Smart Agricultural Technology 12 (2025) 101333

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Assessing deep learning model robustness for banana ripeness classification under varying illumination conditions

Luis E. Chuquimarca [a,b,*], Boris X. Vintimilla [b], Sergio A. Velastin [c,d]

[a] FACSISTEL, Universidad Estatal Península de Santa Elena, La Libertad, Ecuador
[b] CIDIS, ESPOL Polytechnic University, ESPOL, Campus Gustavo Galindo, Km. 30.5 Vía Perimetral, Guayaquil, 090902, Ecuador
[c] School of EECS, Queen Mary University of London, London, UK
[d] Dept. of Computer Engineering, University Carlos III, Madrid, Spain

ARTICLE INFO

Keywords:
Banana ripeness classification
Deep learning
Illumination robustness
LIME

ABSTRACT

The classification of banana ripeness is critical for the external assessment of fruit quality in automated inspection systems. This study investigates the robustness of deep learning models under varying illumination conditions, utilizing a dataset of real banana images captured in multiple lighting scenarios. Models were trained and evaluated with gamma-based illumination augmentation generated through the Low-light Image Enhancement model to simulate realistic variability. Although training on pristine images yields high baseline accuracy, the models exhibit pronounced performance drops under altered illumination, highlighting significant overfitting issues. Augmentation substantially improves robustness, but the degree of improvement is architecture-dependent and does not fully mitigate vulnerabilities under extreme lighting conditions. Furthermore, the trade-offs between model complexity, inference efficiency, and generalization capacity are critically analyzed, revealing constraints for real-time and resource-limited applications. In particular, baseline accuracies reached an average of 92.5% under ideal lighting (I0), with ViT scoring 93.13% and InceptionV3 trailing at 89.84%. Under moderate gamma augmentation ($\Gamma = 0.4$), performance dropped by up to 11.8 percentage points (InceptionV3: from 89.84% to 77.63%). Training with augmented data restored accuracy by 8.3 pp on average, with InceptionV3 recovering to 91.88% and ViT maintaining 92.90%. Inference times ranged from 2.08 ms (ViT) to 3.89 ms (InceptionV3), demonstrating feasibility for real-time deployment. These findings emphasize the importance of incorporating environmental variability into evaluation protocols to ensure a reliable deployment of automated grading systems in practical scenarios.

1. Introduction

Assessment of banana ripeness is a key parameter in ensuring quality and market value [1]. Ripeness significantly influences every stage of the production and commercialization chain, determining the handling and the final destination [2,3]. During harvest, bananas are harvested in their physiologically green stage (the minimum level of colorimetry) to prevent damage and extend their shelf life during transport [4]. In the post-harvest phase, the ripeness level defines the selection, classification, and packaging processes, as international markets require fruit that meets specific conditions to withstand long-distance shipping [5,6]. During distribution and marketing, control of ripening with ethylene (internal inspection) or the use of a color reference (external inspection) allows the product to be adapted to market needs, ensuring it reaches the optimal stage for sale and consumption [7]. Finally, ripeness influences consumer choice, as green bananas are preferred for cooking, while riper bananas are ideal for direct consumption or use in the food industry [8]. However, inspections are often performed manually, making them slow and prone to human error [9]. Consequently, developing technological applications that integrate machine vision and deep learning (DL) models could enable more precise classification and digitalization of processes for the benefit of all stakeholders in the supply chain [10–12].

The external quality inspection of bananas generally involves a visual evaluation of the color of the peel, the appearance of spots, and the overall condition of the surface. Traditionally, this inspection is performed manually by trained personnel, following criteria defined by established international standards such as USDA, OECD, or similar ripening guide-

* Corresponding author at: FACSISTEL, Universidad Estatal Península de Santa Elena, La Libertad, Ecuador.
E-mail address: lchuquimarca@upse.edu.ec (L.E. Chuquimarca).

https://doi.org/10.1016/j.atech.2025.101333
Received 29 May 2025; Received in revised form 5 August 2025; Accepted 18 August 2025

Available online 26 August 2025

lines. However, manual inspection is subjective, labor intensive, and prone to inconsistencies, highlighting the need to implement automated and efficient methods [13].

Automating this classification using DL models not only provides consistency and reliability, but also enhances efficiency by significantly reducing human errors and subjectivity in quality inspections [14].

DL models, such as Convolutional Neural Networks (CNN) and Vision Transformer (ViT), have been successfully applied to classify fruit ripeness during the post-harvest stage.

Saranya et al. [15] presents a proposal to classify four stages of banana ripeness using a customized CNN model, which is compared to the VGG16 and ResNet50 models, achieving a validation accuracy of 96.14% with both original and enhanced data. The study shows that the proposed model uses fewer parameters and requires less training time than the state-of-the-art models, while also benefiting from data augmentation. For the classification, four stages of ripeness (ripe, partially ripe, very ripe, and overripe) [16], with a total of 273 images (104 in the first stage, 48 in the second, 88 in the third and 33 in the fourth). However, the limited amount of data and the unbalanced class distribution affect the generalization of the model to more varied scenarios. Likewise, variations in lighting and real field conditions remain a challenge.

Chuquimarca et al. [17] focuses on classifying the levels of ripe bananas using CNN models and proposes building a robust dataset that combines real and synthetic images. The authors introduce a simple custom CNN architecture, initially trained with synthetic data and later refined through transfer learning, then evaluated with real data. Multiple models were tested including ResNet50, VGG19, InceptionV3, and InceptionResNetV2, with the proposed model achieving a higher accuracy of 91.70%. The study highlights that generating synthetic banana images is more cost-effective and efficient than collecting large volumes of real images, although it notes the limitation of an imbalance in the number of real images for each ripeness level and the absence of environmental considerations.

Knott et al. [18] proposes an approach based on pre-trained ViT to facilitate banana ripeness classification and apple defect detection without the need to train traditional CNNs, achieving competitive accuracy (within 1% of the best CNN) and requiring up to three times fewer training samples to reach 90% accuracy. It also emphasizes the importance of Green AI by reducing costly computational demands and large data volumes. However, current research does not cover Transformer optimization or fine-tuning, nor does it address the method's adaptability to environments with limited data and basic hardware.

Arunima et al. [19] presents a CNN-based model to classify the ripeness of Nendran bananas during the post-harvest stage, using a data set of 4,320 images and comparing its performance against established models such as VGG16, VGG19, InceptionV3, ResNet50, and EfficientNetB0. The study highlights that the proposed CNN, consisting of nine layers, achieves an outstanding accuracy of 95%, outperforming the other evaluated architectures. However, although the research demonstrates promising results, it lacks a more detailed critical analysis regarding key aspects such as the model's robustness against variations caused by irregular illumination on the fruit.

In Chang et al. [20] a hybrid physics based and learning based color constancy method was used to preprocess images of unharvested palm fruit bunches before training a YOLOv8 detector yielding a 1.5% increase in mAP under variable lighting. This work underscores the importance of illumination normalization in reducing color shifts and enhancing model robustness in outdoor scenarios and provides a methodological precedent for our illumination augmentation strategy in banana ripeness classification.

In a recent study [21] propose a hybrid attention convolutional network for avocado ripeness classification on resource constrained devices combining spatial channel and self attention modules to capture local blemish and texture features alongside global dependencies Through transfer learning on EfficientNetB3 and MobileNetV3 they achieve over 91% accuracy while preserving inference speed and memory footprint suitable for smartphones This methodology underscores the power of attention mechanisms to improve the robustness and efficiency of lightweight models under visual variability and provides valuable guidance for banana ripeness classification under varying illumination where precision and deployment on modest hardware must be carefully balanced.

Correcting low-light images is critical, because banana ripeness is primarily assessed by the green-to-yellow chromatic ratio and the appearance of browning. Classical histogram-based methods and several recent DL approaches often raise brightness at the cost of oversaturating those channels, thereby shifting the color distribution [22–24]. Consequently, any illumination-enhancement mechanism must preserve the image's original colorimetry. To evaluate the suitability of different illumination-enhancement algorithms, we reviewed the most influential methods published in recent years. Table 1 compares them in terms of training requirements, main limitations, and technical advantages, with particular attention to maintaining the green-to-yellow color balance that is vital for estimating banana ripeness.

The comparison shows that DL model enhancers such as Retinex-Net, Zero-DCE, EnlightenGAN, LLFlow and SCI achieve impressive perceptual quality, but rely on large training datasets and often introduce color artefacts either oversaturation or texture smoothing. In contrast, LIME operates without annotated data, preserves the critical green-to-yellow chromatic balance required for the evaluation of ripeness and has a lightweight implementation, making it the most suitable and cost-effective option for enhancing banana images captured under uncontrolled lighting.

De-Arteaga et al. [29] identify two main challenges for deploying computer vision and deep learning in post-harvest quality assessment:

- *Dataset limitations:* Images captured with smartphones along the production chain lack a standardized acquisition protocol, producing lighting shifts that undermine model robustness. The limited number of real samples—due to the high cost and effort of acquiring images throughout the ripening process—and the unpredictable variability in operational environments further constrain the representativeness of training data.
- *Resource constraints:* Farmers and packing-house operators often cannot afford specialized hardware, so models must run efficiently on standard devices such as mobile phones or low-cost cameras to enable practical adoption.

This study addresses these issues by employing lightweight DL architectures capable of real-time execution on commodity hardware and by extending previous work [17] through the introduction of controlled illumination variations to rigorously evaluate model stability under realistic lighting changes.

The main contributions of this article are:

- The introduction of lighting variations into the dataset for banana ripeness classification, aiming to evaluate and improve model robustness and stability of the model under varying illumination conditions.
- A detailed comparison of DL models considering the number of parameters, inference time, and efficiency, providing clear criteria for selecting the most efficient and suitable model for real-world applications.

The paper is structured as follows. Section 2 outlines the overall methodology, beginning with the definition of ripeness levels and dataset construction, followed by the illumination-variation procedure, the transfer-learning approach, the DL models considered and the evaluation protocol. Section 3 presents the experimental findings under various lighting conditions, while Section 4 highlights the main contributions and practical implications of the study.

Table 1
Comparison of recent low-light enhancement methods and their suitability for banana-ripeness classification.

