR[14]. Focusing on the classification of specific fruit varieties, Ratha et al. proposed a system for identifying Indian mango varieties using MobileNet-v2 and ShuffleNet-based deep features integrated with various ML classifiers. Among the classifiers, Cubic SVM paired with MobileNet-v2 features yielded a CAR of 99.50% and an AUC of 1, highlighting its potential for high-precision classification[15]. Similarly, Taner et al. applied ML techniques to apple variety classification, evaluating seven CNN architectures. Their study found DenseNet201 to be the most effective, achieving a CAR of 97.48%. Using DenseNet201 features for traditional classifiers, SVM achieved 98.28%, while dimensionality reduction followed by MLP further improved CAR to 99.77%[16]. Another notable contribution was made by Katarzyna et al. who introduced a two-track method for fruit variety classification. Their approach involved two CNNs with identical architectures but different

weight matrices: one trained on full images with backgrounds, and the other trained on regions of interest. Using a certainty factor to aggregate the results, the system achieved an CAR of 99.78% when tested on six apple varieties, demonstrating its robustness in handling ambiguous classifications under variable conditions[17]. Finally, Zaki et al. developed a novel framework for date fruit classification by combining image processing, CNNs and explicable AI techniques. Their approach included normalization, augmentation and evaluation of several transfer learning models, with ResNet50 achieving the highest CAR of 94.00%. In addition, the use of Grad-CAM provided interpretable results by highlighting key image regions, providing actionable insights for agricultural product classification[18].

In the field of agricultural image classification, bananas have a significant nutritional and economic value, making them the focus of various studies[19–21]. Research on bananas has mainly focused on disease detection in banana fruit or leaves. For example, Saranya et al. proposed a CNN-based approach to classify four stages of banana ripeness, overcoming the inherent challenges of differentiating visual features across ripeness stages. Their model, trained on both original and extended datasets, achieved a validation CAR of 96.14%, demonstrating its robustness against state-of-the-art CNNs[22]. Similarly, Mohamedon et al. developed a mobile application using transfer learning and a CNN classifier to identify the ripeness of bananas. This application, tailored for the Malaysian market, achieved 98.25% CAR and used TensorFlow Lite for seamless deployment on Android devices, improving real-time usability[23]. For disease detection in banana leaves, Narayanan et al. introduced a hybrid CNN model to detect diseases early and provide actionable insights to farmers in India. Their approach achieved 99% CAR, significantly outperforming other models, and helped mitigate productivity losses[24]. Thiagarajan et al. addressed the challenges of scale and rotation invariance by integrating Artificial Neural Network with SIFT in one model and Histogram of Oriented Gradients (HOG) with Local Binary Pattern in another. These methods effectively improved early detection of banana leaf diseases by capturing complex local and global patterns[25]. Furthermore, Sujithra et al. investigated the detection of diseases affecting both banana and sugarcane crops using CNN, Feedforward Neural Network and Radial Basis Function Neural Network classifiers. Their CNN model achieved 97.00% CAR for banana diseases, demonstrating its potential for real-time agricultural applications[26].

In addition to disease detection, research was carried out on banana species identification. Vijayalakshm et al. developed a fruit identification model using a five-layer CNN for feature extraction, coupled with RF and k-NN classifiers for classification. Their model, tested on a dataset of 5887 images, achieved 96.98% CAR with the RF, outperforming k-NN and HOG-based methods[27]. Widodo et al. implemented a back-propagation neural network for banana type and ripeness identification, achieving 80.00% CAR for type recognition and 90.00% for ripeness assessment through feature extraction and rigorous training[28]. Finally, Rangkuti et al. compared several CNN architectures, including ResNet50V2, InceptionV3 and EfficientNet, for banana image recognition. Of these, EfficientNet showed superior performance with an CAR of 89.00%, followed by VggNet16 with 83.80%[29].

Despite significant advances in the application of AI to fruit classification and disease detection, the majority of existing studies have focused on specific fruit types or diagnostics, without addressing the broader need for more robust and scalable methods[30]. In particular, banana classification and disease detection research has predominantly focused on either ripeness assessment or leaf disease identification, often overlooking the potential of integrating state-of-the-art feature extraction and classification models. This gap in the literature highlights the need for more innovative approaches that can leverage advanced DL architectures to provide highly accurate and interpretable solutions.

To address this critical gap, in this study proposes a novel methodology that combines ViT-based feature fusion with SVM classification. By leveraging the feature extraction capabilities of ViT and the robust classification performance of SVM, this approach aims to significantly improve the accuracy and reliability of banana classification tasks. We validated proposed methodology on two comprehensive datasets: a 6-class BananaImageBD[31] and a 4-class BananaSet[32] dataset. Features were extracted using ViT after normalization based on the mean and standard deviation of the image datasets. The datasets were then classified using SVM, GBM, MLP and CNN-based classifiers and evaluated under a 5-fold cross-validation (FCV). The proposed approach yielded exceptional results, achieving a CAR of 99.86% $\pm$ 0.099 for BananaSet and 99.70% $\pm$ 0.17 for BananaImageBD. Metrics such as confusion matrices, CAR, F1 score and RCL further validated the effectiveness of the model. These results highlight the transformative potential of proposed methodology in overcoming the limitations of existing approaches, thereby making a significant contribution to the field of agricultural image classification and advancing the state of precision farming.

Materials and methods

Description of dataset

BananaImageBD

Ferdaus et al. developed the BananaImageBD dataset, a comprehensive collection specifically designed to classify banana varieties and detect maturity stages in Bangladesh[31]. This dataset addressed a critical gap in agricultural image resources by providing high quality images of four widely grown banana varieties: Sagor Kola, Shabri Kola, Bangla Kola and Champa Kola. Bananas were sourced from various locations across Bangladesh, including prominent markets such as Karwan Bazar and Rangpur City Bazar, to ensure a representative sample of commonly consumed varieties. Images were taken using high-resolution smartphone cameras under controlled lighting conditions to reduce background noise and minimize color distortion. Each banana was photographed from four different angles to ensure comprehensive coverage of its visual characteristics. Rigorous manual quality checks excluded poor quality images to maintain the high standard of the dataset.

The BananaImageBD dataset contained 2471 original images, organized into sub-folders for each variety. To improve diversity and generalizability, Ferdaus et al. applied enhancement techniques including flipping, rotation, brightness adjustment and noise addition, resulting in a total of 7413 images for variety classification[31].

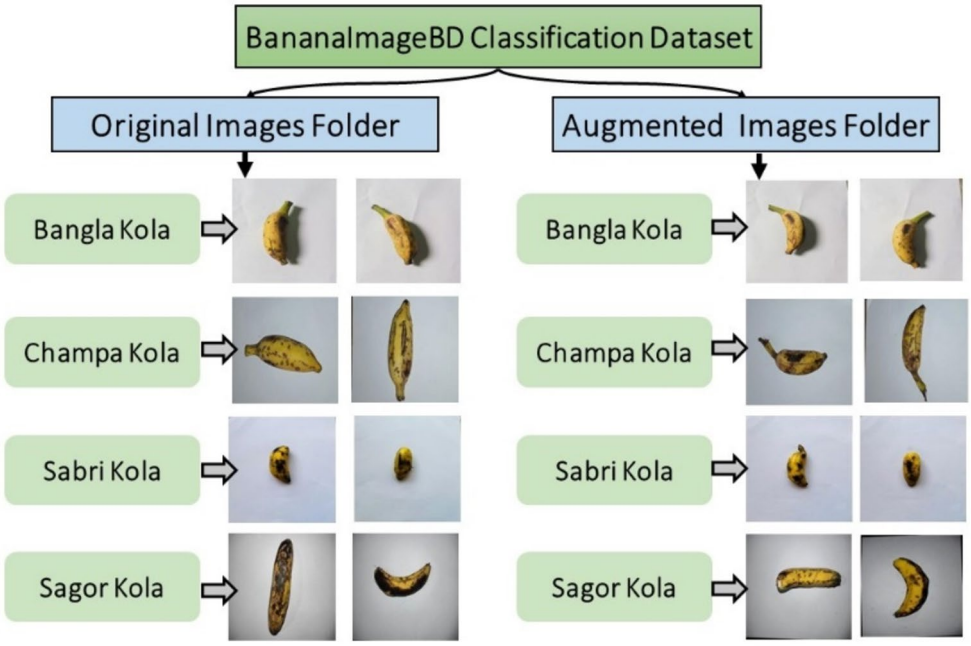

Fig. 1. Comprehensive visualization of BananaImageBD organization: folder structure and representative example images for each banana variety[31].

Class Name	BananaImageBD	
	Number of original images	Number of augmented images
Bangla Kola	444	1332
Champa Kola	994	2982
Sabri Kola	506	1527
Sagor Kola	524	1572
Total	2471	7413

Table 1. Distribution of images in the BananaImageBD by class.

The folder structure and example images for each class are illustrated in Fig. 1, while the number of images per class is detailed in Table 1. In this study, analyses were conducted using an augmented dataset.

