OENO One
vine & wine

DOI: https://doi.org/10.20870/oeno-one.2025.59.2.9306

OPEN ACCESS

SPECIAL ISSUE

Artificial intelligence-driven classification method of grapevine major phenological stages using conventional RGB imaging

Rubén Íñiguez[2,3], Fikile Wolela[1], María Ignacia Gonzalez Pavez[1,4], Ignacio Barrio[2,3], Javier Tardáguila[2,3], Talitha Venter[1] and Carlos Poblete-Echeverría[1,*]

[1] South African Grape and Wine Research Institute (SAGWRI), Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
[2] Televitis Research Group, University of La Rioja, 26006 Logroño, Spain
[3] Affiliation Institute of Grapevine and Wine Sciences (University of La Rioja, Consejo Superior de Investigaciones Científicas, Gobierno de La Rioja), 26007 Logroño, Spain
[4] Research and Extension Center for Irrigation and Agroclimatology (CITRA), Faculty of Agricultural Sciences, Universidad de Talca, Campus Talca, Chile

This article is an original research article published in cooperation with the 23rd GiESCO International Conference, July 21-27, 2025, hosted by the Hochschule Geisenheim University in Geisenheim, Germany.

Guest editors: Laurent Torregrosa and Susanne Tittmann.

Article number: 9306

*correspondence:
cpe@sun.ac.za

Associate editor:
Laurent Jean-Marie Torregrosa

Received:
15 March 2025

Accepted:
23 May 2025

Published:
16 June 2025

This article is published under the **Creative Commons licence** (CC BY 4.0).

Use of all or part of the content of this article must mention the authors, the year of publication, the title, the name of the journal, the volume, the pages and the DOI in compliance with the information given above.

ABSTRACT

Accurate monitoring of grapevine phenological stages is essential for optimising vineyard management. This study evaluates the performance of three deep learning architectures (ResNet-34, YOLOv11-Classification and Vision Transformer (ViT)) for automated classification of vineyard canopy images into four key phenological stages: i) Shoot and inflorescence development (E-L 12–18), ii) Flowering (E-L 19–26), iii) Berry formation (E-L 27–33), and iv) Berry ripening (E-L 35–38). These categories correspond to broad developmental periods that may span several E-L stages. A dataset comprising 4,381 images was used to train and validate the models, incorporating data augmentation techniques to improve robustness. Results indicate that all three models achieved high classification accuracy, with ResNet-34 obtaining the highest accuracy (97.4 % validation, 95.6 % test), reinforcing its strong feature extraction capabilities. However, its lower F1-score (95.3 % validation, 91.8 % test) suggests challenges in handling class imbalances. YOLOv11-Classification demonstrated the most balanced classification performance, achieving a high F1-score (93.6 % validation, 91.8 % test) while maintaining the fastest training time, making it particularly suitable for real-time applications. ViT exhibited competitive classification performance but had higher computational demands, limiting its feasibility for real-time vineyard monitoring. A confusion matrix analysis highlighted misclassification trends, particularly between early shoot development and flowering, due to their visual similarities. Despite these challenges, the study confirms that AI models can effectively automate vineyard phenology classification, reducing manual assessment efforts, and contributing to more efficient viticultural decision-making.

KEYWORDS: deep learning, vineyard phenology, YOLOv11, ResNet-34, Vision Transformer (ViT), precision viticulture, automated classification, GiESCO 2025

INTRODUCTION

The phenological stage of the grapevine (*Vitis vinifera* L.) represents a fundamental element in vineyard management, since it determines key practices such as fertilisation, irrigation, phytosanitary interventions, and optimal harvest time (Mullins *et al.,* 1992). Phenology can be understood as the study of the phenological events, or the stages of plant development that occur during their active lifecycle, in response to climatic conditions. The phenological development of grapevines is mainly influenced by climatic elements, such as temperature, solar radiation, and precipitation, which also influence the production and the quality of grape berries. Temperature is the main forcing element in the phenology of grapevines, in this sense, the projected increases in temperature under future likely climate change scenarios may lead to the advancement of 6 to 25 days for different grapevine varieties in Mediterranean climate regions (Reis *et al.,* 2020).

To standardise the description of these stages, several phenological scales have been developed, including the Baggiolini scale, the Eichhorn-Lorenz (E-L) scale, and the extended BBCH scale. The Baggiolini scale, initially used for planning pesticide applications, was limited in scope as it only covered early development stages. In contrast, the E-L scale introduced 47 numerical codes describing 22 phenological stages from winter bud to leaf fall, providing greater detail and flexibility to incorporate sub-stages (Coombe, 1995). The extended BBCH scale increased precision by detailing phenological macro and micro-stages, facilitating its application across multiple crops and standardising its use internationally (Lorenz *et al.,* 1995).

Despite advances in these descriptive tools, phenological identification traditionally relies on manual observations by an experienced technical person. This method is time-intensive, subjective, and often unable to capture spatial variability within the vineyard, which can lead to suboptimal decisions being made (Verdugo-Vásquez *et al.,* 2018). Although reference scales such as the E-L and BBCH systems facilitate identification, practical implementation across large and heterogeneous vineyard blocks still presents logistical and consistency challenges. In addition, climatic variations, edaphic conditions, and agronomic practices can further complicate correct identification (Altimiras *et al.,* 2024).

In this context, emerging technologies, such as computer vision and deep learning, offer a promising solution. These tools provide a possibility to automatise the classification and monitoring of phenological stages by capturing and analysing large-scale areas covered by multiple images. Algorithms based on Convolutional Neural Networks (CNNs) have proven to be effective for image classification techniques in agriculture, with accuracies exceeding 88 % in applications such as grapevine phenological stage identification (Schieck *et al.,* 2023). In addition, these technologies have also been successfully used in the detection of grape bunches under various occlusion conditions (Íñiguez *et al.,* 2024) and in the assessment of diseases such as downy mildew (Hernández *et al.,* 2021). Systems that integrate proximity sensors and Internet of Things (IoT) platforms are integrating the capacity for continuous and real-time monitoring in vineyards, reducing costs and improving decision making (Mendes *et al.,* 2022).