Method	Training required	Main limitations	Key advantage for this study
LIME [25]	No	Slight edge sharpening; no curve fine-tuning	Preserves the green–yellow chromaticity critical for ripeness assessment
Retinex-Net [26]	Yes	Requires low/high-light pairs and a large labeled dataset	Recovers fine detail in shaded regions
Zero-DCE [22]	Yes	Sensitive to uniform textures; needs thousands of images to generalize	Compact network; pair-free training
EnlightenGAN [23]	Yes	Over-saturation in R and G channels; large unpaired dataset needed	Global perceptual enhancement while preserving structure
LLFlow [27]	Yes	High VRAM usage; extensive data for fine-tuning	SOTA visual quality; consistent brightness via invertible flow
SCI [28]	Yes	Manual curve adjustment per batch; requires wide lighting diversity in training	Fast inference on GPU; pair-free training

2. Materials and methods

This section details the experimental procedure. First, a four-level banana ripeness taxonomy (Green, Partially ripe, Ripe and Over-ripe) is established as the reference for all subsequent analyses. Next, we describe the construction of a dataset comprising 3,495 real images and 161,280 synthetic images generated with a 3D banana model in Unreal Engine. Then, we apply the Low-light IMage Enhancement (LIME) algorithm to create seven brightness enhancement variations ($\Gamma = 0.1$ to 0.7) per image while preserving the critical green-to-yellow chromatic relationship. A transfer learning strategy reuses weights pretrained on the synthetic data to accelerate convergence and improve the generalization of the DL model. In this work, we evaluate six architectures (ResNet50, VGG19, InceptionV3, Inception-ResNetV2, CIDIS and Vision Transformer) on the augmented dataset across varying brightness levels. Finally, we define an evaluation protocol that includes accuracy, precision, recall, and F1 score, as well as model parameter counts and per-image inference time under each illumination condition, employing stratified splits to rigorously compare the models.

2.1. Banana ripeness

Banana ripeness plays a crucial role in determining fruit quality, directly influencing post-harvest management, market acceptance, and consumer preference. An accurate classification of banana ripeness ensures that the fruit meets international quality standards, affecting its marketability, shelf life, and final use [30].

Conventionally, bananas are classified into multiple ripeness stages based primarily on the coloration and appearance of the peel. Although there are up to seven standard ripeness levels, they are commonly grouped into fewer categories to simplify and clarify the process. In this research, four main ripeness levels have been established, aligned with industry practices and market requirements [17] (see Fig. 1):

- **Level A (Green):** Completely green peel, firm texture, predominantly starch composition.
- **Level B (Partially ripe):** Predominantly green peel with initial yellow spots appearing, less firm texture, and beginning of starch-to-sugar conversion.
- **Level C (Ripe):** Completely yellow peel, optimal softness for immediate consumption, predominantly sugar composition.
- **Level D (Overripe):** Significant presence of brown spots or areas on the peel, softer texture, high sugar content, suitable for immediate consumption or processing.

2.2. Dataset generation

In this study, we consider the use of the publicly available data set at https://github.com/luischuquim/BananaRipeness, which provides images of bananas in four different ripe stages [17]. This dataset includes both real and synthetic images, expanding the possibilities for training and evaluating DL models. However, for this study, greater emphasis is placed on real data, as it more accurately reflects the natural variations in the appearance of bananas under different conditions of light and ripeness.

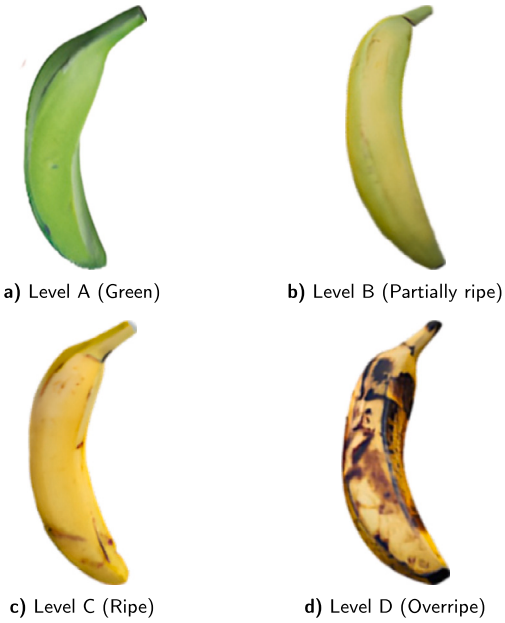

Fig. 1. Banana image samples representing the four ripeness levels used in this work: a) Level A (Green), b) Level B (Partially ripe), c) Level C (Ripe), and d) Level D (Overripe).

The dataset consists of 3,495 real images of Cavendish bananas, captured within a controlled climatic environment at temperatures between 15 °C and 18 °C over a period of 28 days, which approximates the full ripening duration of this banana variety from the first day of the post-harvest stage [31]. In addition, four levels of banana ripeness were considered, corresponding to one level of ripeness per week. Consequently, the real data set has an imbalanced number of images at each ripeness level due to acquisition complexity, which hinders the efficiency of DL models.

The development of this real data set is costly and tedious, as it requires dedicated personnel to supervise the data acquisition process throughout the banana ripening period. In addition, the environmental conditions where bananas are stored must be carefully controlled. Consequently, to obtain a sufficiently large number of images necessary for training DL models, the acquisition process must be repeated multiple times, resulting in significant costs and considerable staff time. Currently, there are technological tools that facilitate the generation of synthetic datasets; therefore, such tools, including the Unreal Engine, are being explored to generate a greater number of images per ripeness level [32,33].

Synthetic datasets represent an important complement for fruit data augmentation due to their low cost and the ease of generating large amounts of high-quality synthetic images similar to real images [34,35]. The process of generating the synthetic banana image dataset for differ-

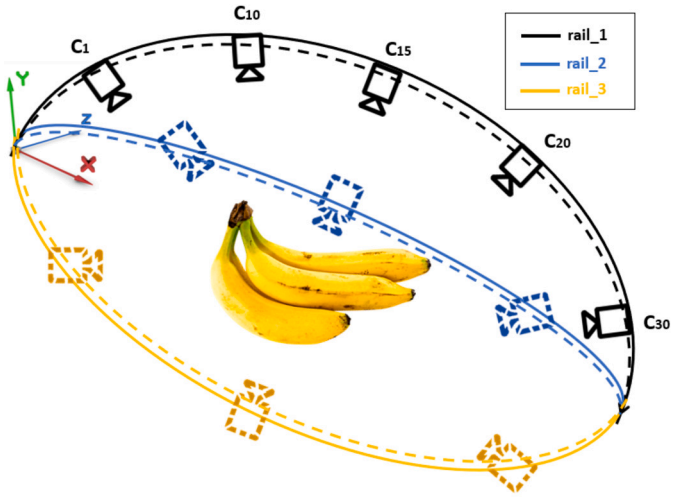

Fig. 2. Virtual scenario for the generation of the synthetic images using Unreal Engine [17].

Table 2
Summary of the banana dataset and CIDIS model performance [17].

Banana Ripeness Level	Number of Real Images	Number of Synthetic Images
Level A	1,429	40,320
Level B	815	40,320
Level C	559	40,320
Level D	692	40,320

ent ripeness levels is carried out by creating a virtual scenario using the Unreal Engine software tool, in which a 3D banana model is employed according to a given ripeness level to produce synthetic images [17]. The artificial appearance of banana ripeness is generated using layers derived from real images. The virtual environment includes three rails (rail-1, rail-2, and rail-3), each equipped with cameras mounted at 30 different positions and angles (positions C1–C30), enabling the acquisition of synthetic 2D RGB images of bananas at specific ripeness stages (see Fig. 2). The size of the synthetic images is 224x244 pixels, the same as the real dataset.

To provide greater variability to the virtual scenario, the synthetic dataset is configured with eight different backgrounds: orange, purple, brown, light blue, asset platform, basic wall, concrete tiles, and rock marble. Thus, considering all proposed scenario combinations, the total number of synthetic images generated is 161,280, approximately 40 times greater than the number of images in the real dataset, with 40,320 synthetic images per maturity level (see Table 2).

Variability in lighting conditions poses a critical challenge in banana ripeness classification, as illumination fluctuations are inevitable in industrial post-harvest environments and directly affect color perception in RGB images. Reproducing this diversity of lighting scenarios in real world settings entails high costs, prolonged acquisition times, and considerable logistical effort, limiting the representativeness of the collected data. In this context, synthetically augmenting illumination variations using the LIME model is justified as an effective strategy to simulate realistic lighting conditions in a controlled and cost-efficient manner. This intervention is applied exclusively to the real dataset, as it more faithfully captures the complexities inherent to the ripening process and the natural conditions of data acquisition, in contrast to the synthetic dataset. Consequently, the models' ability to generalize under variable lighting scenarios can be enhanced since the diversity of the training set is increased without the need for additional data acquisition processes, and the practical relevance of the experiments for real-world operational contexts is preserved.

2.3. Lighting variation model

Images captured in low-light conditions often exhibit poor visibility, negatively affecting not only their visual quality, but also the performance of DL models that rely on high-quality inputs [36,37]. To address this issue, this work utilizes the Low-light Image Enhancement (LIME) method, originally proposed to improve the luminosity of images.

The LIME method enables the generation of lighting variations in the real banana dataset, thus increasing both the quantity and diversity of data available to train intelligent ripeness classification models. Specifically, LIME first estimates the illumination of each pixel by determining the maximum value among the R, G, and B channels of each image [25]. Subsequently, this initial illumination map is refined by applying structural constraints to achieve a more consistent final illumination map [38]. Using this enhanced map, the illumination of the original images is effectively improved.

The application of the LIME method to the real dataset allows an increase in the amount of available data through the generation of images with controlled lighting variations, thus potentially enhancing the generalization capability of DL models under diverse lighting conditions [39]. In addition, previous studies have shown that LIME offers superior performance compared to other state-of-the-art methods in terms of image quality and processing speed, ensuring improvements in related tasks such as object detection, recognition, and classification [36,40]. Therefore, the use of LIME contributes significantly to the robustness and versatility of DL models developed for classification under real and varying conditions.

Initially, LIME generates a preliminary illumination map by identifying the maximum value among the RGB channels for each pixel. Then, this initial map is optimized using a structural smoothing model to preserve visual coherence, solving an optimization problem via Fast Fourier Transforms (FFTs). Finally, the original image is corrected by dividing each RGB channel by this optimized illumination map, adjusting the intensity further through a specific gamma parameter to generate different lighting variations. This process enhances image visibility and visual quality, making the images more suitable for applications in DL models. Additionally, by varying the gamma value, it is possible to obtain multiple enhanced versions of the same image, thereby enriching the diversity and quantity of the training dataset.

Algorithm 1 LIME applied to banana dataset augmentation.