BananaSet

Islam et al. developed the BananaSet dataset, a comprehensive resource specifically designed for classifying six widely consumed banana varieties in Bangladesh[32]. This dataset was meticulously curated to address the challenges associated with banana classification in agriculture, offering both raw and augmented high-resolution images for ML and DL applications. Bananas were sourced from diverse regions, including Faridpur and Mirpur-1, representing rural orchards and urban markets to ensure a rich and diverse sample. Six banana varieties—Shagor, Shabri, Champa, Anaji, Deshi, and Bichi—were photographed using a VIVO V25 5G smartphone under controlled conditions. A total of 1166 raw images were captured, each in JPG format at a resolution of 4608 $\times$ 3456 pixels, ensuring exceptional visual clarity.

Recognizing the need for larger datasets in DL, Islam et al. applied rigorous augmentation techniques, including scaling, rotation, shearing, zooming, and shifting[32]. This process increased the dataset to 6000 images, with 1000 augmented images per class. The augmented dataset preserved the resolution and quality of the original images, providing robust training data for ML models. The dataset was organized into six subfolders, with the dataset folder structure shown in Fig. 2, each corresponding to a specific banana variety to ensure clarity and usability. Additionally, the number of images for each class was detailed in Table 2. In this study, analyses were conducted using an augmented dataset.

Research methodology framework

In this research, a robust and comprehensive methodology was developed to classify images from the BananaImageBD and BananaSet datasets. The BananaImageBD dataset consists of four classes, while the BananaSet dataset contains six classes. The aim of the study was to accurately classify these banana varieties

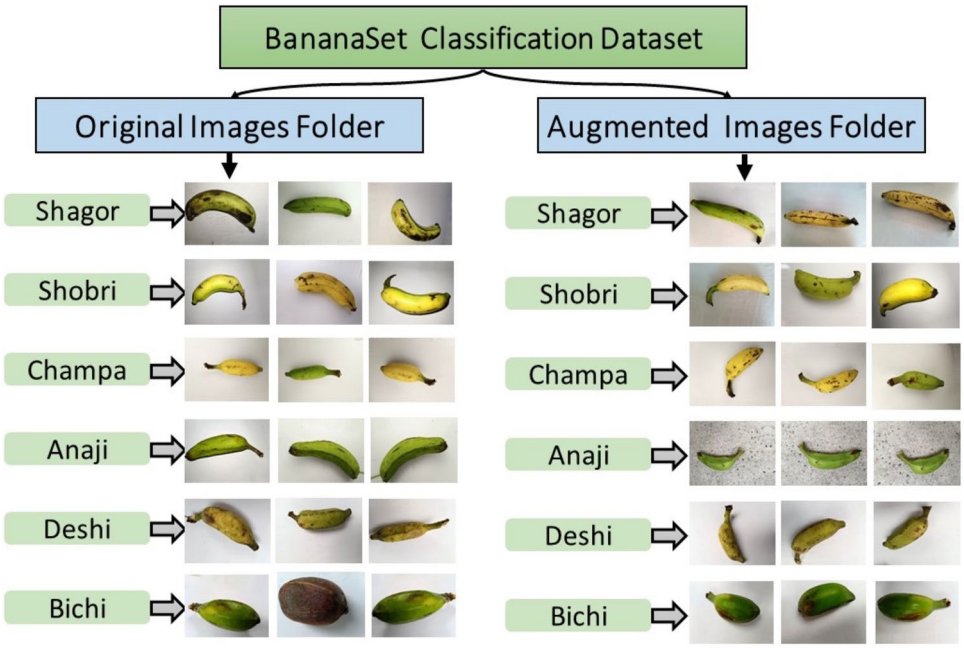

Fig. 2. A detailed overview of the BananaSet dataset's organization, showcasing the folder structure and sample images representing each banana variety[32].

Class Name	BananaSet	
	Number of original images	Number of augmented images
Shagor	239	1000
Shobri	163	1000
Champa	136	1000
Anaji	209	1000
Deshi	237	1000
Bichi	182	1000
Total	1166	6000

Table 2. Distribution of images in the BananaSet by class.

using advanced ML techniques. As shown in Fig. 3, the proposed methodology is detailed and involves several critical steps.

First, images and their corresponding labels were collected from each dataset. For each image, preprocessing steps were performed to transform the data into a format suitable for input into the ViT model. These preprocessing steps included resizing the images to 224 × 224 pixels, normalizing their color values using the mean and standard deviation of the dataset, and converting the images into tensors. The dataset was then split using 5-FCV, where separate training and test sets were created for each fold. For each fold, features were extracted using the ViT model, which allowed the conversion of image data into feature representations suitable for classification. After feature extraction, the extracted features were classified using SVM, GBM, MLP and CNN. The performance of the models was evaluated using various metrics including CAR, F1 score and RCL. In addition, confusion matrices were computed to further evaluate the classification performance across different classes. The methodology was implemented with great attention to detail using Python to ensure accurate calculations and seamless integration of the various components.

Vision transformer

In this study, a comprehensive feature extraction strategy was employed using the ViT, which has recently emerged as a groundbreaking architecture for image classification[33]. By extending the attention-based mechanisms that have proven successful in natural language processing, ViT was adapted to the visual domain. The study focused on applying ViT to classify BananaImageBD and BananaSet images, with particular emphasis on different varieties of bananas. This approach aimed to leverage the model's ability to capture complex relationships and dependencies within the images for more accurate classification. In the ViT model, as illustrated in Fig. 4 and further elaborated in Algorithm 1, the process of feature extraction and image classification was systematically

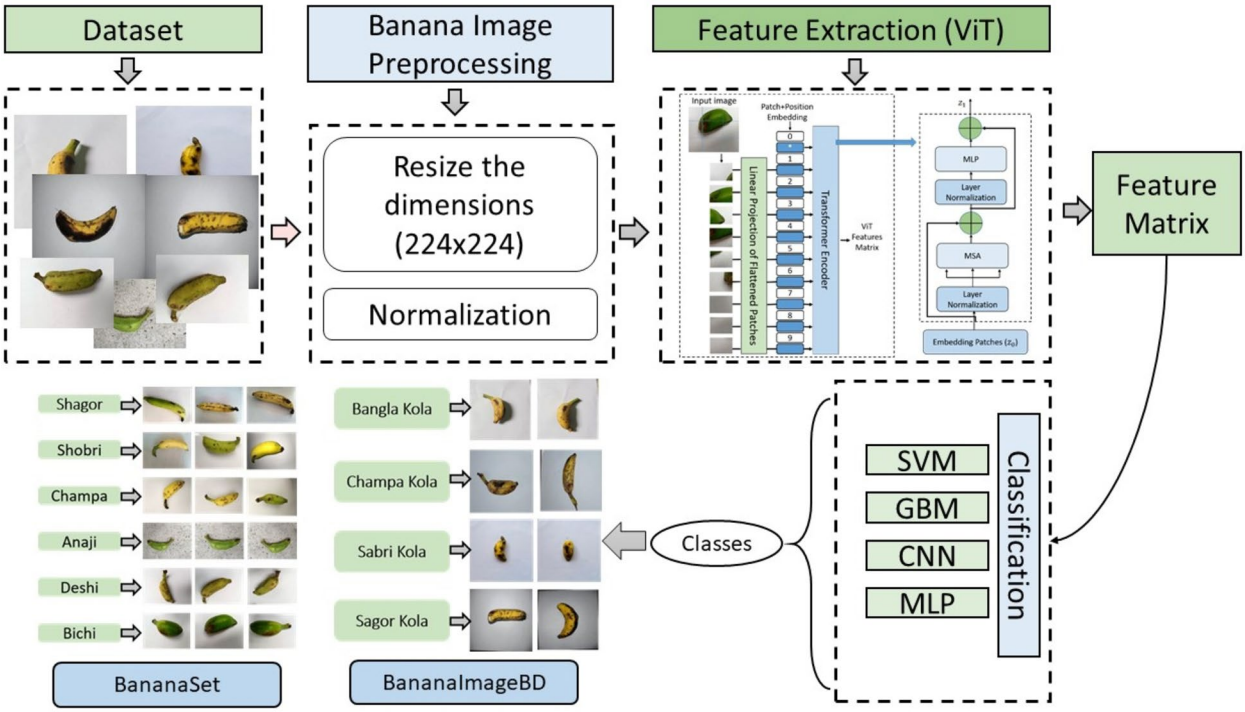

Fig. 3. The comprehensive workflow of the proposed methodology: preprocessing, feature extraction with ViT, and classification for banana image datasets.