The objective of this study is to develop an Artificial Intelligence-driven classification method of grapevine phenology using conventional RGB imaging. The aim is to train AI models capable of accurately identifying major grapevine phenological stages through image analysis under real field conditions. The practical objective is to implement a system that provides speed, objectivity, and scalability, addressing the limitations of traditional methods based on manual observations.

MATERIALS AND METHODS

1. Description of experimental sites

The experiment was conducted in two vineyards located in different countries and hemispheres: Spain (Site SP) and South Africa (Site SA).

Site SP: in Spain, the study was carried out at the University of La Rioja's experimental vineyard in Logroño (42° 27' 42.9" N 2° 25' 40.2" W). Data were collected from Tempranillo vines trained to two systems: a Vertical Shoot Positioning (VSP) system with double cordon Royat (Figure 1a) and a free-cordon system with simple cordon Royat (Figure 1b). The vineyard was planted in 2010, grafted on Richter 110 rootstock, and equipped with a drip irrigation system. The cordon height was approximately 0.6 m. The vineyard is situated at an elevation of 384 m above sea level, with a north-south row orientation, and vine spacing of 2.8 × 1.1 m. A total of 80 vines were selected across eight rows (four per training system), with 10 consecutive vines per row, and images were collected from the same individual vines throughout the 2024 growing season.

Site SA: in South Africa, the study was conducted at Stellenbosch University's Welgevallen Experimental Farm, Stellenbosch (33° 56' 26" S 18° 51' 56" E). Data were collected from Cabernet-Sauvignon (Figure 1c) and Chenin Blanc (Figure 1d) vines both trained to a Vertical Shoot Positioning (VSP) system. The vineyard, established in 2020, is grafted on Richter 110 rootstock and managed under a drip irrigation system. The cordon height was approximately 0.6 m. The vineyard is situated at 157 m above sea level with a north-south row orientation and vine spacing of 2.7 × 1.5 m. A total of 240 vines were selected across four rows (two per variety), with 60 consecutive vines per row, and images were collected from the same individual vines during the 2024-2025 growing season.

2. Image acquisition

Image acquisition was conducted during the 2024 growing season in two vineyard sites: La Rioja (Spain) and Stellenbosch (South Africa). In the Spanish site, images were collected from 25 March to 23 August 2024, covering

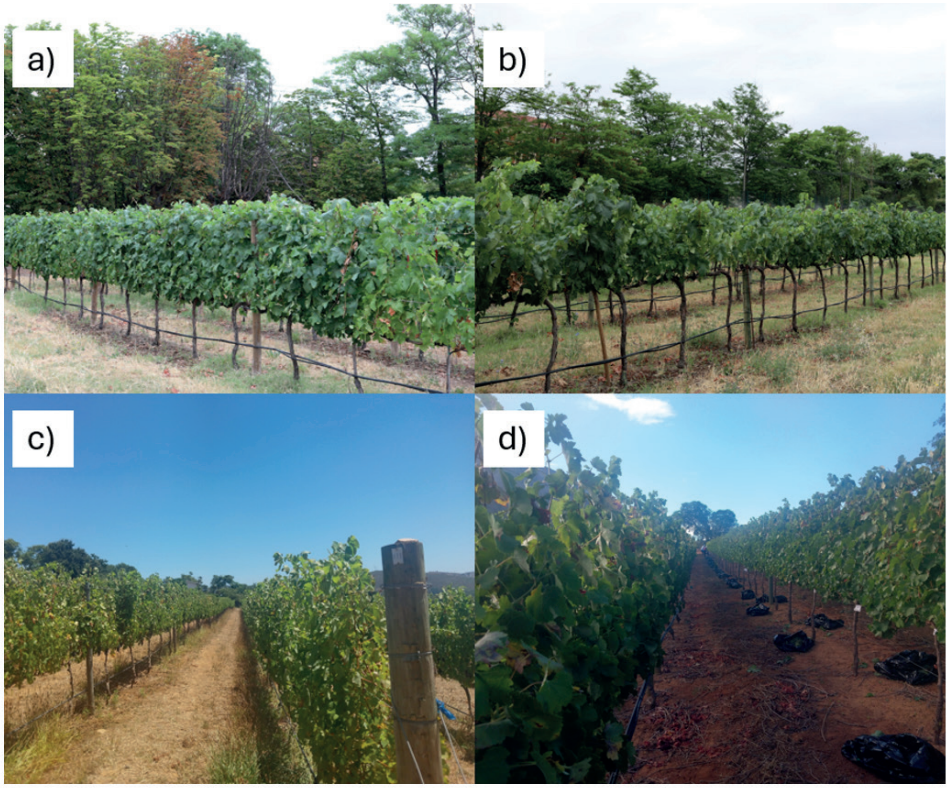


FIGURE 1. Experimental vineyard sites used for image acquisition. (a) Tempranillo trained in a Vertical Shoot Positioning (VSP) system in Spain. (b) Tempranillo trained in a free-cordon system in Spain. (c) Cabernet-Sauvignon trained in a VSP system in South Africa. (d) Chenin Blanc trained in a VSP system in South Africa.

the stages from early shoot development to the onset of ripening. In the South African site, image acquisition took place from 15 November to 27 December 2024, capturing a comparable phenological range in the southern hemisphere.

Images were captured one to two times per week, depending on the vine growth rate and field conditions, ensuring sufficient temporal resolution to represent the progression through each phenological period.

Canopy images were captured using conventional digital cameras under natural lighting conditions, reflecting uncontrolled ambient environments. In SP site the images were taken with a Digital Single Lens Reflex (DSLR) RGB camera (Canon EOS 5D Mark IV, Canon Inc., Tokyo, Japan) featuring a full-frame CMOS sensor (30.4 MP) equipped with a Canon EF 20 mm F/2.8 USM lens. The camera was mounted on a tripod, positioned 1.0 m from the row axis, and elevated to 1.2 m above ground level. In the SA site, the images were primarily captured using a compact digital camera: a Sony Cyber-shot DSC-W800 (Sony Corporation, Tokyo, Japan). A Canon PowerShot ELPH 160 (Canon Inc., Tokyo, Japan) with similar specifications was used as a backup device in case of battery depletion or technical contingencies, ensuring continuity during field acquisition. Both cameras feature 20-megapixel CCD sensors and optical zoom capabilities, and all images were taken under natural lighting conditions without the use of artificial illumination. Images were taken manually without a tripod, maintaining a consistent distance of 1.5 m from the canopy, and an elevation of 1.3 m above ground level. No artificial lighting was applied in either location.