Require: Original dataset images $\{I_n\}_{n=1}^{N}$, gamma values $\Gamma = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$
Ensure: Augmented dataset images with illumination variations
for $n = 1$ to N **do**
  $I \leftarrow$ Load image I_n
  for γ in Γ **do**
    Initialize illumination map $\hat{T}$ as the maximum of R, G, B channels per pixel
    Refine $\hat{T}$ by applying structural constraints:
    Solve the following iterative optimization:
    $T^{(t+1)} \leftarrow$ Optimize illumination map
    $G^{(t+1)} \leftarrow$ Update gradient subproblem
    $Z^{(t+1)} \leftarrow$ Update dual variable
    $u^{(t+1)} \leftarrow \rho \cdot u^{(t)}$
    Repeat until convergence is reached
    Apply gamma correction: $T \leftarrow (T^{(final)})^{\gamma}$
    Enhance image: $I_{\text{enhanced}} \leftarrow I/T$
    Normalize and convert I_{enhanced} to `uint8` format
    Save enhanced image with identifier according to γ
  end for
end for

Algorithm 1 describes the process of augmenting the banana dataset through LIME. Initially, each original image from the dataset (I_n) is individually loaded, and a set of predefined gamma values ($\Gamma = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$) is utilized to generate multiple enhanced

Table 3
Number of real banana images per gamma level (Γ), with brightness progressively enhanced using the LIME model.

Banana Ripeness Level	Γ = 0.1	Γ = 0.2	Γ = 0.3	Γ = 0.4	Γ = 0.5	Γ = 0.6	Γ = 0.7	Total Images
Level A	1,429	1,429	1,429	1,429	1,429	1,429	1,429	10,003
Level B	815	815	815	815	815	815	815	5,705
Level C	559	559	559	559	559	559	559	3,913
Level D	692	692	692	692	692	692	692	4,844

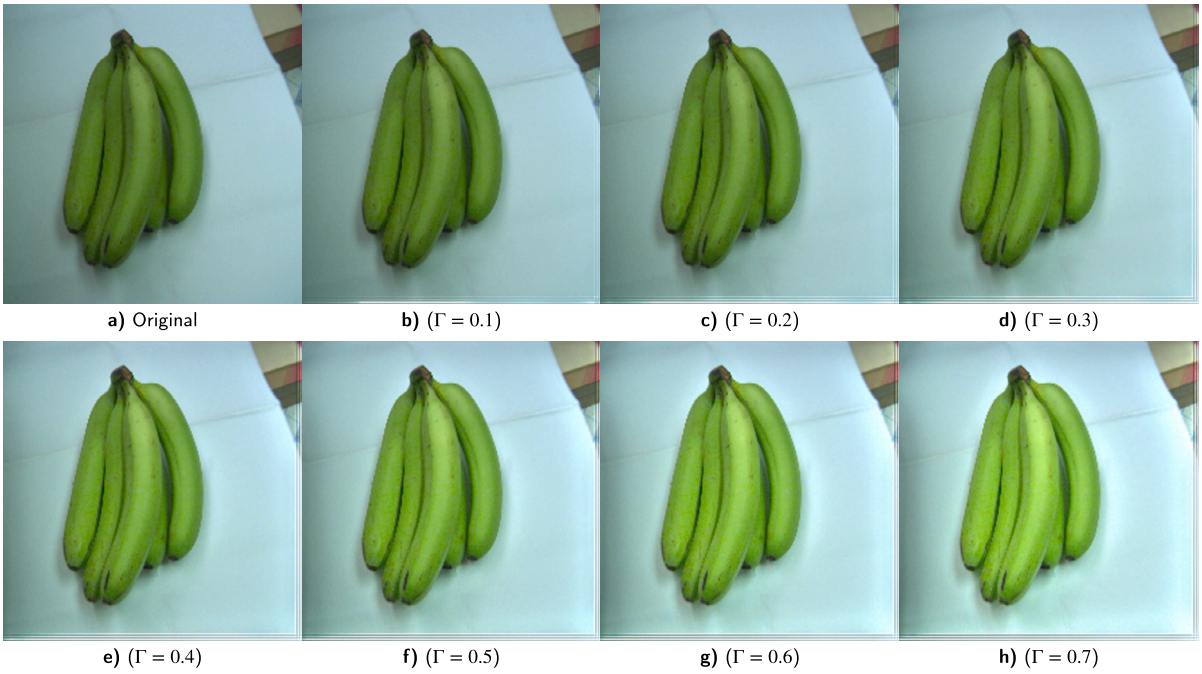

Fig. 3. Banana images showing progressive brightness enhancement as the gamma value (Γ) increases from 0.1 to 0.7, using the LIME model.

versions with varying illumination conditions. In this context, higher gamma values result in brighter outputs, while lower values produce more subtle enhancements.

For each gamma value, the algorithm first estimates an initial illumination map ($\hat{T}$) by computing the maximum intensity among the RGB channels for each pixel, providing a baseline for further refinement. Subsequently, structural constraints are applied to refine this illumination map, solving iteratively an optimization problem to ensure illumination consistency and image detail preservation. This iterative process updates the illumination map ($T^{(t+1)}$), gradient subproblem ($G^{(t+1)}$), dual variable ($Z^{(t+1)}$), and scaling parameter ($u^{(t+1)}$), continuing until convergence criteria ars satisfied, typically based on the Frobenius norm.

Once optimized, gamma correction is applied to the final illumination map to adjust the brightness level depending on the selected gamma value, using the transformation $T \leftarrow (T^{(final)})^{\gamma}$. The original image is then enhanced by dividing it pixel-wise by the gamma-corrected illumination map, significantly improving visibility under low-light conditions. Finally, enhanced images are normalized, converted to 8-bit unsigned integer format, and saved with filenames that indicate the applied gamma values.

This method effectively generates an augmented dataset with controlled lighting variations, aiming to enhance the robustness and generalization capabilities of DL models for banana ripeness classification under diverse illumination conditions (see Table 3 and Fig. 3).

2.4. Transfer learning

Transfer learning techniques are applied, utilizing synthetic image datasets to reduce the dependency on large volumes of real images. This methodology involves initially training DL models with synthetic images, followed by fine-tuning with real images to train and validate the models. This approach allows the models to leverage weights obtained during the transfer phase [41], to improve their generalization ability and accuracy in classifying banana ripeness levels (see Fig. 4).

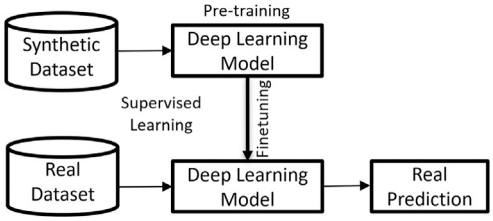

Fig. 4. Transfer learning using synthetic dataset.

2.5. Deep learning models

The work here is based on applying DL models to extract essential features for banana ripeness classification during external fruit quality inspection [42]. This approach involves selecting DL models designed to accurately classify four banana maturity levels.

These DL models are selected based on a state-of-the-art review and trained with a suitable dataset of images for each banana ripeness level, following the indications provided by international standards for external fruit quality inspection. Additionally, the use of techniques such as transfer learning and dataset variability is incorporated to enhance model effectiveness.

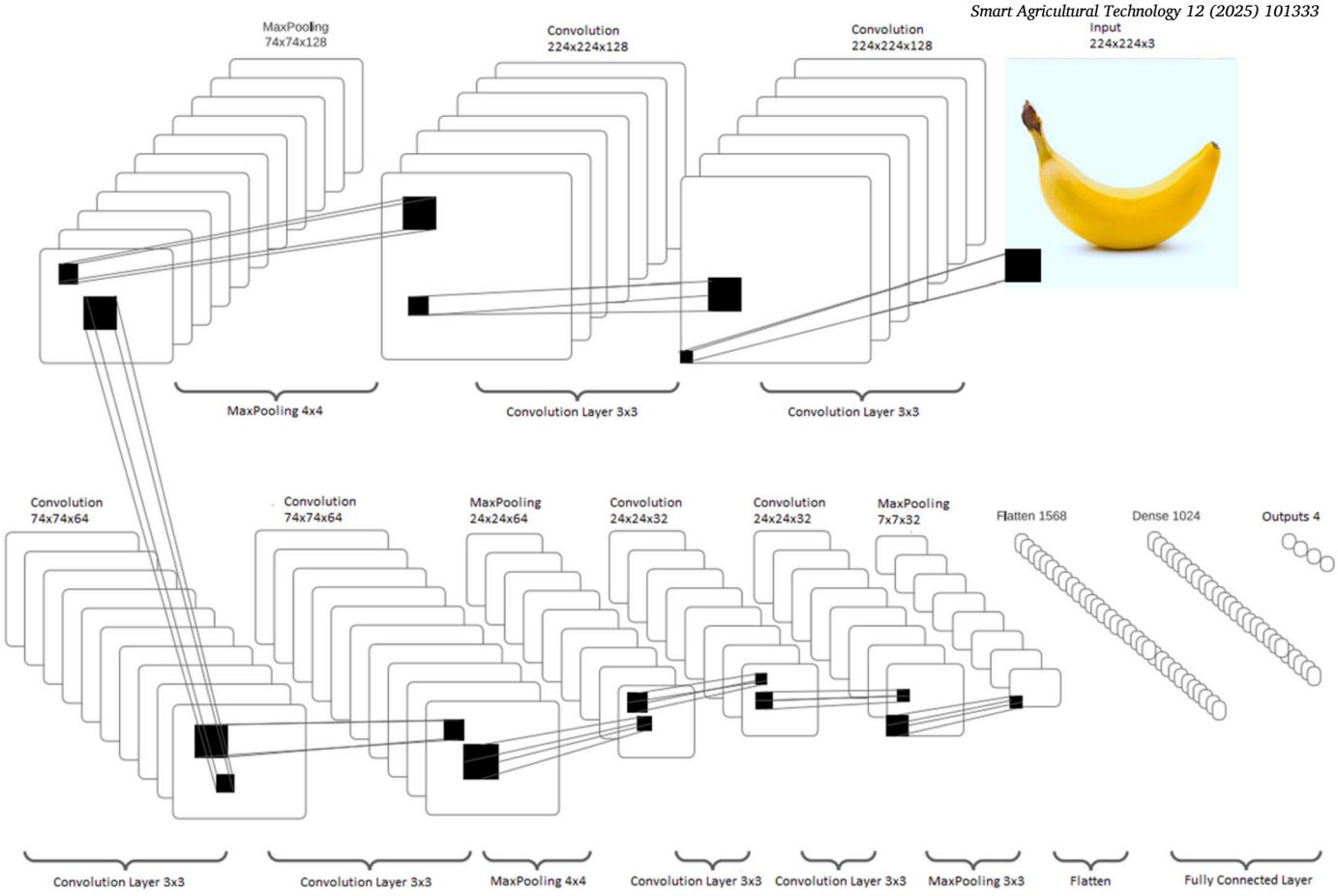

Fig. 5. CIDIS model architecture.