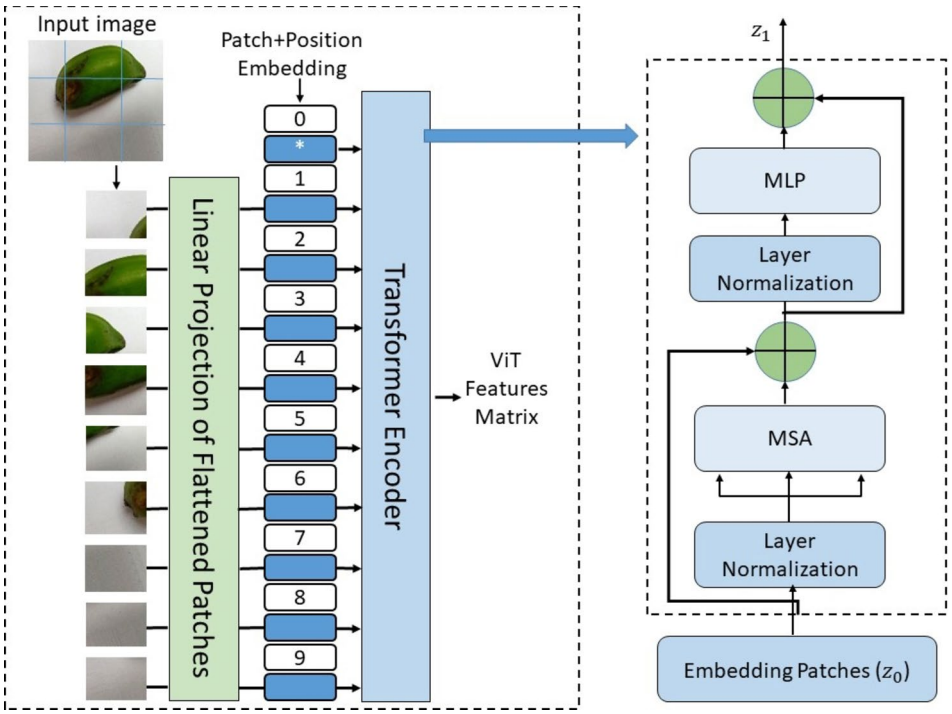

Fig. 4. Overview of the ViT architecture with the structure of the transformer encoder.

executed, providing a clear roadmap for how the model processes the input data and generates the necessary outputs.

The initial step in the ViT model involves converting the input image into a sequence of patches. Each image $x \in R^{Y \times P \times C}$ is divided into fixed-size non-overlapping patches[33]. These patches are then flattened to form

a 2D sequence $x_b \in R^{N \times B^2 C}$, where Y represents the height, P represents the width, C is the number of channels, and B is the resolution of each image patch. The number of patches N can be calculated using the Eq. (1)[34].

$$N = \frac{Y \times P}{B^2} \tag{1}$$

Once the patches are created, they are projected into a higher-dimensional space using a linear projection. The projection is performed by multiplying the patches by an embedding matrix $\mathrm{E} \in R^{B^2 C \times S}$, where S is the dimension of the embedding space. The resulting embedding $x_b \mathrm{E}$ represent the image as a sequence of patches in the higher-dimensional space[34]. Additionally, positional embedding $\mathrm{E}_{Pos} \in R^{(N+1) \times S}$ are added to each patch embedding to preserve spatial information within the image. The final sequence of patch embedding, z_0 , which includes a learnable class token x_{class}, is computed as Eq. (2). Where x_N represents the N-th image patch and E is the embedding matrix. The class token x_{class} is crucial for the classification task, as it aggregates information from all the patches and ultimately predicts the image class.

$$z_0 = [x_{class}; x_1 E; x_2 E; \ldots ; x_N E] + \mathrm{E}_{pos} \tag{2}$$

The next step in the ViT architecture is the Transformer Encoder. The encoder is composed of multiple identical encoder blocks, each containing two primary sub-layers: Multi-head self-attention (MSA) layer and the MLP layer. The process within each encoder block is as follows: Layer normalization, the input z^{l-1} to the l-th encoder block undergoes layer normalization to stabilize the training process and improve performance. This normalization operates across the feature dimension. MSA, after normalization, the sequence is passed through the MSA layer. The MSA layer computes the relationships between all patches using attention scores[34]. For each head in the MSA layer, the input sequence z is projected into three separate matrices: query q, key k, and value v. These projections are defined as Eq. (3)[34].

$$[q,\ k,\ v] = [zU_q,\ zU_k,\ zU_v] \tag{3}$$

where $U_q,\ U_k,\ U_v$ are learnable weight matrices of size $D \times D_h$, and D_h is the dimension of each attention head. In each attention head, the attention score is computed by calculating the dot product between the query and key matrices, followed by an softmax Eq. (4).

$$A = softmax\left(\frac{q \cdot k^T}{\sqrt{D_h}}\right) \tag{4}$$

where $A \in R^{N \times N}$ represents the attention matrix. In each attention head, the attention score is computed by taking the dot product between the query and key matrices, which measures the similarity between different input elements. To prevent excessively large values that could negatively impact the softmax function, the result is scaled by the square root of the head dimension, D_h. This normalization ensures numerical stability and improves training efficiency. The softmax function is then applied to transform the scaled scores into a probability distribution, where higher values indicate stronger attention weights. The resulting attention matrix, A, determines the contribution of each input element to the final representation. This matrix is subsequently multiplied by the value matrix, v, to generate the refined output for each attention head[35,36]. This output is concatenated across all heads and passed through a linear transformation with Eq. (5). Where U_{MSA} is a learnable weight matrix that projects the concatenated output into the original embedding space.

$$MSA\,(z) = [SA_1\,(z)\,; SA_2\,(z)\,; \ldots\ ; SA_K\,(z)]\,U_{MSA} \tag{5}$$

Feed-Forward MLP, the output of the MSA layer is then processed through a fully connected MLP, which consists of two layers with a non-linearity in between. The output of the MLP is added to the original input via a residual connection, facilitating the flow of gradients through the network. The process for each encoder block can be summarized as Eqs. (6) and (7)[35,36].

$$z_0^l = MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{6}$$

$$z^l = MLP\left(LN\left(z_0^l\right)\right) + z_0^l \tag{7}$$

Furthermore, the image size was set to 224 × 224 pixels, which is a standard input dimension that allows the model to effectively capture essential visual features. Each image was divided into patches with a patch size of 16×16 pixels, ensuring that the spatial structure of the image was preserved while reducing computational complexity. The model used 12 encoder layers, each containing 12 attention heads, to allow rich representation learning and capture complex relationships between image patches. These design choices were carefully implemented to maximize the model's ability to accurately classify the BananaImageBD and BananaSet datasets.

1: Input: $\boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{Y}\times\boldsymbol{P}\times\boldsymbol{C}}$: Input image tensor where $\boldsymbol{Y}, \boldsymbol{P}$, and $\boldsymbol{C}$ represent height, width, and number of channels, respectively
2: Compute N, the number of patches, using $\boldsymbol{N} = \frac{\boldsymbol{Y}\times\boldsymbol{P}}{\boldsymbol{B}^2}$: $\boldsymbol{B}$ is the patch size
3: Encode patches using PatchEncoder
4: **for** i ← 1 to N_transformer_layers do
5: t1 ← Apply layer normalization on encoded patches
6: t2 ← Compute multi-head attention on t1 with the specified number of attention heads and projection dimensions
7: t3 ← Add x1 and attention outputs: Skip connection 1
8: t4 ← Apply layer normalization on t3
9: t5 ← Pass t4 through a MLP with predefined units and dropout rate
10: t6 ← Add t3 and MLP outputs: Skip connection 2
11: Update encoded patches with t6
12: **end for**
13: Perform final layer normalization on encoded patches
14: Flatten the representation
15: Apply dropout regularization to the flattened features
17: Classify the features using SVM, GBM, MLP and CNN models
18: Output: Class predictions
19: **end procedure**

Algorithm 1. Vision transformer.

Classification stage and evaluation metrics

In this study, SVM, GBM, MLP and CNN classifiers were used to categorize the features extracted from the ViT model. Each classifier brought different advantages to the classification process, ensuring a robust evaluation of the extracted features. The classifiers are described in detail below, followed by a description of the evaluation metrics used to assess their performance.

The SVM is a powerful supervised learning algorithm that is widely used for classification tasks[37]. Its core principle is to find an optimal hyperplane that maximally separates data points of different classes in a high-dimensional feature space. For non-linearly separable data, SVM uses kernel functions, such as the radial basis function, to map input features to a higher-dimensional space where linear separation becomes possible. Mathematically, the SVM decision function can be expressed as Eq. (8)[37] where α_i is the Lagrange multipliers, y_i represents the class labels, $K(x_i, x)$ denotes the kernel function, and b is the bias term. The sign function in this equation plays a crucial role in determining the final class label by mapping the computed decision value to either a positive or negative class. Essentially, it ensures that the output of the decision function corresponds to one of the predefined categories. SVM's ability to handle high-dimensional data with excellent generalization makes it a suitable choice for the ViT-based feature representation.

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \tag{8}$$

GBM, on the other hand, is an ensemble learning technique that builds a strong predictive model by sequentially combining weak learners, typically decision trees[38]. At each iteration, GBM minimizes a differentiable loss function by adding a new tree that corrects the errors of the previous models. This iterative boosting process enhances classification performance by reducing bias and variance. The final model is represented as Eq. (9)[38]. Where h_m represents the individual decision trees, γ_m denotes their corresponding weights, and M is the total number of trees. The adaptability of GBM to complex data distributions makes it a highly effective classifier for the extracted features[38].

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{9}$$

CNN, a DL model renowned for its ability to capture spatial and hierarchical patterns in data, was also employed as a classifier. Unlike traditional methods, CNN leverages convolutional layers to automatically learn spatial features, followed by pooling layers for dimensionality reduction[39]. The learned features are passed through fully connected layers for final classification. CNN's architecture, when used as a classifier, benefits from its capability to learn complex patterns in high-dimensional feature representations, making it complementary to the ViT-extracted features.