3. Reference dataset

The dataset was pre-processed by manually classifying canopy images based on the E-L major stages described by Coombe (1995). This activity was done by personnel with training in viticulture. The images were classified and organised into four phenological stages:

i) Shoot and Inflorescence Development (1_INF): E-L stages 12 to 18, encompassing early shoot elongation and inflorescence development (2,099 images).

ii) Flowering (2_FLO): E-L stages 19 to 26, covering the beginning of flowering to cap-fall completion (349 images).

iii) Berry Formation (3_BER): E-L stages 27 to 33, corresponding to early berry development and bunch closure (1,695 images).

iv) Berry Ripening (4_RIP): E-L stages 35 to 38, depicting the ripening process of grape berries (238 images).

The fifth E-L stage, senescence, was excluded since no post-harvest images were taken. Abbreviations (1_INF, 2_FLO, 3_BER, 4_RIP) were used to name the image folders and are also used in the figures for consistency. This classification ensured the dataset was accurately labelled and suitable for training and validation.

4. Dataset partitioning for training, validation, and testing images

The dataset curated for this study comprised a total of 4,381 canopy images, manually classified into four phenological stages based on the E-L major stages. To ensure a robust and unbiased evaluation of the deep learning models, the dataset was systematically divided into three subsets: training, validation, and testing (Table 1).

The training set consisted of 3,061 images, accounting for 70 % of the total dataset, and was used for model learning. The validation set included 867 images (20 % of the dataset), and was used for hyperparameter tuning and performance monitoring during training. The remaining 453 images (10 %) formed the testing set, which was exclusively used to evaluate the final model's performance on unseen data (Table 1).

The partitioning of images was carried out in a stratified manner. Specifically, the 1_INF stage comprised 2,099 images, 2_FLO (349 images), 3_BER (1,695 images), and 4_RIP (238 images). The number of images per phenological stage reflects the natural dynamics of grapevine development. Early shoot and inflorescence development (1_INF), and berry formation (3_BER) are extended phenological phases, allowing for more frequent image acquisition. Conversely, flowering (2_FLO) and berry ripening (4_RIP) are shorter periods, inherently limiting the number of images collected during these stages.

5. Data augmentation for improved model generalisation

To enhance the robustness and generalisation capacity of the dataset, an offline data augmentation strategy, which was common for all the models, was applied prior to training. This ensured that all deep learning architectures evaluated in this study were trained with the same augmented dataset, thereby maintaining consistency across model comparisons. By artificially increasing the diversity of training samples, data augmentation mitigates overfitting and improves the model's ability to generalise to new, unseen vineyard canopy images.

The augmentation process entailed the application of probabilistic transformations to each image, thereby introducing controlled variations while preserving the essential phenological characteristics. The following augmentations were applied, iterating each image with the specified probabilities:

- Random rotation: each image was randomly rotated within a range of –30° to +30° with a probability of 0.5, allowing the model to learn invariance to slight orientation differences in canopy images.
- Horizontal flip: with a probability of 0.5, images were mirrored horizontally, simulating different perspectives of vineyard rows.
- Vertical flip: applied with a probability of 0.3, this transformation introduced further variability in image orientation.
- Colour jitter: adjustments to brightness (±20 %), contrast (±20 %), saturation (±20 %), and hue (±10 %) were randomly introduced with a probability of 0.5, helping the models adapt to varying lighting conditions in real-world vineyard environments.
- Colour inversion: to simulate extreme colour variations, 10 % of images underwent complete colour inversion, exposing the models to unnatural but challenging input conditions.

Figure 2 illustrates the application of various data augmentation techniques to vineyard canopy images. These transformations, including rotation, flipping, colour jittering, and colour inversion, introduce controlled variations that enhance model robustness and generalisation while preserving key phenological features.

In the rare event that no transformations are applied (7.88 % probability), no new images are added to the augmented dataset since no modifications would have been performed on that particular image. To ensure reproducibility, the random seed was fixed, yielding a total of 6,032 training images for this experiment. By precomputing the augmented dataset offline, these transformations were consistently applied across all training procedures, enabling fair comparisons among the evaluated deep learning architectures. The validation and testing datasets were left unchanged, preserving their original characteristics to ensure an unbiased assessment of the performance of all evaluated models.

TABLE 1. Distribution of the images in the datasets (Training, Validation, and Testing) across the four phenological stages (1_INF, 2_FLO, 3_BER, and 4_RIP).

Dataset	1_INF	2_FLO	3_BER	4_RIP
Total images	2,099	349	1,695	238
Training set (70 %)	1,470	240	1,185	166
Validation set (20 %)	416	75	330	46
Testing set (10 %)	213	34	180	26