2.5.1. CIDIS model

The CIDIS model consists of repeated layers, including two convolution layers followed by a max-pooling layer, repeated three times, incorporating ReLU activation in hidden and fully connected layers (see Fig. 5). It receives RGB images of 224x224 pixels as input and produces four outputs, considering, for example, four levels of banana ripeness. The transfer learning technique is applied to this model, using weights learned from synthetic images, with the last fully connected layers removed and only these layers trained. The transfer learning model is optimized by adjusting hyperparameters, using dropout layers, modifying learning rates and batch sizes, and choosing between the Nadam and Adagrad optimizers. These optimizations lead to a robust model that predicts banana ripeness levels, which is finally evaluated on a dataset of real images. Previous results have indicated that the transfer learning model outperforms the model without transfer learning, trained only on real images [17].

2.5.2. Vision transformer model

The ViT architecture has emerged as a promising alternative to traditional CNN models commonly used in image classification tasks. Recent studies indicate that ViT models can achieve superior performance compared to current state-of-the-art CNN architectures. One distinctive advantage of ViT architectures is their ability to capture global context from the very beginning of the network, unlike CNN models whose initial receptive fields are inherently local and limited. Additionally, ViT models offer a high degree of parallelization, enabling more efficient computation, especially beneficial when processing large-scale image datasets. [43–45].

The attention mechanism enhances the relevant parts of the input data while fading out the rest. A self-attention module replaces the convolutional layer, allowing the model to interact with pixels that are distant from its location. Self-attention is a mechanism that allows each element in a sequence to interact with others, determining to what they should give more attention. This distinctive behavior arises from including certain inductive biases in CNN models. In contrast, ViT models lack these biases, which they can leverage to quickly grasp the nuances of the analyzed images, even capturing global relationships, usually at the expense of requiring more extended data training [46,47].

The ViT model comprises multiple blocks. As a self-attention method, the MultiHeadAttention layer is applied to the sequence of patches. The Transformer blocks generate a tensor of shape (batch size, number of patches, projection dimension), which a SoftMax classifier head processes to produce the final class probability distributions for the output classes.

A detailed framework diagram of the ViT employed in this paper is presented in Fig. 6. Pioneering work by Han et al. [44] introduced the Transformer model to the computer vision domain, resulting in the development of the ViT architecture. This architecture consists of three core components: patch embedding, feature extraction using stacked Transformer encoders, and a classification head. ViT processes images by dividing them into smaller fixed-size “patches”, significantly reducing image dimensionality and enabling efficient handling of higher-resolution images compared to traditional models. Unlike conventional CNN-based approaches, ViT utilizes self-attention mechanisms to extract global image features, resulting in superior training performance even with limited data.

To illustrate, the initial input image, like that of a banana, is subdivided into image patches (e.g., 16x16). Subsequently, these groups of image patches are embedded into encoded vectors and introduced into the Transformer encoder network. The Transformer encoder learns the features of these embedded patches through a stack of Transformer encoders [48]. The encoder primarily comprises multi-head attention

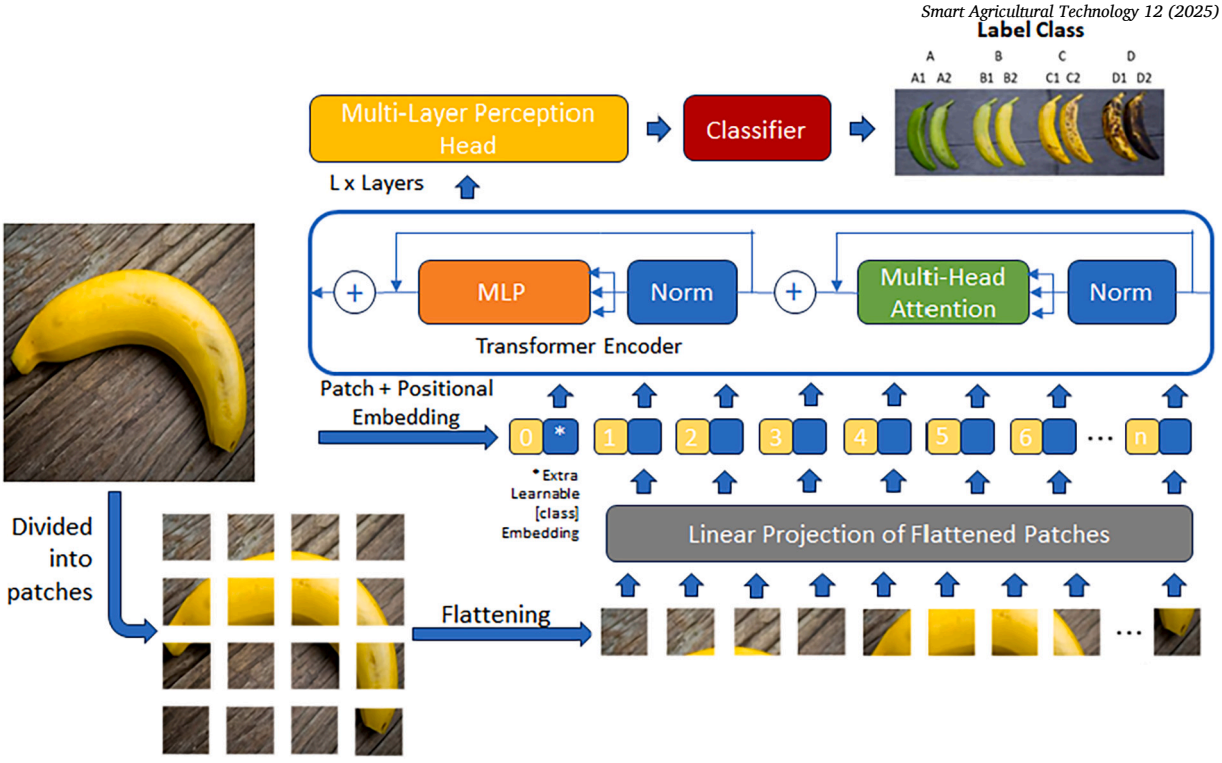

Fig. 6. Vision transformer model architecture [49,44].

(MHA) and a 2-multi-layer perceptron (MLP) with layer normalization and residual connections. The final MLP block, the MLP head, serves as the Transformer's output. In the context of image classification, a Softmax function at the output generates the classification results. Our study used the ViT-B/16 model (where B stands for Base, signifying a relatively small dataset; 16 indicates the input patch size of 16x16) trained using ViT.

The ViT models transform a sequence of image segments into a semantic label through the Transformer encoder module designed for classification tasks. Unlike traditional CNN architectures, which commonly utilize filters with a limited receptive field, ViT employs an attention mechanism capable of selectively focusing on different regions of the image and interpreting information across its entirety. Due to these characteristics, ViT became the first image recognition model capable of consistently outperforming conventional CNN architectures, especially those relying on constrained filter sizes. Structurally, the ViT architecture consists of three principal modules: an Embedding Layer, multiple Encoder blocks, and the final classification head layers [50]. Additionally, ViT's performance notably benefits from pre-training on large datasets, followed by fine-tuning for specific downstream tasks. To further reduce training duration and enhance computational efficiency, transfer learning methodologies are typically employed in training original ViT models [51].

2.5.3. VGG19 model

The VGG19 model is a classic design in the field of image recognition, demonstrating high performance and precision across a wide range of tasks. It consists of 19 layers of depth, including 16 convolutional layers and 3 fully connected layers. A key feature of VGG19 is its use of small 3x3 filters, allowing the network to be very deep, and learn hierarchical representations of images. VGG19 has shown notable performance with high accuracy and relatively low training times, especially when compared to more complex models such as InceptionResNetV2 [52]. However, while effective, VGG19 can be less efficient in terms of computational consumption compared to newer models, due to its relatively simple architecture and large number of parameters.

2.5.4. ResNet50 model

ResNet models (Residual Networks) are known for their ability to train very deep networks without encountering the vanishing gradient problem, which is achieved by using residual connections. ResNet50, with its 50 layers, is used in this study for its ability to handle deeper networks without losing information, which is common in traditional networks when more layers are added. Residual connections allow information to flow directly through layers, making it easier to train deeper models and improving accuracy. In particular, ResNet-50 is well suited for tasks such as fruit ripeness classification, as it can learn high-level features from complex data. Furthermore, the combination of transfer learning and residual learning in this model has been reported to optimize network parameters and enhance system development, leading to better generalization and reducing overfitting [53].

2.5.5. Inception-ResNetV2 model

The Inception-ResNetV2 model combines two powerful architectures: Inception and ResNet. Inception is known for its ability to capture information at different scales by using multiple filter sizes within the same layer, allowing the model to learn richer representations of images. On the other hand, ResNet contributes by enabling deeper networks through residual connections. Inception-ResNetV2 incorporates both ideas, making it even more efficient and accurate. This model has 164 layers, making it one of the deepest and most complex models in the field of computer vision. Its ability to handle different filter sizes and computational efficiency makes it ideal for complex image classification tasks, such as banana ripeness classification. Inception-ResNetV2 has proven to be highly effective in reducing loss during training and has outperformed other Inception models in terms of accuracy [54]. Furthermore, by incorporating 1x1 convolutions and residual connections, it optimizes computational resource usage and helps prevent overfitting [55], making it a solid choice for tasks that require both precision and efficiency.

2.5.6. InceptionV3 model

The InceptionV3 model is a more efficient version of the Inception model, designed to reduce computational power consumption while

improving efficiency without compromising performance. This model is more efficient than previous versions, such as VGGNet and InceptionV1, due to its focus on reducing computational complexity through techniques such as dimensionality reduction with 1x1 convolutions. InceptionV3 optimizes the use of computational resources by combining multiple operations in a single layer, allowing faster training speeds and reduced memory requirements while maintaining high accuracy. In the context of the estimation of banana ripeness, InceptionV3 has been shown to be more efficient than other traditional models, delivering competitive results in terms of accuracy while keeping computational costs low [56].