The performance of each classifier was evaluated using standard metrics, including CAR, F1 score, and RCL. CAR measured the proportion of correctly classified instances relative to the total number of samples, while the F1 score, defined as the harmonic mean of precision (PRC) and RCL, provided a balanced evaluation of the model's performance, particularly for imbalanced datasets. RCL assessed the ability of the model to correctly identify all relevant instances[35,40]. These metrics were calculated separately for the 4-class and 6-class problems to ensure a comprehensive assessment of classification performance. Additionally, confusion matrices were constructed to analyze misclassification rates across different classes, offering further insights into the effectiveness of each classifier. CAR is the ratio of correctly classified instances to the total number of instances (TNI) as given Eq. (10)[41,42]. The F1 score is the harmonic mean of PRC and RCL, defined as in Eq. (11)[35,40].

$$CAR = \frac{Number\ of\ Correct\ Predictions\ Image}{TNI} \tag{10}$$

$$F1\ Score = 2 \cdot \frac{PRC \cdot RCL}{PRC + RCL} \tag{11}$$

PRC and RCL are critical for assessing the effectiveness of a model in handling imbalanced datasets or multi-class problems[40]. Precision measures the proportion of correctly predicted positive instances out of all predicted positive instances, emphasizing the model's ability to avoid false positives. This is mathematically represented as Eq. (12)[35,40].

$$PRC = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{12}$$

RCL, on the other hand, evaluates the model's capacity to identify all relevant positive instances, focusing on minimizing false negatives[40]. It ensures that the model captures as many true positive instances as possible, which is particularly important in applications where missing a positive class instance could have significant consequences. RCL is mathematically expressed as Eq. (13)[35].

$$RCL = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{13}$$

Results

In this study, the proposed methodology demonstrated the effectiveness of ViT-based feature extraction in image processing tasks, particularly in the classification of the four-class BananaImageBD and six-class BananaSet datasets. Although the ViT model uses an MLP-based classifier within its architecture, this study introduced a novel approach by integrating a SVM for classification, and achieved promising results. First, all images in the datasets were resized to 224 × 224 pixels to match the input requirements of the ViT model. The images were then converted into tensors and normalized based on the mean and standard deviation values used in the pre-training of the ViT model. This normalization ensured compatibility and optimal performance of the model during feature extraction. The ViT model was then applied using a 16 × 16 image patching mechanism. Each input image (224 × 224) was divided into non-overlapping patches of 16 × 16 pixels, resulting in 14 × 14=196 patches per image. These patches were individually transformed into feature vectors suitable for model processing. The embeddings represented high-dimensional feature information that was extracted from the output layer of the ViT model as a feature matrix for subsequent classification tasks. The extracted features were classified using four different classifiers: SVM, GBM, MLP and CNN. The performance evaluation was performed using the CAR, F1 score and RCL metrics. For both the BananaImageBD and BananaSet datasets, a 5-FCV strategy was applied to ensure a robust performance evaluation. The results of the F1, RCL, and CAR metrics for each dataset and for SVM, GBM, MLP, and CNN are presented in Fig. 5.

As shown in Fig. 5 for the BananaImageBD dataset, the SVM classifier consistently outperformed its counterparts across all metrics. In particular, it achieved the highest CAR across all folds, exceeding 99.30% in every instance and peaking at 99.93% in Fold 3. MLP also exhibited strong classification performance, with CAR values consistently above 97.50%, closely approaching SVM in multiple folds. GBM maintained competitive results but showed slightly lower CAR, fluctuating between 96.42% and 97.98%. CNN, on the other hand, displayed the lowest CAR, particularly in Folds 1 and 4, where its CAR fell below 95.00%. These trends were further reflected in the F1 score results. SVM recorded the highest F1 score across all folds, consistently nearing 1.0000, highlighting its robust balance between precision and RCL. MLP followed closely, achieving nearly identical performance in several folds. GBM displayed moderate results, maintaining F1 scores in the 0.9600–0.9700 range, while CNN lagged behind, with the lowest F1 scores observed in Folds 1 and 4. A similar pattern emerged in the RCL, where SVM achieved nearly perfect scores across all folds, reaffirming its reliability in correctly identifying positive instances. MLP closely mirrored this performance, particularly in Fold 4, where it attained 0.9991. GBM exhibited stable RCL, though slightly lower than SVM and MLP, while CNN remained the weakest performer, with RCL ranging between 0.9300 and 0.9700.

A similar pattern was observed for the BananaSet dataset. The SVM classifier consistently achieved the highest CAR across all evaluation metrics. It maintained a CAR above 99.70% in every fold, reaching a perfect score of 100.00% in Fold 3. MLP followed closely SVM, demonstrating nearly equivalent performance, with CAR values consistently exceeding 99.50%. GBM exhibited stable results, maintaining CAR values between 97.50% and 98.60%, while CNN lagged behind, recording the lowest CAR, particularly in Folds 1, 2, and 5, where it remained around 92.50%. The F1 score results reinforced these trends. SVM achieved near-perfect

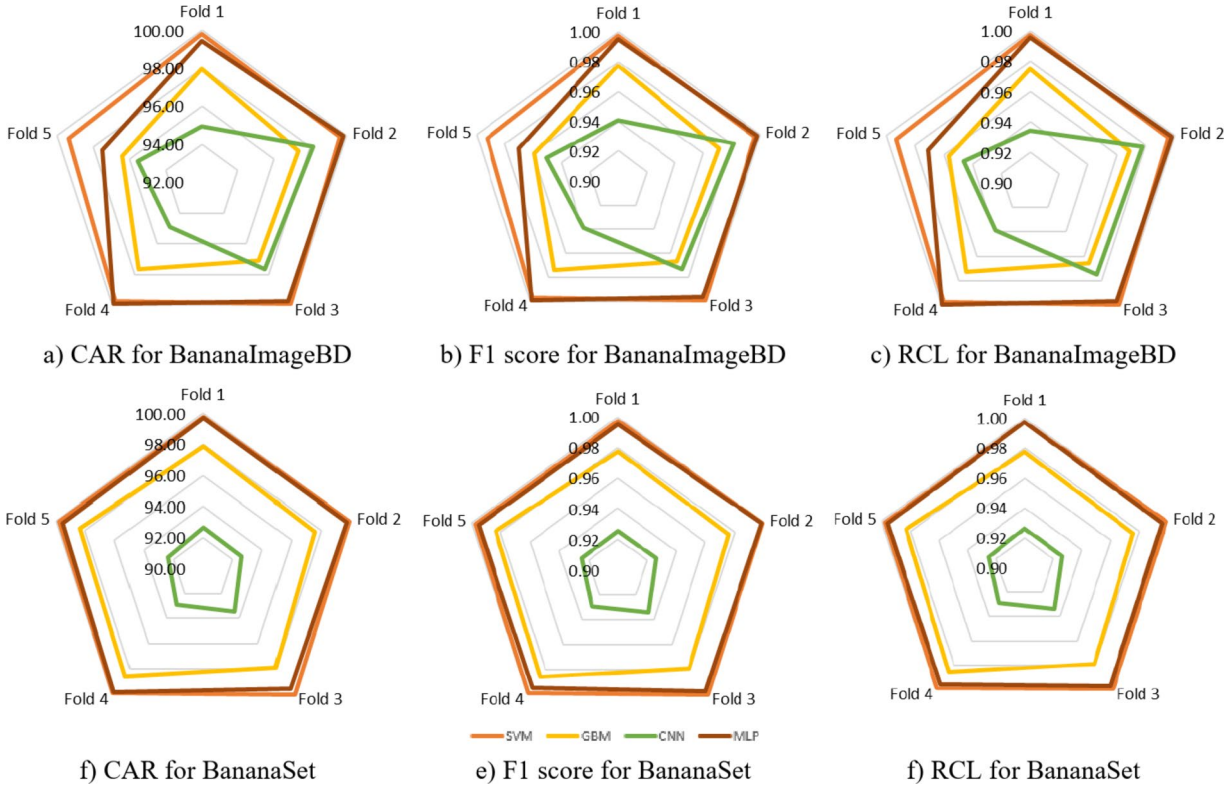

Fig. 5. Results of the 5-FCV, detailing the F1 score, RCL, and CAR metrics for each fold in the BananaImageBD and BananaSet datasets.

F1 scores in every fold, peaking at 1.0000 in Fold 3, further highlighting its effectiveness. MLP also delivered strong performance, with F1 scores consistently above 0.9900. GBM maintained respectable results, with values ranging from 0.9757 to 0.9859, while CNN struggled to achieve competitive scores, remaining below 0.9350 in all folds. Similarly, the RCL values demonstrated that SVM was the most reliable classifier, consistently scoring above 0.9970 and reaching 1.0000 in Fold 3. MLP closely mirrored this trend, achieving nearly identical RCL, particularly in Fold 4, where it surpassed 0.9950. GBM exhibited stable yet slightly lower RCL, while CNN recorded the lowest RCL across all folds, with the weakest performance observed in Folds 1, 2, and 5.