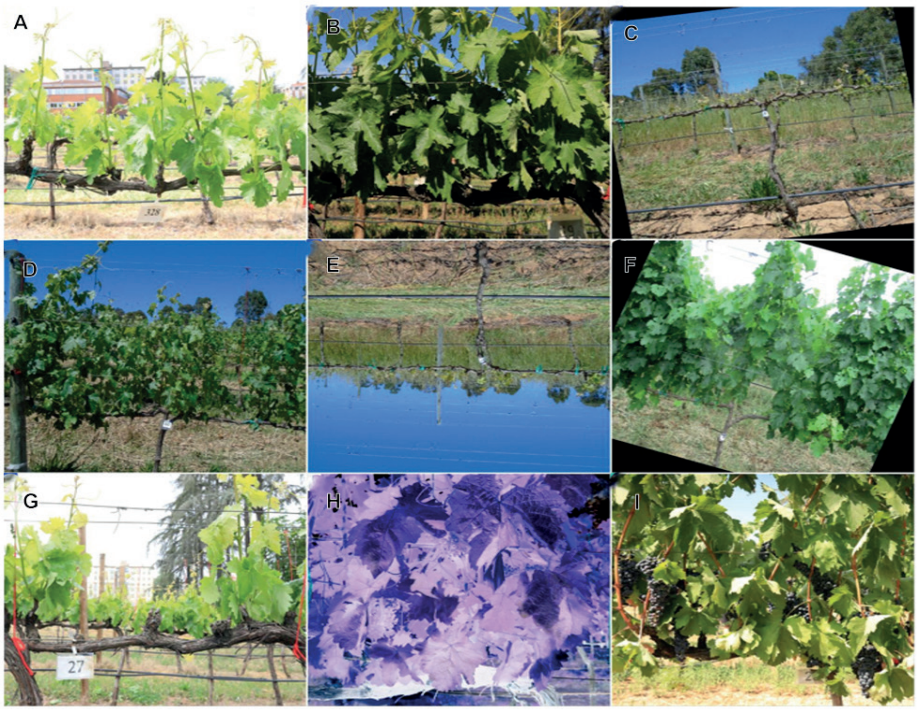


FIGURE 2. Example of data augmentation techniques applied to vineyard canopy images, including rotation, flipping, colour jittering, and colour inversion. The images illustrate different transformations: decreased brightness due to colour jitter (B, D); random rotation within ±30° (C, F); horizontal flip (E); increased brightness from colour jitter (A, G, I), and colour inversion simulating extreme spectral variation (H).

6. Deep learning architecture for image classification

The classification of canopy images into phenological stages was performed using three deep learning architectures: a Convolutional Neural Network (CNN) based on ResNet-34, the YOLOv11 Classification model (YOLOv11-Classification), and a Vision Transformer (ViT). These models were selected for their complementary strengths in image classification tasks, enabling a comprehensive evaluation of different deep learning paradigms for viticultural applications (Silva *et al.,* 2024).

6.1. Convolutional Neural Networks: ResNet-34

Convolutional Neural Networks (CNNs) are among the most widely used deep learning models for image classification due to their ability to extract hierarchical features from images (LeCun *et al.,* 2015). For this study, we implemented ResNet-34, a deep residual network that utilises residual connections to mitigate the vanishing gradient problem and facilitate the training of very deep architectures (He *et al.,* 2016).

ResNet-34 consists of 34 layers with residual skip connections, allowing the network to learn complex patterns in canopy images while maintaining computational efficiency. The model processes images through sequential convolutional layers, extracting both low-level features (*e.g.*, edges, textures) and high-level semantic features (*e.g.*, canopy structure, grape bunches development). These hierarchical features enable robust classification of images into the predefined phenological stages.

6.2. YOLOv11-Classification

The You Only Look Once (YOLO) model is traditionally designed for real-time object detection, but recent adaptations have optimised its architecture for image classification tasks (Redmon *et al.,* 2016). In this study, we employed YOLOv11-Classification (YOLOv11-C), an advanced CNN-based model designed for fast and accurate image categorisation. Unlike traditional CNNs such as ResNet, which process an image through sequential convolutional layers, YOLO treats classification as a single forward pass problem, significantly improving inference speed.

YOLOv11-C divides the input image into a fixed grid, applying convolutional transformations to predict class probabilities across the entire image. This global feature extraction approach enhances the model's ability to recognise subtle variations in canopy structure and phenological indicators, making it well-suited for high-throughput vineyard monitoring applications.

6.3. Vision Transformers (ViT) for image classification

While CNN-based models rely on convolutional operations to extract spatial features, Vision Transformers (ViTs) introduce an alternative paradigm based on self-attention mechanisms (Dosovitskiy *et al.,* 2020). ViTs treat an image as a sequence of non-overlapping patches, applying multi-head self-attention to model long-range dependencies within the image. This approach differs fundamentally from CNNs, as it does not assume local spatial priors, allowing the model to capture global contextual relationships more effectively.

In this study, we implemented the ViT-Base-Patch16-384 model from Google (Hugging Face, 2024), pre-trained on large-scale image datasets and fine-tuned for vineyard canopy classification. This model partitions each input image into a grid of 16 × 16-pixel patches, applies positional embeddings, and processes the sequence through a Transformer encoder. This architecture is conceptually similar to BERT (Bidirectional Encoder Representations from Transformers) in natural language processing, where input text is tokenised into word embeddings and processed through a Transformer encoder to capture long-range dependencies (Devlin *et al.,* 2019). The self-attention mechanism in ViT enables the model to focus on relevant phenological traits, such as leaf morphology, shoot elongation, and berry formation, which are crucial for accurate classification.

ViTs have demonstrated superior performance in complex visual recognition tasks, particularly when trained with large datasets. However, they typically require more computational resources compared to CNN-based architectures such as ResNet and YOLO (Yu *et al.,* 2023).

7. Computational setup and training procedure

The deep learning models were trained on a virtual machine running in a high-performance compute server equipped with an AMD Ryzen Threadripper 3970X 32-Core CPU, 256 GB of ECC DDR4 SDRAM, and several NVIDIA GeForce RTX 4090 (24 GB VRAM) GPUs. The resources allocated for this experiment were 16 CPU cores, 100 GB of RAM and a single RTX 4090 GPU via IOMMU, and PCIe passthrough. This hardware configuration allowed for efficient processing of the dataset and optimisation of deep learning models while minimising computational bottlenecks.

The training process was conducted using PyTorch as the primary deep learning framework, with model implementations adapted from the Torchvision and Ultralytics YOLO libraries. The three architectures (ResNet-34, YOLOv11-C, and ViT) were trained separately using the same dataset partitioning scheme to enable direct performance comparisons.

For all the models, pretrained weights for the respective neural networks were used in order to reduce training times through fine tuning and transfer learning. For ResNet-34 and YOLOv11-C all the weights of the network were trained, for ViT all the layers except for the classification ones were frozen in order to keep time and memory requirements on levels adequate for the available hardware resources.