2.6. Performance evaluation

This section presents the evaluation strategy used to analyze the effectiveness and robustness of DL models under varying illumination conditions. Performance was assessed using standard classification metrics, which provide a comprehensive overview of DL model behavior, particularly in the presence of class imbalance [57,58]. In addition, the number of trainable parameters and the average inference time per image were considered to evaluate the computational efficiency of each DL model. These indicators enable a balanced assessment between predictive performance and practical feasibility in real-world deployment scenarios [29].

Accuracy measures the proportion of correctly predicted instances out of the total number of examples in the dataset and is defined in Equation (1). Precision, Recall, and F1-score, defined in Equations (2), (3) and (4) respectively, are complementary metrics commonly used in classification tasks [59]. Given the imbalanced nature of the real-world dataset and the tendency of accuracy to favor the majority class, which can lead to biased results [60], this comprehensive evaluation approach ensures a more reliable assessment of model performance.

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

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

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

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP refers to true positives, TN to true negatives, FP to false positives, and FN to false negatives.

Additionally, to evaluate the computing requirements of the models under varying lighting conditions, we considered not only the number of parameters but also the inference time during testing. These factors are relevant for assessing how cost effectively a model can handle real-world scenarios where lighting conditions may change dynamically [61]. The performance of the model under varying lighting conditions is not only about accuracy but also about how quickly and efficiently it can process the input data. In real-world applications, lighting variations can significantly affect performance [62], and the ability to maintain accuracy while processing efficiently is a key measure of robustness.

The number of parameters (P) in a model influences the complexity and capacity of the model to capture intricate patterns (see Equation (5)), but also directly impacts the computational cost during training and inference [63]. The larger the model, the more resources it requires. Inference time ($T_{inference}$) is critical to evaluate the speed at which the model processes each input (see Equation (6)), especially when working with large datasets or in real-time applications [64]. In our evaluation, we calculated the inference time per image for each model under different lighting conditions.

$$P = \sum_{l=1}^{L} \left(n_l^{\text{in}} \times n_l^{\text{out}} + n_l^{\text{out}} \right) \tag{5}$$

$$T_{\text{inference}} = \frac{t_{\text{total}}}{N} \tag{6}$$

where P is the total number of trainable parameters, L is the number of layers, n_l^{in} and n_l^{out} are the input and output units of layer l, $T_{\text{inference}}$ is the average inference time per image, t_{total} is the total time to process N images, and N is the number of images evaluated.

3. Results and discussion

This section presents the results obtained by evaluating several DL models under varying illumination conditions for the classification of banana ripeness. Initial models (denoted as M0) were trained exclusively on data without illumination variations. These baseline models are evaluated first on a test dataset without lighting variation (I0), followed by seven additional test sets (I1 to I7) that introduce incremental illumination variations defined by Γ, ranging from 0.1 (I1) to 0.7 (I7). A comparative analysis of model performance is presented, using accuracy, precision, recall, and F1-score metrics to identify the behavior of these models in response to gradual changes in image illumination.

For the initial conditions without illumination variation (I0), all evaluated models demonstrate strong and consistent performance, largely due to the uniform lighting of the test images. This controlled environment helps preserve the green–yellow chromatic balance that is critical for accurately identifying banana ripeness levels.

Analyzing Table 4, we observe that under conditions without illumination variation (I0), all models exhibit strong and stable results in terms of accuracy. Notably, the ViT model achieves the highest accuracy (93.13%), closely followed by ResNet50 (92.99%). In contrast, InceptionV3 shows the lowest accuracy under this standard condition (89.84%). Precision, recall, and F1-score metrics under the baseline condition are consistent with accuracy, maintaining values above 89% across most models.

The ViT model stands out with the highest accuracy (93.13% under I0), a result attributed to its global attention mechanism, which captures broader spatial and contextual relationships across the image, especially beneficial when training data is limited. In contrast, InceptionV3, which relies on localized receptive fields, achieves only 89.84% under the same conditions and thus performs less effectively in this context. These findings emphasize that, under ideal lighting, color fidelity and dataset quality are key factors driving performance in ripeness classification.

Under low illumination variations (I1 to I3, corresponding to $\Gamma = 0.1$ to 0.3), ResNet50 and CIDIS models maintain consistent and robust performance, each achieving accuracies above 90%, with CIDIS showing peak accuracy at 92.85% in I1. The ViT model, however, exhibits a noticeable decrease in accuracy starting from I2 (83.83%), demonstrating sensitivity to small lighting variations. The VGG19 model maintains acceptable accuracy through I2 but experiences a significant drop at I3, achieving only 79.97% accuracy, suggesting limited robustness to moderate illumination changes.

This behavior can be explained by the architectural design and generalization mechanisms of each model. ResNet50, with its residual connections, facilitates the learning of robust features that remain stable under slight illumination shifts, allowing it to sustain high performance even when lighting conditions are subtly altered. Although CIDIS does not incorporate residual connections, it is specifically designed as a lightweight and efficient model that effectively preserves key discriminative features, which may account for its stability under low illumination variation. In contrast, the ViT model, while excelling under ideal conditions due to its global attention mechanism, exhibits sensitivity to slight lighting changes, possibly because such variations affect the chromatic consistency across the patches it processes. VGG19, lacking mechanisms to compensate for illumination shifts, begins to degrade in performance as images deviate from the original training conditions. This decline reflects a limited capacity to adapt to alterations that directly impact the green–yellow chromatic balance, which is essential for correctly estimating the ripeness level of bananas.

Table 4
Performance metrics of DL models for banana-ripeness classification, trained on the baseline set M0 (no illumination variation) and evaluated on test sets I0–I7, where progressively higher gamma values ($\Gamma = 0.0$–0.7) are applied with LIME model to increase image brightness.

Illumination	Model	Accuracy	Precision	Recall	F1 Score
M0_I0	VIT	0.9313	0.9310	0.9313	0.9310
M0_I1		0.8956	0.8981	0.8956	0.8957
M0_I2		0.8383	0.8425	0.8383	0.8388
M0_I3		0.8813	0.8895	0.8813	0.8822
M0_I4		0.8884	0.8996	0.8884	0.8882
M0_I5		0.7210	0.7479	0.7210	0.7262
M0_I6		0.8598	0.8902	0.8598	0.8601
M0_I7		0.7396	0.8125	0.7396	0.7312
M0_I0	ResNet50	0.9299	0.9318	0.9299	0.9304
M0_I1		0.9227	0.9254	0.9227	0.9235
M0_I2		0.9185	0.9196	0.9185	0.9185
M0_I3		0.8970	0.9062	0.8970	0.8984
M0_I4		0.7639	0.7771	0.7639	0.7651
M0_I5		0.7983	0.8454	0.7983	0.7974
M0_I6		0.8569	0.8796	0.8569	0.8563
M0_I7		0.8169	0.8472	0.8169	0.8171
M0_I0	InceptionResNetV2	0.9242	0.9246	0.9242	0.9243
M0_I1		0.9213	0.9224	0.9213	0.9215
M0_I2		0.9027	0.9050	0.9027	0.9032
M0_I3		0.8727	0.8901	0.8727	0.8749
M0_I4		0.8870	0.9019	0.8870	0.8886
M0_I5		0.8598	0.8808	0.8598	0.8620
M0_I6		0.8469	0.8643	0.8469	0.8477
M0_I7		0.8255	0.8711	0.8255	0.8229
M0_I0	CIDIS	0.9227	0.9237	0.9227	0.9231
M0_I1		0.9285	0.9285	0.9285	0.9283
M0_I2		0.9099	0.9143	0.9099	0.9109
M0_I3		0.9013	0.9071	0.9013	0.9015
M0_I4		0.8712	0.8775	0.8712	0.8718
M0_I5		0.8755	0.8975	0.8755	0.8768
M0_I6		0.6781	0.7401	0.6781	0.6903
M0_I7		0.8255	0.8605	0.8255	0.8220
M0_I0	VGG19	0.9070	0.9062	0.9070	0.9062
M0_I1		0.9199	0.9215	0.9199	0.9203
M0_I2		0.8884	0.8937	0.8884	0.8885
M0_I3		0.7997	0.8068	0.7997	0.7998
M0_I4		0.8255	0.8628	0.8255	0.8250
M0_I5		0.8755	0.8897	0.8755	0.8744
M0_I6		0.7611	0.8224	0.7611	0.7567
M0_I7		0.6381	0.7205	0.6381	0.6498
M0_I0	InceptionV3	0.8984	0.8996	0.8984	0.8986
M0_I1		0.8727	0.8745	0.8727	0.8729
M0_I2		0.9070	0.9129	0.9070	0.9082
M0_I3		0.8870	0.8923	0.8870	0.8879
M0_I4		0.8655	0.8808	0.8655	0.8671
M0_I5		0.8627	0.8708	0.8627	0.8632
M0_I6		0.8512	0.8788	0.8512	0.8541
M0_I7		0.8255	0.8680	0.8255	0.8300

As illumination variation is increased significantly (I4 to I7, corresponding to $\Gamma = 0.4$ to 0.7), all DL models exhibit substantial declines in performance across all evaluated metrics. This degradation is primarily attributed to the distortion of key chromatic features, particularly the green–yellow hue relationship, which is essential for accurately classifying banana ripeness. Excessive brightness tends to saturate these color channels, compromising the discriminative cues that models rely on. For instance, VGG19, which strongly depends on color-based features, is the most affected, reaching its lowest accuracy of 63.81% at I7. ResNet50 suffers a noticeable drop at I4 (76.39%) but demonstrates partial recovery at I6 and I7, likely due to its residual connections that help maintain feature consistency despite input distortions. Similarly, the CIDIS model, while efficient, lacks architectural mechanisms to handle extreme lighting, leading to unstable performance, especially at I6 (67.81%). The ViT model, although initially robust in I4 (88.84%) due to its global attention mechanism, exhibits significant fluctuations under more severe conditions, dropping to 73.96% in I7, likely a result of losing patch-wise attention coherence in unevenly illuminated regions. In contrast, Inception-based models such as InceptionResNetV2 and InceptionV3 show a more gradual accuracy decline, maintaining approximately 82% even under strong illumination shifts, suggesting that their multiscale feature extraction offers some resilience to chromatic distortions.