Figure 6 presenst radar charts illustrating the class-specific metrics, including CAR, F1 score, and RCL, for SVM, CNN and GBM classifier and each dataset. These charts enable a detailed analysis of the performance for individual classes within the BananaImageBD and BananaSet datasets. Notably, CAR was emphasized in this analysis as it provides a clear measure of the overall CAR of each classifier in distinguishing between classes. For the BananaImageBD dataset, the SVM exhibited the most consistent performance across all classes. Specifically, Champa Kola achieved the highest CAR of 99.87%, closely followed by Sabri Kola with 99.67%. These results are corroborated by their corresponding F1 Scores and RCL, all of which exceeded 0.9900, highlighting the robust ability of SVM to accurately identify these classes. Conversely, Sagor Kola recorded a slightly lower CAR of 99.48%, yet its performance remains commendable and well above the thresholds of other classifiers. The GBM also delivered strong results, with Champa Kola achieving a CAR of 99.26%, while other classes, such as Bangla Kola and Sagor Kola, recorded CAR values of 94.84% and 94.71%, respectively. These outcomes suggest that while GBM is effective, it is more sensitive to class variations than SVM. In contrast, CNN demonstrated a more variable performance. For instance, Champa Kola achieved a competitive CAR of 99.39%, indicating CNN's capability in certain classes. However, the CAR for Bangla Kola and Sagor Kola was significantly lower, at 91.93% and 92.79%, respectively. This variability suggests that CNN may require further optimization to handle intra-class differences effectively.

The performance trends observed in BananaSet further validate the superiority of SVM in class-level accuracy. Notably, Shobri achieved a perfect CAR of 100.00%, accompanied by F1 Score and RCL of 1, reflecting flawless classification. Similarly, other classes, such as Anaji and Shagor, exhibited near-perfect CARs of 99.91% and 99.90%, respectively, affirming SVM's reliability across all classes. GBM also performed well in this dataset, with its highest CAR observed for Shobri at 98.78%, closely followed by Bichi and Champa, with CARs of 98.10% and 98.32%, respectively. Despite this strong performance, certain classes, such as Shagor with CAR of 97.49%, displayed minor reductions in CAR compared to SVM. CNN, however, showed pronounced variability in the BananaSet dataset. While Bichi achieved a CAR of 98.70%, other classes, such as Shagor with 91.10% and Shobri with 87.10%, were significantly lower, highlighting a potential limitation of CNN in maintaining uniformity across classes. The radar charts reveal that SVM consistently provides the highest and most uniform CAR values

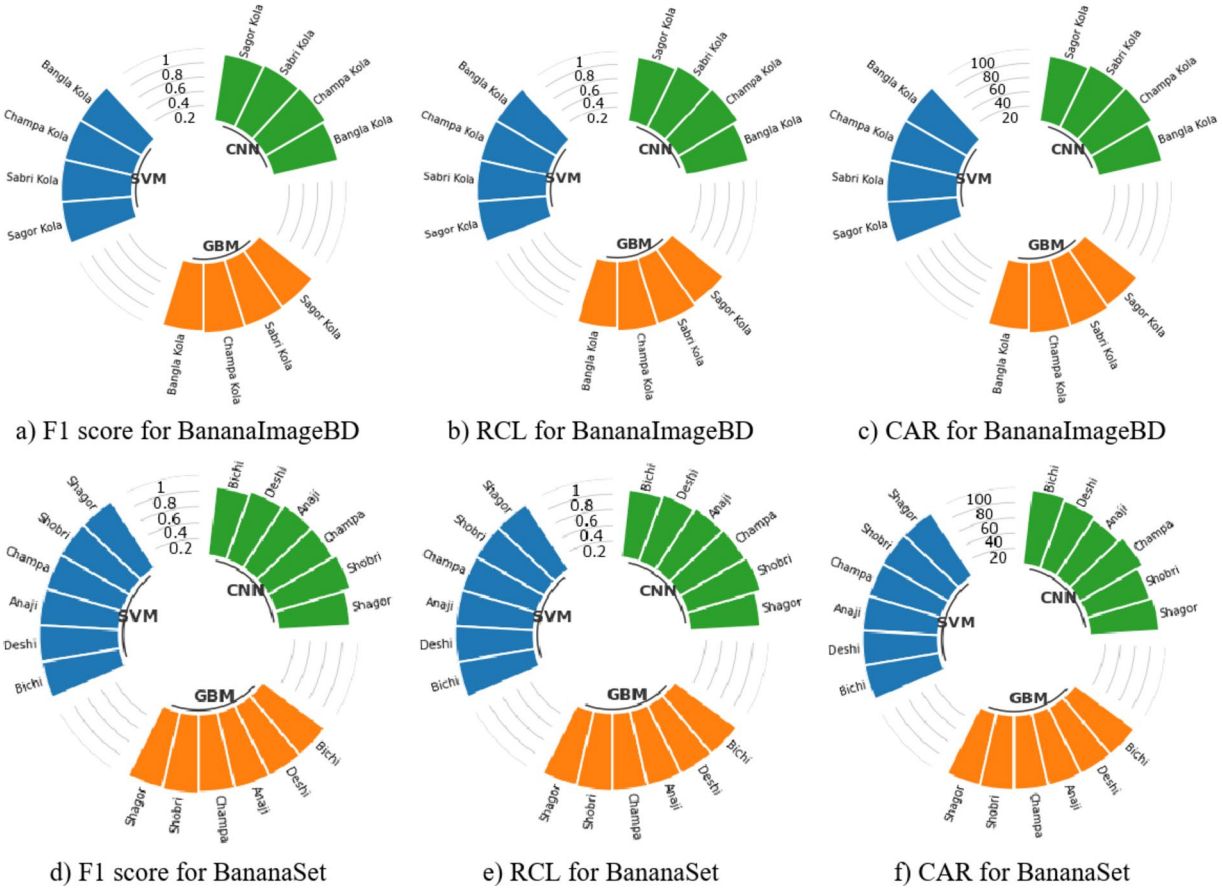

Fig. 6. Radar charts illustrating the class-specific metrics, including CAR, F1 score, and RCL, for each dataset and classifier.

Feature	Classification Method	BananaImageBD			BananaSet		
		CAR (%)	F1 score	RCL	CAR (%)	F1 score	RCL
ViT	SVM	99.70	0.9965	0.9965	99.86	0.9986	0.9986
	GBM	97.39	0.9696	0.9676	98.05	0.9803	0.9802
	CNN	96.23	0.9569	0.9543	92.80	0.9279	0.9281
	MLP	99.29	0.9917	0.9920	99.65	0.9962	0.9963
ResNet	SVM	96.69	0.9636	0.9639	98.95	0.9894	0.9896
	GBM	96.29	0.9572	0.9553	96.76	0.9670	0.9670
	CNN	97.27	0.9689	0.9720	94.66	0.9466	0.9465
	MLP	97.76	0.9756	0.9751	98.93	0.9892	0.9893
SqueezeNet	SVM	97.63	0.9746	0.9734	98.45	0.9844	0.9845
	GBM	93.08	0.9256	0.9188	88.60	0.8869	0.8858
	CNN	88.51	0.8651	0.8693	68.62	0.6707	0.6832
	MLP	95.88	0.9559	0.9541	91.93	0.9195	0.9192

Table 3. Average metrics calculated across the 5-FCV for each dataset for vit. ResNet and squeezenet features.

across all classes and datasets. This suggests that SVM is the most robust classifier in handling both inter-class and intra-class variations, making it a highly reliable choice for these datasets.

This study assessed the effectiveness of the proposed methodology by computing the average performance metrics across a 5-FCV framework, as presented in Table 3. Additionally, classification results were obtained using ViT, ResNet, and SqueezeNet features to facilitate a comprehensive performance comparison. As shown in Table 3, the SVM classifier consistently outperformed other models across all feature sets. When employing ViT features, SVM achieved the highest performance, yielding 99.70% CAR, 0.9965 F1 score, and 0.9965 RCL for the BananaImageBD dataset, while attaining 99.86% CAR, 0.9986 F1 score, and 0.9986 RCL for the BananaSet

dataset. MLP also demonstrated competitive accuracy, with CAR values of 99.29% and 99.65%, respectively. In contrast, GBM and CNN exhibited lower CAR, particularly CNN, which recorded the weakest performance with a 92.80% CAR on the BananaSet dataset. When using ResNet features, SVM remained the top-performing classifier, achieving 98.95% CAR and 0.9894 F1 score on the BananaSet dataset. MLP also yielded competitive results, whereas GBM and CNN performed comparatively worse. Specifically, CNN achieved a 94.66% CAR, which was notably lower than that obtained using ViT-based feature extraction. For SqueezeNet features, SVM once again demonstrated superior classification performance, achieving a 98.45% CAR on the BananaSet dataset, alongside stable F1 score and RCL values. However, CNN exhibited a significant performance drop, obtaining only a 68.62% CAR, indicating that this combination was less effective for accurate classification.

All in all, SVM consistently produced the highest CAR across all feature sets, with ViT features producing the best results. MLP also proved to be a strong alternative, especially when combined with ViT and ResNet features. In contrast, CNN showed lower CAR, especially with SqueezeNet features, suggesting that it may be less suitable for the given classification task.

Furthermore, as shown in Table 3, the results clearly demonstrate that SVM consistently outperformed the other classifiers across all metrics, Establishing it as the most effective method for both datasets. The highest CAR was achieved using vit.based features with SVM, reinforcing the robustness of this combination. To further illustrate the distribution of vit.derived features, a t-SNE (t-Distributed stochastic neighbor Embedding) visualization was generated and is presented in fig. 7. t-SNE is a powerful dimensionality reduction technique that projects high-dimensional feature representations into a two-dimensional space while preserving local similarities. This method is particularly effective for visualizing how well different feature representations separate distinct classes, providing valuable insights into the discriminative capability of the extracted features. The t-SNE visualization in fig. 7 depicts the feature distributions on a class-wise basis, offering a clearer Understanding of the feature separability achieved by vit.