Each model was trained using stochastic gradient descent (SGD) with AdamW optimisation and an initial learning rate of 0.0001, which was reduced dynamically based on validation performance. The main training hyperparameters were configured as follows:

- Batch size: 32
- Image resolution: 768 × 768 pixels
- Number of epochs (maximum): 50

8. Statistical analysis of the model performance

Model performance was evaluated using Python, with scikit-learn (Pedregosa *et al.,* 2011) and Seaborn (Waskom, 2021) for statistical analysis and visualisation. A normalised confusion matrix was generated for each model during validation and testing, providing a detailed summary of classification accuracy. The confusion matrix compares predicted labels with true labels, identifying correct classifications and misclassifications across the four phenological stages.

To facilitate result interpretation, the confusion matrix was normalised to represent values as percentages of total predictions per class. This approach enabled a clearer assessment of model strengths and weaknesses, highlighting phenological stages that were more prone to misclassification with independency of the number of samples. Additionally, accuracy and macro averaged F1-score were computed for each model to quantify classification performance beyond overall accuracy. The macro averaged F1-score provides a balanced measure of precision and recall without weighting the number of samples per class, making it particularly useful for evaluating models trained on datasets with class imbalances.

The statistical analysis focused on comparing the three architectures based on their ability to correctly classify phenological stages. Differences in accuracy and F1-score were analysed to determine the most suitable model for vineyard phenology classification, considering both classification performance and computational efficiency.

RESULTS AND DISCUSSION

1. Performance evaluation of deep learning models

The performance of the deep learning models was evaluated by means of an analysis of validation and test accuracy, as well as F1-score, since these are the most common model performance metrics for image classification with multiple classes. This approach enabled an assessment to be made of both their learning efficiency and generalisation capability. The validation results offer insights into the extent of adaptation of the models to unseen samples within the training domain, while the test results reflect their performance on completely new data, ensuring robustness for real-world vineyard canopy classification. The comparative performance of the three architectures is summarised in Figure 3 across both dataset splits. This figure provides an intuitive interpretation of the models' classification effectiveness on both the validation and testing datasets. The size of each circle corresponds to the magnitude of the performance metric it represents, with larger circles indicating higher accuracy or F1-score. Similarly, the colour intensity follows a gradient where deeper hues denote superior performance, allowing for a quick comparative assessment of each model's strengths across different evaluation metrics.



FIGURE 3. Comparative performance of deep learning models in vineyard canopy classification across validation and testing, where the size of each circle represents the relative magnitude of accuracy or F1-score: larger circles denote superior performance; while the colour gradient intensity follows a scale in which deeper hues indicate higher values.

The three models exhibited very high classification performance, achieving validation and test accuracies above 95 %. ResNet-34 achieved the highest validation accuracy (97.4 %) and test accuracy (95.6 %), demonstrating strong feature extraction capabilities. However, its F1-score (95.3 % validation, 91.8 % test) suggests that it struggles slightly more with class imbalance, misclassifying underrepresented phenological stages in the same amount as the rest of the tested models in the test dataset.

YOLOv11-C demonstrated the best balance between accuracy and class representation, with validation and test accuracies of 96.2 % and 95.4 %, respectively, and the highest F1-score (93.6 % validation, 91.8 % test). This confirms its robustness in handling class imbalances while maintaining high precision and recall. ViT achieved a comparable validation accuracy of 95.7 % and test accuracy of 95.8 %, with an F1-score of 92.4 % validation and 91.9 % test, but at a higher computational cost.

In terms of computational efficiency in the training phase, YOLOv11-C was the fastest model, completing training in 56 minutes, making it the best model for scenarios with limited computational resources. ResNet-34 required 1 hour and 19 minutes, striking a balance between training speed and accuracy. ViT had the longest training time,1 hour and 33 minutes, despite only updating the classification layer weights as opposed to a complete fine tuning, which may limit its deployment in resource-constrained environments. However, this limitation primarily affects training efficiency and is less relevant during inference.

One common limitation observed across all models was the misclassification of the Flowering (2_FLO) stage, which was confused with adjacent phenological stages more often than any of the remaining combinations. This issue likely stems from visual similarities between flowering and early shoot development, as well as some imbalance between classes in the dataset resulting in a considerable underrepresentation of categories 2_FLO and 4_RIP. This is not a problem for the latter since any visible part of a bunch may settle the classification very easily for the model. Despite this limitation, the accuracy was adequate also in the class corresponding to the stage 2_FLO and the three models effectively automated the phenological classification process, demonstrating their potential for precision viticulture.

ViT offers the highest classification accuracy on the test dataset, making it a strong candidate for offline phenology monitoring where no compute power or time restriction exists, but the differences between all the models are very limited in this scenario, making any of the models almost equally suitable for this task. YOLOv11-C stands out as the best all-around option, given its balance between accuracy, F1-score, and fast training time, making it ideal for real-time applications. ViT, while achieving competitive results, remains computationally demanding, suggesting that hybrid CNN-Transformer approaches may be a promising direction for future research.

2. Classification analysis

In order to perform a more thorough analysis of the classification performance of the models, normalised confusion matrices were generated for both the validation and test datasets. These matrices provide a detailed breakdown of the models' predictive performance across the four phenological stages: Shoot and Inflorescence Development (1_INF), Flowering (2_FLO), Berry Formation (3_BER), and Berry Ripening (4_RIP). The subsequent examination of misclassification patterns facilitates the identification of the specific challenges encountered by each architecture in distinguishing between phenological stages.

ResNet-34 (Figure 4) demonstrated strong classification performance, particularly for the 1_INF and 3_BER phenological stages. During validation, the model correctly classified 99 % of 1_INF samples and 99 % of 3_BER samples, indicating that it effectively distinguishes between early shoot development and berry formation. However, 2_FLO exhibited the highest misclassification rate, with 12 % of its samples being incorrectly identified as 1_INF and 7 % as 3_BER. This suggests that ResNet-34 struggles with the subtle visual differences between flowering and adjacent stages.

In the test phase, ResNet-34 maintained a high classification accuracy, correctly predicting 96 % of 1_INF, 98 % of 3_BER, and 100 % of 4_RIP samples. However, 2_FLO remained the most challenging stage, with 12 % of the samples misclassified as 1_INF and 12 % as 3_BER. These results indicate that while ResNet-34 generalises well, distinguishing flowering from early vegetative and fruit development remains a challenge due to phenological similarities and dataset imbalances.