Precision, recall, and F1-score metrics consistently follow the same patterns observed in accuracy across all illumination conditions, reinforcing the conclusion that lighting variation systematically affects the overall performance of DL models. These metrics, which evaluate the model's ability to correctly identify relevant classes and reduce false positives or negatives, are particularly sensitive to distortions in the visual features that guide classification. For example, ResNet50 maintains stable precision and recall values above 84% in most scenarios, except for extreme lighting conditions (I4 and I5), a robustness largely attributed to its residual connections that help preserve consistent feature representations. In contrast, the ViT model exhibits greater variability, especially in recall, which drops below 74% in I5 and I7. This decline suggests that the patch-based self-attention mechanism of ViT may struggle with local brightness inconsistencies, leading to reduced spatial coherence in the learned representations and ultimately affecting classification accuracy. Overall, the degradation of precision, recall, and F1-score under varying lighting confirms the importance of robustness in feature extraction for maintaining classification reliability in non-uniform visual environments.

From a global perspective, the results reveal that all evaluated models perform reliably under ideal lighting conditions (I0), maintaining high accuracy and consistent behavior across metrics. However, when illumination variability increases, particularly from $\Gamma = 0.4$ onward, most models begin to exhibit noticeable decrease in performance. This trend is reflected in reduced accuracy, precision, recall, and F1-score for the majority of the architectures. The varying degrees of sensitivity among the models suggest that their ability to generalize is closely related to the distribution and diversity of lighting conditions encountered during training. Therefore, these results underscore the critical need to incorporate controlled lighting variability into the dataset during training. Doing so enhances the robustness of the models and prepares them for deployment in real-world environments where illumination is rarely consistent.

Fig. 7 illustrates the performance behavior of the ResNet50, ViT, VGG19, CIDIS, InceptionResNetV2, and InceptionV3 models for different levels of illumination variation, ranging from I0 (no variation, $\Gamma = 0$) to I7 (maximum variation, $\Gamma = 0.7$). Accuracy was chosen as the metric for visualization, presented as continuous lines with specific markers for each model to facilitate comparative evaluation.

Under ideal conditions (I0), the graph shows that all models start with high accuracy values, with minimal differences between them. However, when illumination variation increases slightly (levels I1–I3), a clear differentiation in the performance curves of each model emerges. ResNet50 and CIDIS exhibit visually stable performance, characterized by a gentle and steady downward slope, indicating notable robustness against slight illumination changes. In contrast, the ViT model curve displays a visibly sharper decrease in accuracy starting at the I2 level, suggesting early sensitivity to moderate illumination variations. Similarly, VGG19 initially shows stable behavior but rapidly deteriorates at level I3.

At intermediate to high illumination variations (levels I4–I7), the graph highlights diverse patterns across the models. The VGG19 model demonstrates the steepest and most continuous decline, reaching the lowest accuracy level at I7 and indicating limited adaptability to adverse lighting conditions. CIDIS similarly presents a notable negative slope from I4 to I6 but slightly improves visually toward I7. The ViT model is characterized by pronounced fluctuations, visually alternating between significant drops and partial recoveries between consecutive

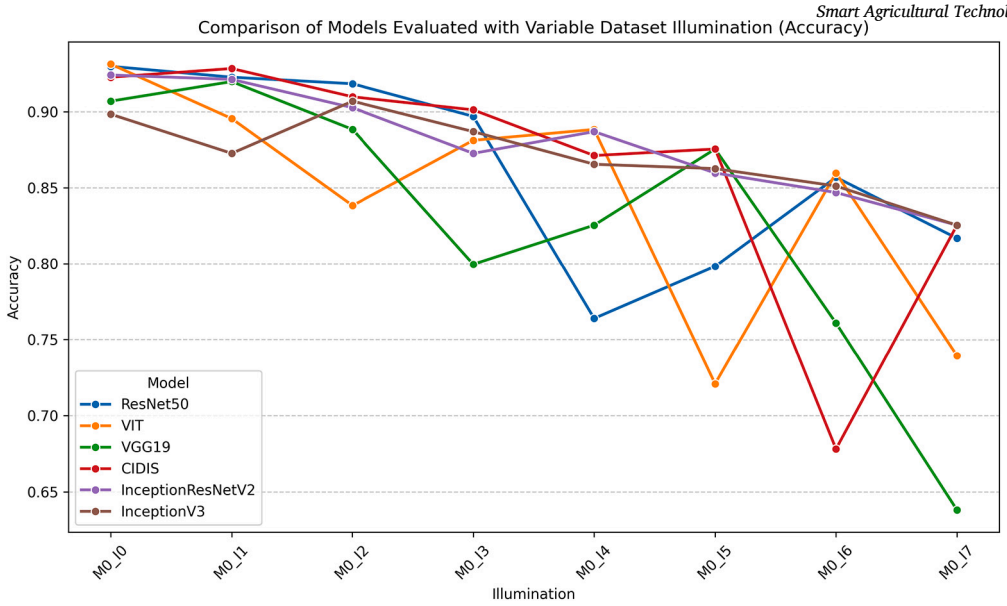

Fig. 7. Accuracy of DL models evaluated under illumination variation levels $\Gamma = 0.0$ to $\Gamma = 0.7$.

variation levels. ResNet50 experiences a clear visual drop at I4 but subsequently stabilizes, displaying slight accuracy recoveries in later conditions (I6–I7). InceptionResNetV2 and InceptionV3 curves appear visually smoother, showing a more gradual and consistent decline across all illumination variations.

Overall, the graph clearly demonstrates that, despite starting from comparably high accuracies under ideal illumination conditions, the models visually diverge significantly as illumination variation increases. ResNet50 and CIDIS stand out visually for their consistent performance under moderate variations ($\Gamma \leq 0.3$), suggesting robust generalization capability under these conditions. InceptionResNetV2 and InceptionV3 curves exhibit smoother and less abrupt reductions, visually indicating relatively greater resilience under substantial illumination variability ($\Gamma \geq 0.4$). Conversely, the ViT model visually reveals high variability and sensitivity to moderate and high illumination fluctuations. The VGG19 model clearly emerges visually as the least robust, exhibiting a continuous and significant decline in accuracy across increasing illumination variations. This visual analysis effectively highlights each model's distinct dynamic behavior, guiding informed model selection depending on anticipated illumination conditions in practical applications.

Table 5 reports the top-1 accuracy obtained by each DL model when trained either on the baseline set without illumination changes (M0) or on the fully augmented set that includes synthetic lighting variation (MG), evaluated for two test conditions: I0 (no variation) and IG (gamma-augmented). These tests allow us to quantify both the robustness loss caused by unseen lighting changes and the robustness gain provided by training with augmented data. The results are graphically represented in Fig. 8.

The results reveal significant differences in how DL architectures respond to illumination variability. DL models such as InceptionV3 exhibit a marked drop in accuracy, suggesting higher sensitivity to changes not represented in the training set. This performance degradation may be attributed to the model's architectural complexity and its tendency to overfit specific features of the data. In contrast, models like ViT and CIDIS, which employ a combination of convolutional and dense layers, demonstrate greater resilience to illumination changes, although they also experience accuracy losses, indicating that their generalization capacity is not entirely robust to complex input transformations. Deeper architectures, such as ResNet50 and InceptionResNetV2, despite their expressive power, show considerable performance drops, suggesting that their ability to learn deep representations makes them more susceptible to overfitting and less capable of generalizing to unseen conditions.

Table 5
Accuracy of each DL model under illumination variation; M0 denotes models trained on non-augmented data, whereas MG refers to models trained with illumination-augmented data.

Illumination	Model	Accuracy
M0_I0	ViT	0.9313
M0_IG		0.8907
MG_I0		0.9299
MG_IG		0.9290
M0_I0	CIDIS	0.9227
M0_IG		0.8872
MG_I0		0.9242
MG_IG		0.9224
M0_I0	ResNet50	0.9299
M0_IG		0.8446
MG_I0		0.9256
MG_IG		0.9272
M0_I0	InceptionResNetV2	0.9242
M0_IG		0.8791
MG_I0		0.9156
MG_IG		0.9219
M0_I0	InceptionV3	0.8984
M0_IG		0.7763
MG_I0		0.9156
MG_IG		0.9188
M0_I0	VGG19	0.9070
M0_IG		0.8718
MG_I0		0.9056
MG_IG		0.9101

Note: "M0_I0" and "M0_IG" refer to models trained on non-augmented data and tested on images with no illumination change or gamma augmentation, respectively. "MG_I0" and "MG_IG" refer to models trained with gamma-based illumination augmentation and tested on images with no change or with gamma augmentation, respectively.

The substantial performance recovery observed when models are trained with gamma-augmented data can be attributed to the enhanced diversity in illumination conditions present during training, which enables the DL models to learn more invariant and generalizable feature representations. Architectures such as InceptionV3, which initially ex-

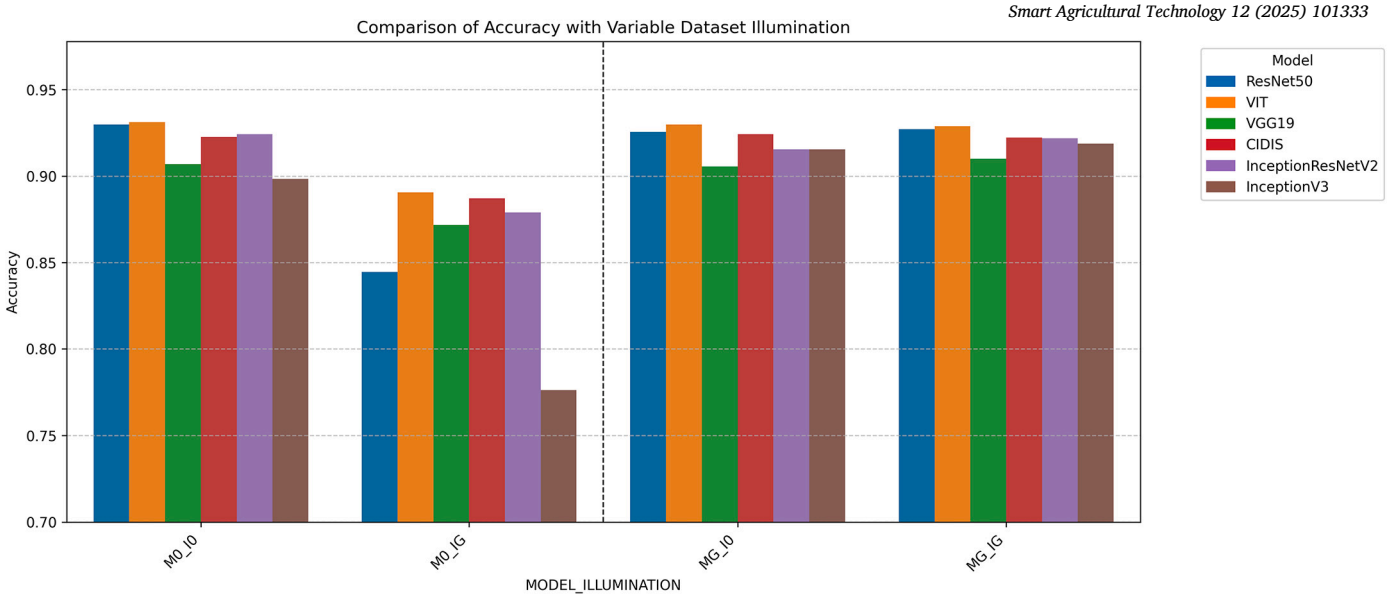

Fig. 8. Accuracy of the six DL models (ResNet50, ViT, VGG19, CIDIS, InceptionResNetV2, and InceptionV3) evaluated on the four illumination conditions used in this study. M0_I0 and M0_IG correspond to models trained without illumination augmentation, whereas MG_I0 and MG_IG denote models trained with gamma-augmented data ($\Gamma = 0.0$–0.7). The dashed vertical line separates the models trained on non-augmented data (left) from those trained with illumination-augmented data (right).