Beyond feature visualization, a comprehensive evaluation of the model's performance was carried out through confusion matrix analysis for each fold of the 5-FCV process, along with the computation of the mean confusion matrix. This approach facilitated a detailed fold-by-fold assessment, ensuring a robust evaluation of classification consistency. Figures 8 and 9 illustrate the confusion matrices computed for individual folds and the aggregated confusion matrices across the entire 5-FCV process for the BananaImageBD and BananaSet datasets, respectively. These matrices provided deeper insights into the model's predictive performance and further validated the effectiveness of SVM with ViT-based features in achieving superior CAR.

The mean confusion matrix for the BananaImageBD dataset as seen in Fig. 8f confirmed the strong discriminative power of the ViT-based features when processed through the SVM classifier. The results demonstrated consistently high CAR across all banana classes, with exceptionally low misclassification rates. Notably, the Champa Kola class achieved the highest correct classification count of 595.60, while misclassification values remained negligible, with a maximum error of 0.80. The balanced distribution of correct classifications across all classes underscored the model's ability to generalize effectively. Even for visually similar banana varieties, the proposed approach exhibited a strong capacity for differentiation, further reinforcing its robustness. The high consistency in classification performance strongly indicated that the integration of ViT features with SVM effectively enhanced predictive reliability. Similarly, the mean confusion matrix for the BananaSet dataset as seen in Fig. 9f provided additional validation of the model's efficacy. The CAR remained exceptionally high, with minimal misclassification observed across all banana varieties. The Anaj, Bichi, Champa, Deshi, Shagor, and Shobri classes all exhibited near-perfect classification scores, with their correct classification counts recorded as 199.80, 200.00, 199.60, 199.80, 199.60, and 199.60, respectively. These figures indicated that the model effectively captured the distinct characteristics of each banana type, leading to highly precise predictions. Furthermore, the misclassification values were minimal, with a maximum error of 0.20, occurring in only a few instances. Minor misclassifications were noted in the Anaj and Deshi classes, yet these errors were insignificant compared to the overall accuracy. The balanced classification performance across all categories suggested that the ViT-

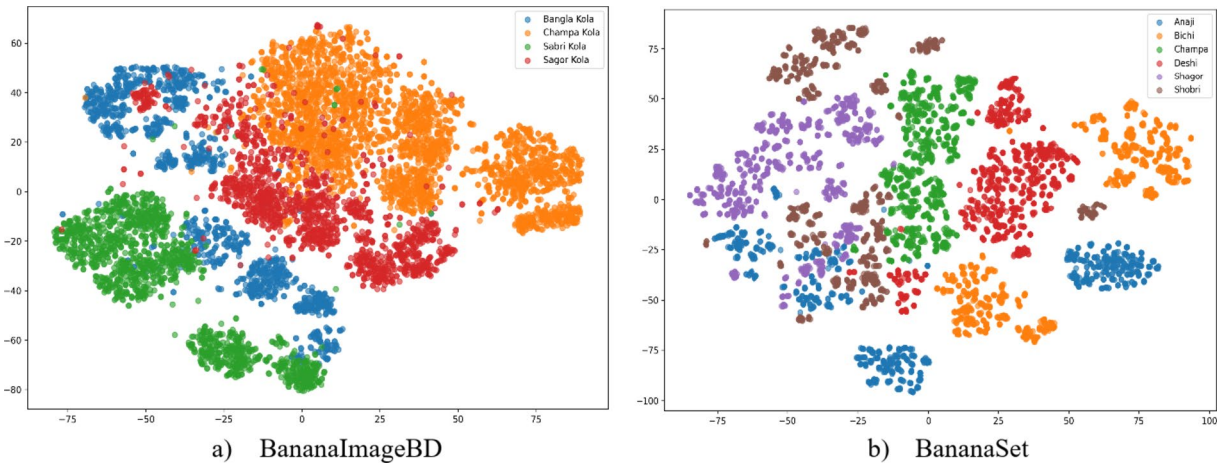

Fig. 7. Detailed t-SNE feature map of ViT extracted features for BananaImageBD and BananaSet datasets.

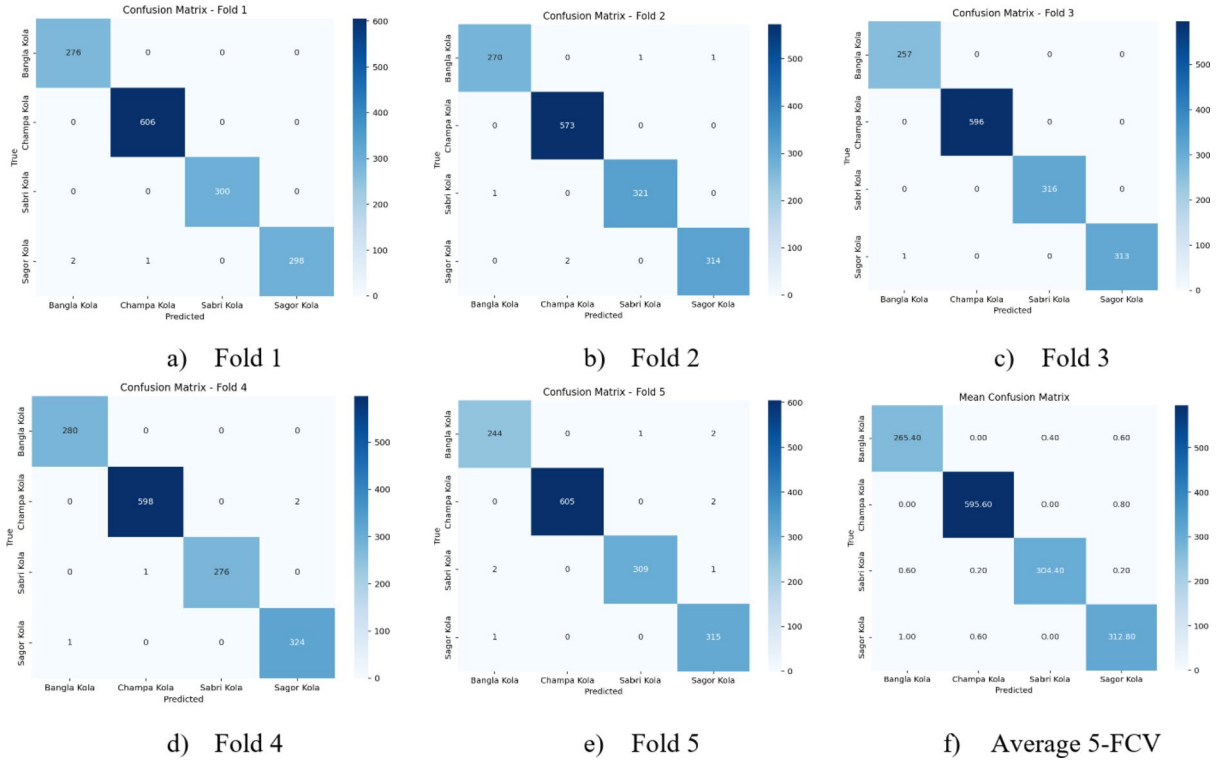

Fig. 8. Detailed confusion matrices for each fold of the 5-FCV process on the BananaImageBD for SVM

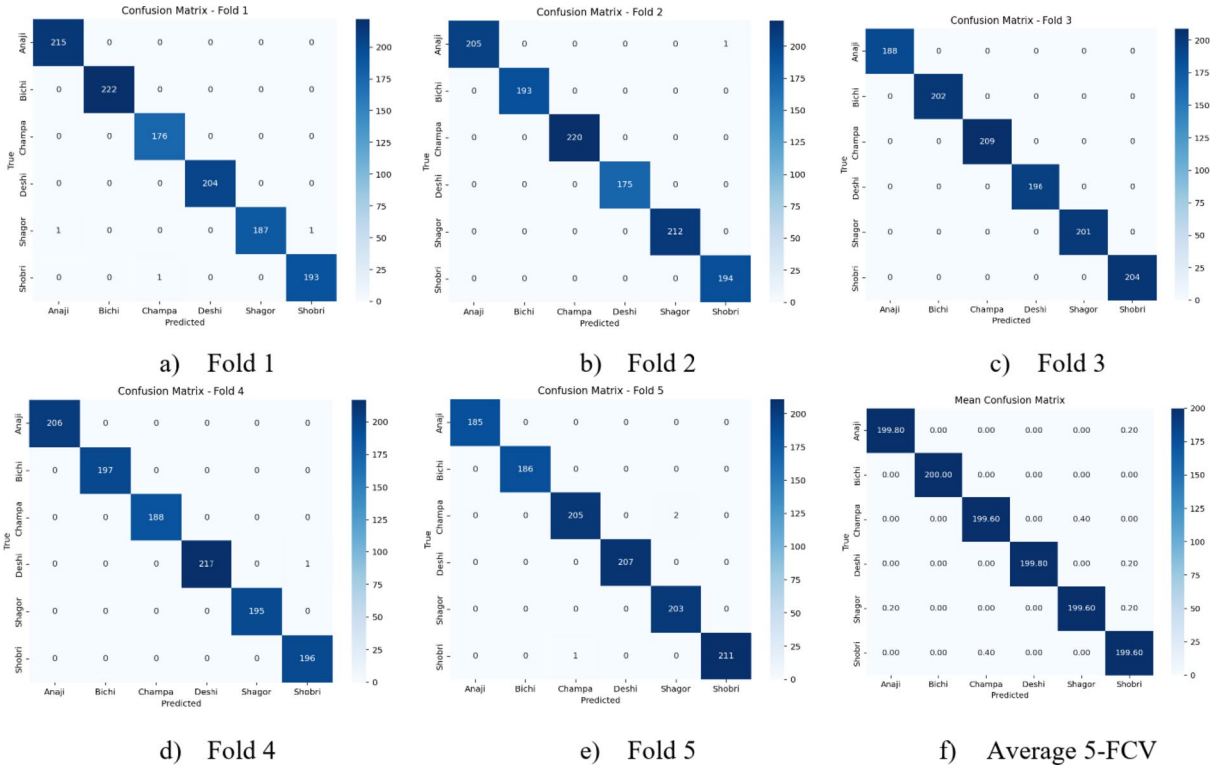

Fig. 9. Detailed confusion matrices for each fold of the 5-FCV process on the BananaSet for SVM.