YOLOv11-C (Figure 5) exhibited the highest F1-score among all models during validation, correctly classifying 1_INF (98 %), 3_BER (99 %), and 4_RIP (100 %) with minimal errors. However, 2_FLO remained the most challenging stage, with 14 % of samples being misclassified, primarily as 1_INF (14 %) and a smaller fraction (3 %) as 3_BER. This suggests that while YOLOv11-C excels in most phenological stages, distinguishing between early shoot development and flowering remains a challenge.

On the test set, YOLOv11-C demonstrated remarkable consistency, with a test accuracy of 95 % and an F1-score of 0.92 (Figure 3). The misclassification pattern for 2_FLO persisted, with 16 % of instances being classified as 1_INF, reinforcing the model's difficulty in differentiating these two stages. Nonetheless, the model's overall stability and strong generalisation capabilities confirm its reliability for vineyard monitoring applications.

ViT (Figure 6) achieved strong classification performance, particularly in identifying 1_INF (98 %) and 3_BER (98 %), similar to ResNet-34 and YOLOv11-C. However, 2_FLO exhibited the highest misclassification rate among all models (21 %), often confused with 1_INF (21 %) and 3_BER (8 %). This pattern aligns with known challenges in transformer-based architectures trained on moderate-sized datasets, where CNNs tend to outperform ViTs in structured image classification tasks.

During the test phase, ViT maintained a stable accuracy of 96 % and an F1-score of 0.92 (Figure 3), closely matching YOLOv11-C. However, the misclassification rate for 2_FLO persisted, indicating that the model still struggles with fine-grained distinctions between early shoot growth and flowering stages. These findings suggest that while ViT captures complex visual features effectively, it may require

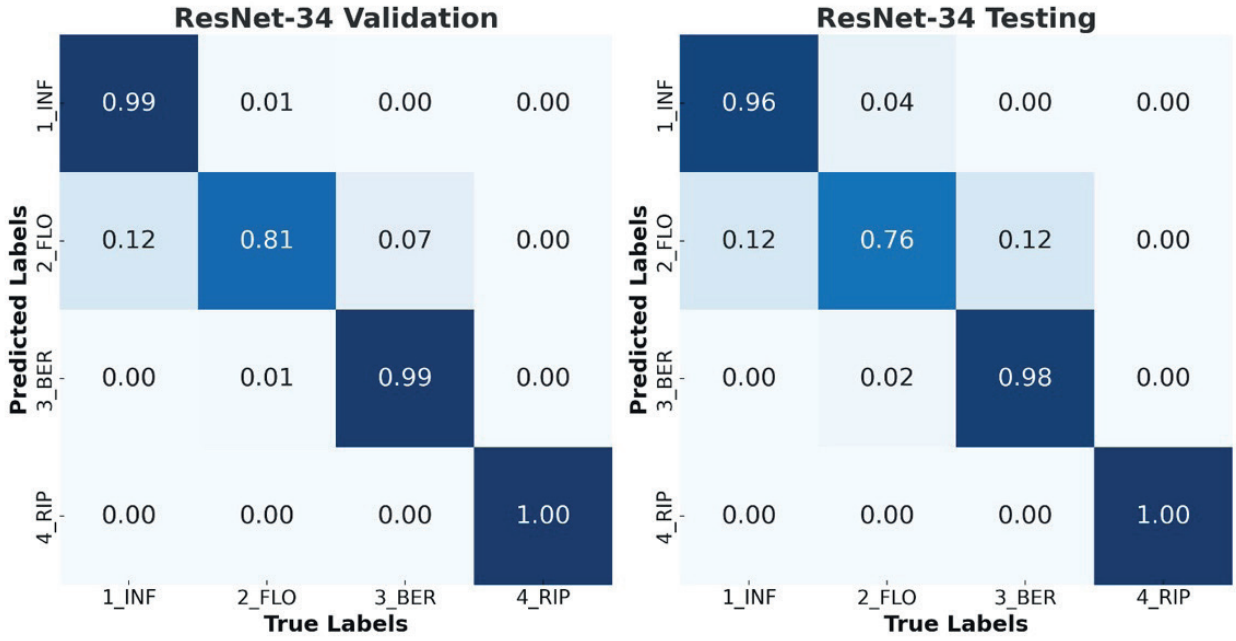


FIGURE 4. Normalised confusion matrices for ResNet-34 during validation (left) and testing (right), showing classification accuracy across the four phenological stages.