hibited the highest sensitivity to illumination changes, benefit the most from this augmentation, as the exposure to a broader range of lighting conditions mitigates their tendency to overfit to specific brightness and contrast patterns. Similarly, ResNet50, CIDIS, and InceptionResNetV2 show significant improvements, suggesting that their feature extraction mechanisms can adapt effectively when provided with sufficiently varied training data. The relatively smaller gain observed in ViT, despite its strong baseline performance, may indicate that its attention-based mechanism already captures robust global features that are less affected by local lighting variations.

Fig. 8 provides a visual counterpart to Table 5, showing the same accuracy values as grouped bar charts.

The left block (M0) clearly evidences the vulnerability of all models when exposed to illumination conditions not seen during training. The observed performance drops highlight a consistent lack of robustness, particularly in architectures that appear to overfit to the original lighting distribution. Although some models exhibit a relatively milder decline, none are immune to the effects of gamma-based augmentation, underscoring a fundamental limitation in their ability to generalize under realistic lighting variability. These findings stress the importance of incorporating illumination diversity during training to mitigate sensitivity to domain shifts.

The right block (MG) demonstrates the clear benefits of training with illumination-augmented data. All architectures show a notable improvement in robustness under challenging lighting conditions, while maintaining strong performance under standard settings. This consistency suggests that the models effectively internalize illumination-invariant features when exposed to a broader range of visual conditions during training. Notably, the relative differences between models narrow, indicating that data augmentation helps level the playing field and mitigates the vulnerabilities previously observed in more sensitive architectures.

Together with the numerical evidence, the figure makes it visually explicit that illumination augmentation converts the pronounced accuracy drop of the M0 models into a negligible gap, underscoring the need to use diverse lighting data when deploying classifiers of banana ripeness in real-world settings.

Table 6 summarizes the average inference time per image (in milliseconds) and the model complexity (in number of parameters, millions) for each architecture on the IG test set. For every model, two rows are presented: M0_IG corresponds to training on the baseline dataset without illumination augmentation, while MG_IG denotes the same architecture retrained with gamma-augmented data.

Table 6
Average inference time per image on the IG test set and model complexity (in number of parameters, millions). Each block compares the model trained without illumination augmentation (M0_IG) versus the same model trained with gamma-augmented data (MG_IG).

Illumination	Model	Avg. inference (ms)	# Params (M)
M0_IG	ViT	3.266	10.7
MG_IG		**2.083**	
M0_IG	VGG19	2.535	21.07
MG_IG		**2.520**	
M0_IG	CIDIS	2.689	1.9
MG_IG		**2.664**	
M0_IG	InceptionResNetV2	2.678	56.43
MG_IG		**2.748**	
M0_IG	ResNet50	3.504	26.21
MG_IG		**3.383**	
M0_IG	InceptionV3	3.891	24.42
MG_IG		**3.790**	

Note: M0_IG refers to the average inference time for models trained on non-augmented data and tested on gamma-augmented images (IG). MG_IG refers to the average inference time for models trained with gamma-based illumination augmentation and tested on gamma-augmented images. "# Params (M)" indicates the total number of trainable parameters in millions.

The combined analysis of inference time and model complexity indicates that the introduction of illumination augmentation does not substantially compromise computational efficiency across the evaluated architectures. ViT exhibits the most noticeable improvement, highlighting the potential of attention-based models to enhance processing efficiency when exposed to variability in visual conditions. ResNet50 and InceptionV3 demonstrate minor improvements, suggesting that deeper convolutional architectures benefit modestly from data augmentation without incurring additional computational costs. Conversely, CIDIS and VGG19 present negligible variations in inference time, implying an intrinsic robustness to illumination changes but also suggesting limited capacity for further optimization through augmentation. Notably, InceptionResNetV2 experiences a slight increase in inference time, which may be attributed to the higher computational demands associated with

Table 7
Comparative summary of model complexity, hyperparameters, inference time, transfer-learning accuracy (*Acc (reported)*) and illumination-variation performance for banana-ripeness classifiers.

Model	#Params (M)	Optimizer	Dropout	LR	Batch	Inf M0_IG (ms)	Inf MG_IG (ms)	Acc (reported) [17]	Acc M0_IG	Acc MG_IG
VGG19	21.07	Adagrad	0.2	1e-3	64	2.54	2.52	0.562	0.8718	0.9101
ResNet50	26.21	Adagrad	0.2	1e-3	64	3.50	3.38	0.816	0.8446	0.9272
InceptionResNetV2	56.43	Adagrad	0.2	1e-3	64	2.68	2.75	0.869	0.8791	0.9219
InceptionV3	24.42	Adagrad	0.2	1e-3	64	3.89	3.79	0.849	0.7763	0.9188
CIDIS (proposed)	1.90	Adagrad	0.2	1e-3	50	2.69	2.66	0.917	0.8872	0.9224
ViT	10.70	AdamW	0.1	1e-4	64	3.27	2.08	–	0.8907	0.9290

its hybrid architecture, where the integration of inception modules and residual connections intensifies the complexity of feature extraction under augmented conditions.

Building on these observations, Table 7 provides a consolidated comparison of each architecture's key characteristics and performance trade-offs. It details the number of trainable parameters (in millions), the transfer-learning hyperparameters (optimizer, learning rate, dropout rate, and batch size), and the average inference times under non-augmented (M0_IG) and illumination-augmented (MG_IG) conditions. The column "Acc (reported)" reports the best transfer-learning accuracies on real data from Chuquimarca et al. [17], while "Acc M0_IG" and "Acc MG_IG" display test accuracies without and with synthetic lighting augmentation, respectively. This integrated view highlights how model complexity, computational efficiency, and predictive robustness interact to inform the selection of the most suitable DL architecture for practical, real-time post-harvest deployments.

ViT trained with illumination augmentation (MG_IG) achieves over 92.9% accuracy with an average inference time of 2.08 ms per image, demonstrating the most effective balance between robustness and computational efficiency for real-time banana ripeness grading under variable lighting conditions.

LIME surpasses traditional preprocessing methods such as CLAHE, Retinex, Color Jitter, white balance correction and HSV channel extraction by adaptively enhancing illumination without generating artifacts or distorting the critical green to yellow chromatic balance. CLAHE can introduce local over enhancement while Retinex and Color Jitter may shift hue distributions. LIME does not require paired training data and it avoids oversaturation of vital channels. White balance correction addresses only global tone shifts and HSV extraction sacrifices luminance cues. In contrast, LIME preserves original color fidelity and produces realistic brightness variations in a controlled manner, resulting in stable illumination invariant features that strengthen ripeness classification without compromising the essential chromatic information.

4. Conclusion

Training banana-ripeness classifiers on pristine images results in high baseline accuracy but reveals a significant flaw: the models exhibit a sharp decline in performance when exposed to gamma-altered inputs. This highlights a critical overfitting issue, where the models are overly reliant on clean data and fail to generalize well under real-world conditions. The stark performance drops when exposed to variable lighting suggest that these architectures are not inherently robust to illumination changes, which is a major limitation for practical deployment in automated grading tasks.

The augmentation strategy using gamma-based transformations significantly improves robustness, with all models, regaining and even exceeding their original accuracy. However, the improvements vary across architectures, with some models like InceptionV3 showing the largest relative gains, yet still struggling with poor performance for challenging lighting conditions. ViT, ResNet50, and CIDIS do show solid improvements, but their underlying architectures seem to remain somewhat ill-suited for dealing with lighting variability. Their gains, although notable, do not completely resolve their vulnerability to illumination changes, indicating a fundamental weakness in their design for this particular task.

Furthermore, although the models exhibit improved robustness without a substantial compromise in inference time, the trade-off between computational efficiency and predictive performance remains nontrivial. ViT and ResNet50 achieve a favorable balance between accuracy and processing speed; however, their elevated computational demands could limit their deployment in real-time or resource-constrained environments. While the benefits of illumination augmentation are evident, they are insufficient to fully overcome the models' vulnerabilities under extreme or unforeseen lighting conditions. Thus, despite advances in robustness, the long-term reliability of these architectures in dynamic, real-world scenarios remains an open challenge.

The ranking shift in models when tested under a wider set of illumination conditions further emphasizes the limitations of traditional evaluation methods. CIDIS and InceptionResNetV2, initially poor performers, show improvement under augmented training, which challenges the idea that performance on clean data should be the primary criterion for selecting models for deployment. It underscores the need for evaluations that reflect the complexities of real-world environments, where lighting variability plays a pivotal role in performance.

CRediT authorship contribution statement

Luis E. Chuquimarca: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Boris X. Vintimilla:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sergio A. Velastin:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was partially supported by the ESPOL-CIDIS-11-2022 project. Additional support from the INCYT-PNF-2023-S76-175 project is also gratefully acknowledged.

Data availability

Data will be made available on request.

References

[1] J. Naranjo-Torres, M. Mora, R. Hernández-García, R.J. Barrientos, C. Fredes, A. Valenzuela, A review of convolutional neural network applied to fruit image processing, Appl. Sci. 10 (2020) 3443.

[2] E.M. Kikulwe, S. Okurut, S. Ajambo, K. Nowakunda, D. Stoian, D. Naziri, Postharvest losses and their determinants: a challenge to creating a sustainable cooking banana value chain in Uganda, Sustainability 10 (2018) 2381.