based features successfully encoded essential visual cues, minimizing ambiguity between different banana varieties. These findings strongly reinforced the reliability and robustness of the proposed ViT-SVM framework, demonstrating its potential for highly accurate and consistent classification performance across diverse datasets.

An ANOVA test was carried out to evaluate whether significant differences existed in the performance of classifiers (SVM, GBM, CNN, MLP) across three evaluation metrics—CAR, F1 score, and RCL—on the two datasets, BananaImageBD and BananaSet. The results of the ANOVA test indicated statistically significant differences among the classifiers across all three metrics in both datasets. For the BananaImageBD dataset, the p-values obtained for CAR, F1 score, and RCL were 0.0275, 0.0121, and 0.0089, respectively. Likewise, for the BananaSet dataset, the corresponding p-values were 0.0293, 0.0108, and 0.0076. These findings strongly suggest that the performance differences among the classifiers are statistically significant for both datasets and across all evaluation metrics. To gain a deeper understanding of these differences, a post-hoc Tukey's Honest Significant Difference (HSD) test was conducted. The analysis of the Tukey's HSD test results revealed key insights into pairwise classifier comparisons. For instance, the mean difference in CAR between SVM and GBM on the BananaImageBD dataset was 2.31% ($p = 0.0012$, confidence interval: [1.12, 3.50]). Furthermore, the difference in F1 score between SVM and CNN for the BananaSet dataset was 0.0707 ($p < 0.0001$, confidence interval: [0.052, 0.089]). These results underscore the superior performance of SVM compared to other classifiers. Additionally, CNN, particularly on the BananaSet dataset, demonstrated a significant decline in performance, highlighting its relative inefficiency in comparison to SVM, GBM, and MLP. These findings not only provide a clearer view of the performance dynamics between the classifiers but also reaffirm the robustness of SVM in delivering consistently superior results across both datasets.

To enhance the geographical generalization capability of the study, additional work was conducted using an alternative dataset. Due to the limited availability of suitable datasets for banana variety classification, a new dataset, focused on banana ripening stages, was sourced from Indonesia[43]. This dataset encompasses two banana varieties, Cavendish and Ambon, and includes images of bananas at four ripening stages: raw, half-ripe, ripe, and overripe. In total, the dataset consists of 1865 images, each with a resolution of 224 $\times$ 224 pixels. To assess the model's generalization performance, 5-FCV was applied, and classification was carried out using ViT-based SVM model. The results revealed a CAR of 98.61%, an F1 score of 0.9854, and an RCL of 0.9858. These results surpass the findings of Pangestu et al.[44], where a ViT model achieved a CAR of 91.61%. This significant enhancement in performance highlights the effectiveness of the proposed model in classification tasks. The results underscore the robustness of the proposed model, demonstrating that it not only performs exceptionally well on geographically specific data but also excels across various banana varieties with different ripening stages.

Conclusion

In conclusion, this study presents an innovative approach to banana classification, addressing significant challenges in agricultural image analysis. By combining the advanced feature extraction capabilities of the ViT model with the robust classification power of SVM, this framework achieves outstanding results across multiple datasets, including the four-class BananaImageBD and the six-class BananaSet datasets. This represents a notable advancement in DL for agricultural diagnostics, providing an efficient and scalable solution for the accurate identification of banana varieties.

The practical implications of this work are substantial. Achieving a CAR of 99.86% for the BananaSet and 99.70% for the BananaImageBD represents a significant improvement over traditional methods, surpassing previous benchmarks by a margin of 1.77%[45]. This improvement is not merely theoretical but translated into tangible real-world outcomes. This level of accuracy ensures that the proposed model can be deployed effectively in agricultural settings, offering farmers the ability to quickly and accurately identify banana varieties at various stages of ripening. For instance, a CAR of 99.86% could reduce errors in classification, thereby minimizing the risk of misidentifying crop varieties, which in turn could lead to more targeted interventions and better resource management. Furthermore, such precision could aid in predicting harvest times more accurately, ensuring optimal yield and reducing crop loss, ultimately enhancing farm productivity.

Additionally, this study has highlighted the SVM classifier's ability to generalize well across both inter- and intra-class variations, emphasizing its potential for real-world deployment. The robustness of the framework is reinforced by the consistent performance of the ViT-based SVM model, which can handle complex agricultural tasks with high reliability. The exceptional results obtained through 5-FCV not only provide confidence in the model's effectiveness but also demonstrate its ability to overcome potential challenges such as data imbalance and variability in agricultural data. To further emphasize the practical value, the proposed system could be integrated into precision farming applications, where it could monitor and manage crop health and development over time. This integration would not only improve accuracy but also streamline processes such as disease detection and crop quality monitoring, enabling early intervention. For example, by accurately identifying disease outbreaks in bananas with an impressive 99.86% CAR, farmers can take timely action to prevent the spread of infection, thereby reducing potential losses and increasing crop yield.

Moreover, the scalability of the system provides a significant advantage for broader agricultural contexts. While this study has focused on banana classification, the flexibility of the ViT and SVM model allows for easy adaptation to other agricultural products. This suggests that the proposed framework could be extended to various crops, contributing to more efficient and automated agricultural systems worldwide. As the technology evolves, future research can further enhance this system, exploring new applications in automated systems and precision agriculture. In essence, this study represents a substantial step forward in the development of intelligent agricultural solutions, paving the way for future innovations in the application of AI to agriculture. Through these results, the study not only contributes to the advancement of banana classification techniques but also sets a foundation for improving real-world agricultural practices, making a significant impact on the efficiency and sustainability of farming operations.

Data availability

The experimental datasets are publicly available datasets, https://data.mendeley.com/datasets/ptfscwtnyz/2 for BananaImageBD[32], https://data.mendeley.com/datasets/35gb4v72dr for BananaSet[33] and https://dataverse.telkomuniversity.ac.id/dataset.xhtml?persistentId=doi:10.34820/FK2/GJBZ0X for Banana Ripeness[44].

Received: 23 January 2025; Accepted: 21 March 2025
Published online: 26 March 2025