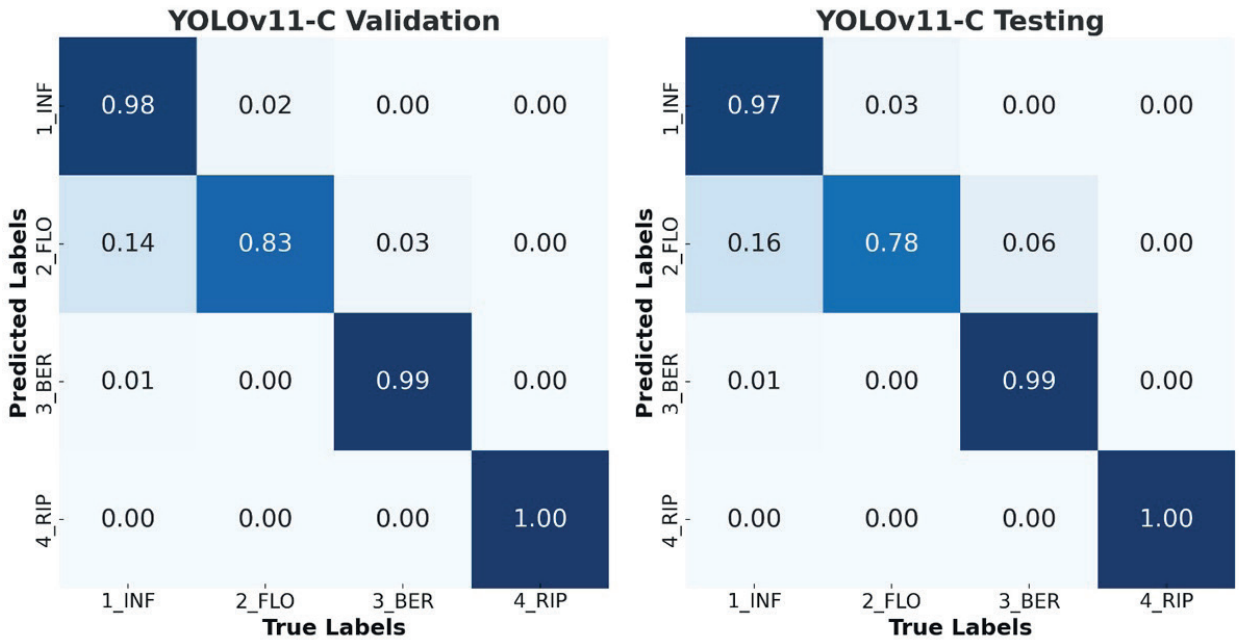


FIGURE 5. Normalised confusion matrices for YOLOv11-C during validation (left) and testing (right), illustrating the model's classification performance across the four phenological stages.

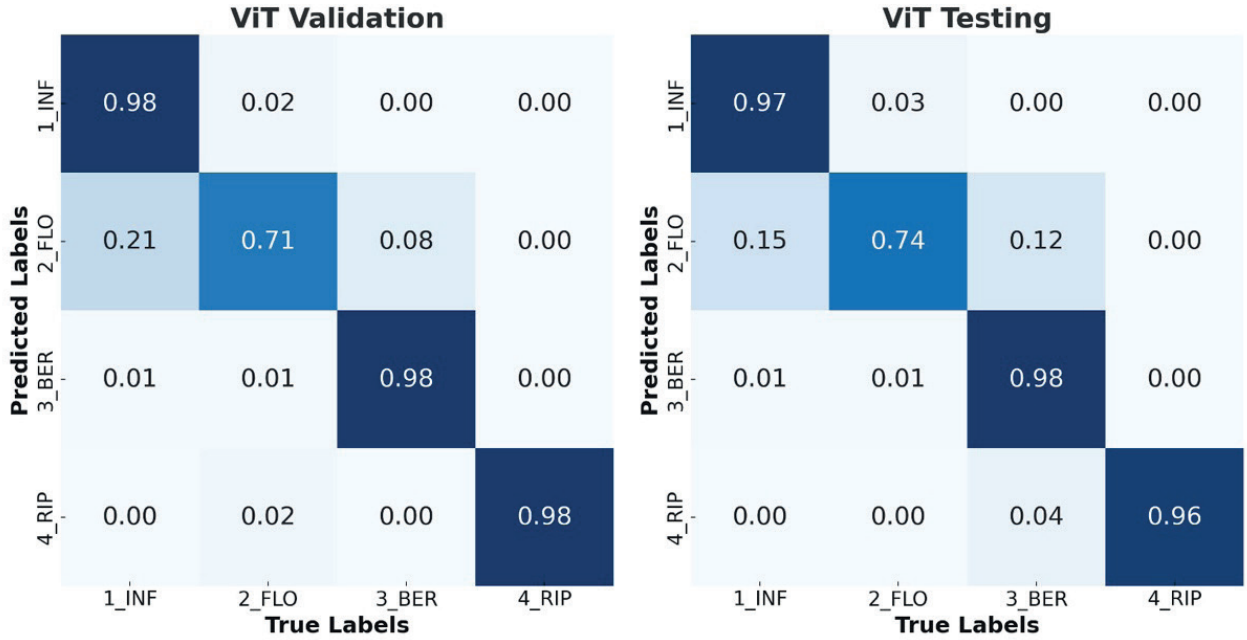


FIGURE 6. Normalised confusion matrices for ViT during validation (left) and testing (right), highlighting classification accuracy and misclassification patterns across phenological stages.

larger datasets to achieve the same class balance proficiency as CNN-based models.

The confusion matrices highlight that all models struggle to differentiate between 1_INF and 2_FLO, likely due to subtle visual similarities during transition phases. This trend has been observed in prior studies (Schiek *et al.,* 2023), emphasising the challenges of classifying phenological stages with gradual morphological changes. Previous research suggests that accuracy alone is insufficient for evaluating classification models, particularly for imbalanced datasets (Buda *et al.,* 2018). Complementary metrics such as precision, recall, and F1-score provide a more nuanced performance picture (Aguiar *et al.,* 2021).

Accuracy alone is not always sufficient to evaluate model performance in imbalanced datasets. As highlighted by Buda *et al.* (2018), the effectiveness of CNN models in classification tasks is directly influenced by the choice of evaluation metrics. In this study, the inclusion of the F1-score provides a more balanced assessment of model performance across all phenological stages, mitigating potential biases caused by class imbalances.

While accuracy and F1-score are key metrics, previous studies have emphasised the need for additional evaluation criteria for a more comprehensive analysis. Aguiar *et al.* (2021) demonstrated that incorporating average precision (AP) and mean average precision (mAP) improves the understanding of classification performance, reporting values ranging from 9.3 % to 49 %, and 12 % to 45 %, respectively. These findings highlight the variability in model performance across phenological stages and reinforce the importance of using multiple metrics to ensure a robust evaluation in crop classification.

Figure 7 illustrates examples of the model's predictions for the validation dataset, highlighting its ability to classify vineyard phenological stages under varying environmental conditions, including different lighting scenarios, canopy structures, and cultivars. Each image is labelled with its predicted phenological stage, providing a qualitative assessment of classification robustness.

The figure reveals that the models effectively distinguish between Shoot and Inflorescence Development (1_INF) and Berry Formation (3_BER), with minimal confusion between these stages. This aligns with the high classification accuracy observed in the confusion matrices, indicating that the models successfully capture distinct morphological traits at these developmental stages. However, Flowering (2_FLO) remains a challenging class, often visually resembling early shoot development (1_INF) or berry formation (3_BER), leading to some misclassifications. These results are consistent with previous studies emphasising the difficulty of detecting flowering stages due to phenotypic similarities and dataset limitations (Schieck *et al.,* 2023; Luoni *et al.,* 2024).