[3] R. Kanjilal, J.E. Saenz, I. Uysal, Large-scale data-driven uniformity analysis and sensory prediction of commercial banana ripening process, Postharvest Biol. Technol. 219 (2025) 113203.

[4] H. Huang, G. Jing, H. Wang, X. Duan, H. Qu, Y. Jiang, The combined effects of phenylurea and gibberellins on quality maintenance and shelf life extension of banana fruit during storage, Sci. Hortic. 167 (2014) 36–42.

[5] P.M. Toivonen, E.J. Mitcham, L.A. Terry, Postharvest care and the treatment of fruits and vegetables, in: Horticulture: Plants for People and Places, vol. 1: Production Horticulture, 2014, pp. 465–483.

[6] M. Al-Dairi, P.B. Pathare, R. Al-Yahyai, H. Jayasuriya, Z. Al-Attabi, Postharvest quality, technologies, and strategies to reduce losses along the supply chain of banana: a review, Trends Food Sci. Technol. 134 (2023) 177–191.

[7] U.L. Opara, I.O. Ogra, An introduction to postharvest handling technology of fresh fruits and vegetables, in: Sustainable Postharvest Technologies for Fruits and Vegetables, CRC Press, 2025, pp. 3–41.

[8] N.R. Giuggioli, S. Ollani, R. Zanchini, B. Danielle, A. Sparacino, S. Massaglia, V.M. Merlino, The appeal of bananas: a qualitative sensory analysis and consumers' insights into tropical fruit consumption in Italy, J. Agric. Food Res. 16 (2024) 101110.

[9] M. Rajini, V. Persis, Developing an iot and ml-driven platform for fruit ripeness evaluation and spoilage detection: a case study on bananas, e-Prime - Adv. Electr. Eng. Electron. Energy 100896 (2025).

[10] L. Chuquimarca, B. Vintimilla, S. Velastin, Classifying healthy and defective fruits with a multi-input architecture and cnn models, in: 2024 14th International Conference on Pattern Recognition Systems (ICPRS), IEEE, 2024, pp. 1–7.

[11] O. Coello, M. Coronel, D. Carpio, B. Vintimilla, L. Chuquimarca, Enhancing apple's defect classification: insights from visible spectrum and narrow spectral band imaging, in: 2024 14th International Conference on Pattern Recognition Systems (ICPRS), IEEE, 2024, pp. 1–6.

[12] M. Kondoyanni, D. Loukatos, C. Templalexis, D. Lentzou, G. Xanthopoulos, K.G. Arvanitis, Computer vision in monitoring fruit browning: neural networks vs. stochastic modelling, Sensors 25 (2025) 2482.

[13] A. Sanaeifar, A. Bakhshipour, M. De La Guardia, Prediction of banana quality indices from color features using support vector regression, Talanta 148 (2016) 54–61.

[14] C. Kanimalar, M. Karthikeyan, Advances in machine learning and deep learning for automated banana disease detection: a comprehensive survey, in: 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), IEEE, 2024, pp. 1160–1168.

[15] N. Saranya, K. Srinivasan, S.P. Kumar, Banana ripeness stage identification: a deep learning approach, J. Ambient Intell. Humaniz. Comput. 13 (2022) 4033–4039.

[16] F.M. Mazen, A.A. Nashat, Ripeness classification of bananas using an artificial neural network, Arab. J. Sci. Eng. 44 (2019) 6901–6910.

[17] L.E. Chuquimarca, B.X. Vintimilla, S.A. Velastin, Banana ripeness level classification using a simple cnn model trained with real and synthetic datasets, in: VISIGRAPP (5: VISAPP), 2023, pp. 536–543.

[18] M. Knott, F. Perez-Cruz, T. Defraeye, Facilitated machine learning for image-based fruit quality assessment, J. Food Eng. 345 (2023) 111401.

[19] P. Arunima, P.P. Gopinath, P.G. Lekshmi, M. Esakkimuthu, Digital assessment of post-harvest nendran banana for faster grading: cnn-based ripeness classification model, Postharvest Biol. Technol. 214 (2024) 112972.

[20] C. Chang, R. Parthiban, V. Kalavally, Y.M. Hung, X. Wang, Unharvested palm fruit bunch ripeness detection with hybrid color correction, Smart Agric. Technol. 9 (2024) 100643.

[21] S. Nuanmeesri, Enhanced hybrid attention deep learning for avocado ripeness classification on resource constrained devices, Sci. Rep. 15 (2025) 3719.

[22] C. Guo, C. Li, J. Guo, C.C. Loy, J. Hou, S. Kwong, R. Cong, Zero-reference deep curve estimation for low-light image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780–1789.

[23] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, Enlightengan: deep light enhancement without paired supervision, IEEE Trans. Image Process. 30 (2021) 2340–2349.

[24] G.F. Hong, S. Nair, C.Y. Lin, C.S. Kuan, S.J. Chen, Deep learning-based detection of green-ripe pineapples via bract wilting rate analysis, Smart Agric. Technol. 100949 (2025).

[25] X. Guo, Y. Li, H. Ling, Lime: low-light image enhancement via illumination map estimation, IEEE Trans. Image Process. 26 (2016) 982–993.

[26] J. Wang, W. Tan, X. Niu, B. Yan, Rdgan: retinex decomposition based adversarial learning for low-light enhancement, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 1186–1191.

[27] Y. Wang, R. Wan, W. Yang, H. Li, L.P. Chau, A. Kot, Low-light image enhancement with normalizing flow, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2604–2612.

[28] L. Ma, T. Ma, R. Liu, X. Fan, Z. Luo, Toward fast, flexible, and robust low-light image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5637–5646.

[29] M. De-Arteaga, W. Herlands, D.B. Neill, A. Dubrawski, Machine learning for the developing world, ACM Trans. Manag. Inf. Syst. 9 (2018) 1–14.

[30] P. Baglat, A. Hayat, F. Mendonca, A. Gupta, S.S. Mostafa, F. Morgado-Dias, Non-destructive banana ripeness detection using shallow and deep learning: a systematic review, Sensors 23 (2023) 738.

[31] Y.A. Ramadhan, E.C. Djamal, F. Kasyidi, A.T. Bon, Identification of cavendish banana maturity using convolutional neural networks, in: Proceedings of the International Conference on Industrial Engineering and Operations Management, United Arab Emirates, Dubai, 2020, pp. 10–12.

[32] M. Ivanovs, K. Ozols, A. Dobrajs, R. Kadikis, Improving semantic segmentation of urban scenes for self-driving cars with synthetic images, Sensors 22 (2022) 2252.

[33] S. Zhao, S. Hao, B. Zi, H. Xu, K.Y.K. Wong, Bridging different language models and generative vision models for text-to-image generation, in: European Conference on Computer Vision, Springer, 2024, pp. 70–86.

[34] R. Pacheco, P. González, L.E. Chuquimarca, B.X. Vintimilla, S.A. Velastin, Fruit defect detection using cnn models with real and virtual data, in: VISIGRAPP (4: VISAPP), 2023, pp. 272–279.

[35] T.D. Beltran, R.J. Villao, L.E. Chuquimarca, B.X. Vintimilla, S.A. Velastin, Fruit deformity classification through single-input and multi-input architectures based on cnn models using real and synthetic images, in: Iberoamerican Congress on Pattern Recognition, Springer, 2024, pp. 46–62.

[36] W. Wang, X. Wu, X. Yuan, Z. Gao, An experiment-based review of low-light image enhancement methods, IEEE Access 8 (2020) 87884–87917.

[37] C. Li, C. Guo, L. Han, J. Jiang, M.M. Cheng, J. Gu, C.C. Loy, Low-light image and video enhancement using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 9396–9416.

[38] S. Yang, D. Zhou, J. Cao, Y. Guo, Lightingnet: an integrated learning method for low-light image enhancement, IEEE Trans. Comput. Imaging 9 (2023) 29–42.

[39] F. Lv, Y. Li, F. Lu, Attention guided low-light image enhancement with a large scale low-light simulation dataset, Int. J. Comput. Vis. 129 (2021) 2175–2193.

[40] W. Kim, Low-light image enhancement: a comparative review and prospects, IEEE Access 10 (2022) 84535–84557.

[41] L. Yang, B. Cui, J. Wu, X. Xiao, Y. Luo, Q. Peng, Y. Zhang, Automatic detection of banana maturity—application of image recognition in agricultural production, Processes 12 (2024) 799.

[42] L.E. Chuquimarca, B.X. Vintimilla, S.A. Velastin, A review of external quality inspection for fruit grading using cnn models, Artif. Intell. Agric. (2024).

[43] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. 54 (2022) 1–41.

[44] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 87–110.

[45] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open (2022).

[46] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, S.J. Oh, Rethinking spatial dimensions of vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11936–11945.

[47] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, A. Veit, Understanding robustness of transformers for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10231–10241.

[48] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[50] M.A.E. Zeid, K. El-Bahnasy, S. Abo-Youssef, Multiclass colorectal cancer histology images classification using vision transformers, in: 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, 2021, pp. 224–230.

[51] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (2020) 43–76.

[52] S.K. Behera, A.K. Rath, P.K. Sethy, Maturity status classification of papaya fruits based on machine learning and transfer learning approach, Inf. Process. Agric. 8 (2021) 244–250.

[53] A. Helwan, M.K. Sallam Ma'aitah, R.H. Abiyev, S. Uzelaltinbulat, B. Sonyel, Deep learning based on residual networks for automatic sorting of bananas, J. Food Qual. 2021 (2021).

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[56] V. Kurama, A review of popular deep learning architectures: resnet, inceptionv3, and squeezenet, Consult. (August 2020).

[57] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (2019) 1–54.

[58] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.

[59] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (2009) 427–437.

[60] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263–1284.

[61] R. Padilla, W.L. Passos, T.L. Dias, S.L. Netto, E.A. Da Silva, A comparative analysis of object detection metrics with a companion open-source toolkit, Electronics 10 (2021) 279.

[62] M. Palumbo, M. Cefola, B. Pace, G. Attolico, G. Colelli, Computer vision system based on conventional imaging for non-destructively evaluating quality attributes in fresh and packaged fruit and vegetables, Postharvest Biol. Technol. 200 (2023) 112332.

[63] M. Tan, Q. Le, Efficientnetv2: smaller models and faster training, in: International Conference on Machine Learning, PMLR, 2021, pp. 10096–10106.

[64] W. Avery, M. Munir, R. Marculescu, Scaling graph convolutions for mobile vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5857–5865.