References

1. Ibba, P. et al. Supervised binary classification methods for strawberry ripeness discrimination from bioimpedance data. *Sci. Rep.* **11**, 11202. https://doi.org/10.1038/s41598-025-86315-1 (2021).
2. Ukwuoma, C. C. et al. Recent advancements in fruit detection and classification using deep learning techniques. Math. Probl. Eng. 9210947. (2022). https://doi.org/10.1155/2022/9210947 (2022).
3. Sabouri, A. et al. Machine learning techniques for non-destructive Estimation of Plum fruit weight. *Sci. Rep.* **15**, 751. https://doi.org/10.1038/s41598-024-85051-2 (2025).
4. Rizzo, M. et al. Fruit ripeness classification: A survey. *Artif. Intell. Agric.* **7**, 44–57. https://doi.org/10.1016/j.aiia.2023.02.004 (2023).
5. Ergün, E. Deep learning-based multiclass classification for citrus anomaly detection in agriculture. *Signal. Image Video Process.* **18**, 8077–8088. https://doi.org/10.1007/s11760-024-03452-2 (2024).
6. Shahi, T. B. et al. Fruit classification using attention-based MobileNetV2 for industrial applications. *PLoS ONE.* **17**, e0264586. https://doi.org/10.1371/journal.pone.0264586 (2022).
7. Ghazal, S. et al. Analysis of visual features and classifiers for fruit classification problem. *Comput. Electron. Agric.* **187**, 106267. https://doi.org/10.1016/j.compag.2021.106267 (2021).
8. Albattah, W. et al. A novel deep learning method for detection and classification of plant diseases. *Complex. Intell. Syst.* https://doi.org/10.1007/s40747-021-00536-1 (2022).
9. Kunduracıoğlu, İ. & Paçal, İ. Deep learning-based disease detection in sugarcane leaves: evaluating EfficientNet models. *J. Oper. Intell.* **2** (1), 321–235. https://doi.org/10.31181/jopi21202423 (2024).
10. Haridasan, A. et al. Deep learning system for paddy plant disease detection and classification. *Environ. Monit. Assess.* **195**, 120. https://doi.org/10.1007/s10661-022-10656-x (2023).
11. Kunduracioglu, I. CNN models approaches for robust classification of Apple diseases. *Comput. Decis. Making: Int. J.* **1**, 235–251. https://doi.org/10.59543/comdem.v1i.10957 (2024).
12. Kunduracioglu, I. Utilizing ResNet architectures for identification of tomato diseases. *J. Intell. Decis. Mak. Inform. Sci.* **1**, 104–119. https://doi.org/10.59543/jidmis.v1i.11949 (2024).
13. Borra, S. R. et al. Fruit type classification using stacked bi-directional long short-term memory. In 2023 Int. Conf. Evol. Algorithms Soft Comput. Techn. (EASCT), IEEE, 1–4 (2023).
14. Gill, H. S. et al. Fruit type classification using deep learning and feature fusion. *Comput. Electron. Agric.* **211**, 107990. https://doi.org/10.1016/j.compag.2023.107990 (2023).
15. Ratha, A. K. et al. Automated classification of Indian Mango varieties using machine learning and MobileNet-v2 deep features. *Trait Signal.* **41**, 1–10. https://doi.org/10.18280/ts.410210 (2024).
16. Katarzyna, R. & Paweł, M. A vision-based method utilizing deep convolutional neural networks for fruit variety classification in uncertainty conditions of retail sales. *Appl. Sci.* **9**, 3971. https://doi.org/10.3390/app9193971 (2019).
17. Taner, A. et al. Apple varieties classification using deep features and machine learning. *Agric* **14**, 252. https://doi.org/10.3390/agriculture14020252 (2024).
18. Zaki, N. et al. Transfer learning and explainable artificial intelligence enhance the classification of date fruit varieties. In 2023 15th Int. Conf. Innov. Inf. Technol. (IIT), IEEE, 222–227 (2023).
19. Sahu, P. et al. A systematic literature review of machine learning techniques deployed in agriculture: A case study of banana crop. *IEEE Access.* **10**, 87333–87360. https://doi.org/10.1109/ACCESS.2022.3199926 (2022).
20. Raghavendra, S. et al. Deep learning-based dual channel banana grading system using convolutional neural network. *J. Food Qual.* **2022** (6050284). https://doi.org/10.1155/2022/6050284 (2022).
21. Upadhyay, A. et al. Segregation of ripe and Raw bananas using convolutional neural network. *Procedia Comput. Sci.* **218**, 461–468. https://doi.org/10.1016/j.procs.2023.01.028 (2023).
22. Saranya, N. et al. Banana ripeness stage identification: A deep learning approach. *J. Ambient Intell. Humaniz. Comput.* **13**, 4033–4039. https://doi.org/10.1007/s12652-021-03267-w (2022).
23. Mohamedon, M. F., Abd Rahman, F., Mohamad, S. Y. & Khalifa, O. O. Banana ripeness classification using computer vision-based mobile application. In 2021 8th International Conference on Computer and Communication Engineering (ICCCE), 335–338. IEEE. (2021). https://doi.org/10.1109/ICCCE50029.2021.9467190
24. Narayanan, K. L. et al. Banana plant disease classification using hybrid convolutional neural network. Computational Intelligence and Neuroscience, 9153699. (2022). (1) https://doi.org/10.1155/2022/9153699 (2022).
25. Thiagarajan, J. D. et al. Analysis of banana plant health using machine learning techniques. *Sci. Rep.* **14** (1), 15041. https://doi.org/10.1038/s41598-024-63930-y (2024).
26. Sujithra, J. & Ferni Ukrit, M. Performance analysis of D-neural networks for leaf disease classification-banana and sugarcane. *Int. J. Syst. Assur. Eng. Manage.* 1–9. https://doi.org/10.1007/s13198-022-01756-5 (2022).
27. Vijayalakshmi, M. & Peter, V. J. CNN-based approach for identifying banana species from fruits. *Int. J. Inform. Technol.* **13** (1), 27–32. https://doi.org/10.1007/s41870-020-00554-1 (2021).
28. Widodo, D., Fauzi, A. & Sembiring, A. Identification of banana fruit types using the backpropagation method. *J. Artif. Intell. Eng. Appl. (JAIEA).* **3** (1), 300–307. https://doi.org/10.59934/jaiea.v3i1.314 (2023).
29. Rangkuti, A. H., Lau, S. L., Hasbi, V. A., Indallah, F. H. & Aryanto, R. Comparison of CNN models for optimizing banana image classification. In IEEE International Conference on Computing (ICOCO), 456–461. IEEE. (2023). https://doi.org/10.1109/ICOCO.2023.9682710 (2023).
30. Gupta, S. & Tripathi, A. K. Fruit and vegetable disease detection and classification: recent trends, challenges, and future opportunities. *Eng. Appl. Artif. Intell.* **133**, 108260. https://doi.org/10.1016/j.engappai.2024.108260 (2024).
31. Ferdaus, M. H. et al. BananaImageBD: A comprehensive image dataset of common banana varieties with different ripeness stages in Bangladesh. Mendeley Data, V2. (2024). https://doi.org/10.17632/ptfscwtnyz.2
32. Islam, M. M., Sheikh, R., Hossain, M. A., Hossain, M. & Himel, G. M. S. BananaSet: A dataset of banana varieties in Bangladesh. Mendeley Data, V4. (2024). https://doi.org/10.17632/35gb4v72dr.4
33. Abımouloud, M. L., Bensıd, K., Elleuch, M., Ammar, M. B. & Kherallah, M. Vision transformer-based convolutional neural network for breast cancer histopathological images classification. *Multimedia Tools Appl.* 1–36. https://doi.org/10.1007/s11042-024-19667-x (2024).
34. Jiang, X., Wang, S. & Zhang, Y. Vision transformer promotes cancer diagnosis: A comprehensive review. *Expert Syst. Appl.* **252**, 124113. https://doi.org/10.1016/j.eswa.2024.124113 (2024).

35. Kunduracioglu, I. & Pacal, I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. *J. Plant Dis. Prot.* **131** (3), 1061–1080. https://doi.org/10.1007/s41348-024-00896-z (2024).
36. Pacal, I., Kunduracioglu, I., Alma, M. H., Deveci, M., Kadry, S., Nedoma, J., … Martinek,R. A systematic review of deep learning techniques for plant diseases. Artificial Intelligence Review, 57(11), 304. https://doi.org/10.1007/s10462-024-10944-7 (2024).
37. Ergün, E., Aydemir, Ö. & Korkmaz, O. E. Investigating the informative brain region in multiclass electroencephalography and near-infrared spectroscopy-based BCI systems using band power-based features. *Comput. Methods Biomech. BioMed. Eng.* 1–16. https://doi.org/10.1080/10255842.2024.2333924 (2024).
38. Konstantinov, A. V. & Utkin, L. V. Interpretable machine learning with an ensemble of gradient boosting machines. *Knowl. Based Syst.* **222**, 106993. https://doi.org/10.1016/j.knosys.2021.106993 (2021).
39. Ergün, E. Artificial intelligence approaches for accurate assessment of insulator cleanliness in high-voltage electrical systems. *Electr. Eng.* 1–16. https://doi.org/10.1007/s00202-024-02691-3 (2024).
40. Aydemir, T., Şahin, M. & Aydemir, O. Sequential forward mother wavelet selection method for mental workload assessment on N-back task using photoplethysmography signals. *Infrared Phys. Technol.* **119**, 103966. https://doi.org/10.1016/j.infrared.2021.103966 (2021).
41. Yavuz, E. & Aydemir, Ö. Olfaction recognition by EEG analysis using wavelet transform features. In 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA) (pp. 1–4). IEEE. (2016). https://doi.org/10.1109/INISTA.2016.7571856
42. Yavuz, E. & Aydemir, Ö. Classification of EEG based BCI signals imagined hand closing and opening. In 2017 40th International Conference on Telecommunications and Signal Processing (TSP) (pp. 425–428). IEEE. (2017). https://doi.org/10.1109/TSP.2017.8076074
43. Pangestu, A. Banana ripeness image [Dataset]. *Telkom Univ. Dataverse*. **V1** https://doi.org/10.34820/FK2/GJBZ0X (2023).
44. Pangestu, A., Purnama, B. & Risnandar, R. Vision transformer Untuk Klasifikasi Kematangan Pisang. *Jurnal Teknologi Informasi Dan. Ilmu Komputer*. **11** (1), 75–84. https://doi.org/10.25126/jtiik.20241117389 (2024).
45. Sheikh, M. R., Hossain, M. A., Hossain, M., Islam, M. M. & Himel, G. M. S. BananaSet: A dataset of banana varieties in Bangladesh. *Data Brief.*, **54**, 110513. https://doi.org/10.1016/j.dib.2024.110513 (2024).

Acknowledgements

This research was financially supported by the Recep Tayyip Erdogan University Development Foundation (Grant number: 02025001027264). We sincerely appreciate their support, which contributed significantly to the completion of this study.

Author contributions

E. Ergün conceived the manuscript, designed the methodology, developed the software, and authored the initial draft. Additionally, E. Ergün contributed to the software development and validation process.

Funding

This study has been supported by the Recep Tayyip Erdoğan University Development Foundation (Grant number: 02025001027264).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.E.

Reprints and permissions information is available at www.nature.com/reprints.

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