Additionally, YOLOv11's architecture plays a key role in classification efficiency. As noted by Khanam & Hussain (2024), its backbone and neck components enhance feature extraction and spatial attention, improving detection performance under variable lighting and occlusion conditions. This robustness is particularly advantageous for real-world vineyard monitoring, where environmental factors can significantly impact image quality.

Although the four-class framework used in this study provides practical insights aligned with major phenological transitions, it necessarily merges several adjacent E-L stages within each category. This reflects current limitations in dataset annotation and image resolution.

Future research could explore whether AI models can be trained to distinguish narrower phenological intervals; for example, every two or three E-L stages. This would allow for finer-grained phenological monitoring, potentially improving decision-making in vineyard management.

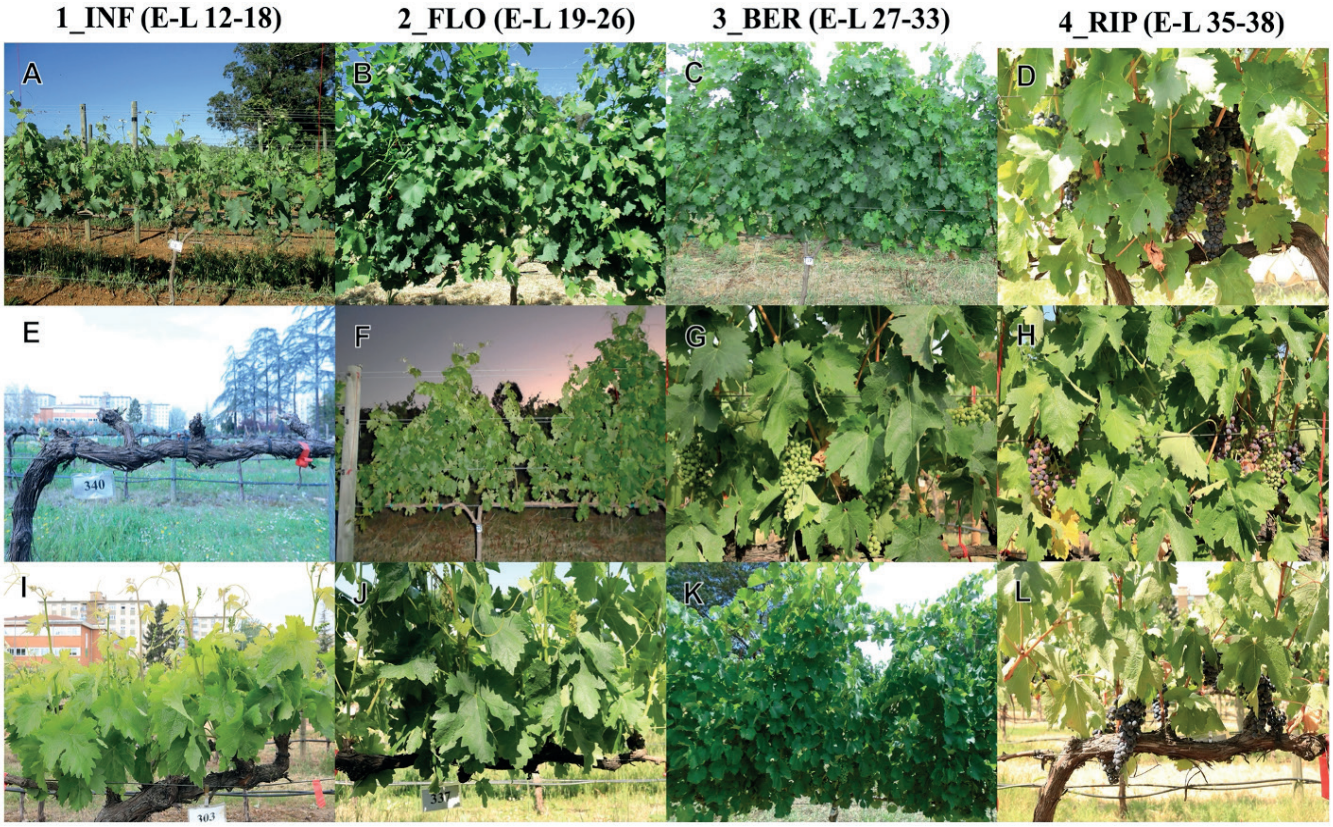


FIGURE 7. Examples of vineyard canopy images predicted by the model across the four major phenological stages, ordered from left to right: Shoot and Inflorescence Development (1_INF, E-L 12–18), Flowering (2_FLO, E-L 19–26), Berry Formation (3_BER, E-L 27–33), and Berry Ripening (4_RIP, E-L 35–38). Rows include different cultivars and training systems under various lighting conditions. Images A, B, and K correspond to Chenin Blanc; C and F to Cabernet-Sauvignon; and D, E, G, H, I, J, and L to Tempranillo.

3. Comparison of deep learning architectures

ResNet-34 achieved the highest validation (0.974) and testing accuracy (0.956), reinforcing its strong feature extraction capabilities. This supports previous studies where ResNet-based models outperformed other CNN architectures in phenological classification (Schieck *et al.,* 2023). However, its F1-score (0.953 validation, 0.918 testing) suggests persistent class imbalance issues, particularly when dealing with minority phenological stages. ViT exhibited a slightly lower accuracy (0.957 validation, 0.958 testing) and F1-score (0.924 validation, 0.919 testing), similar to previous findings where transformer-based models required larger datasets to achieve CNN-like performance (Luoni *et al.,* 2024).

YOLOv11-C demonstrated the most balanced classification performance, with an F1-score of 0.936 (validation) and 0.918 (testing), confirming its effectiveness in handling class imbalances. This aligns with Rodrigues *et al.* (2023), where YOLOv4 outperformed SSD-based models in crop classification (F1-score: 85.5 %). The model's efficiency in processing large datasets while maintaining high accuracy makes it particularly suitable for real-time vineyard monitoring. The present study aligns with the observed trend that single-stage models, such as YOLOv11-C, outperform two-stage approaches, like the YOLOv4-ResNet cascade used by Schiek *et al.* (2023) where ResNet achieved 88 % accuracy in grapevine phenology detection. Our study achieved 99 % accuracy for the same phenological stage using YOLOv11-C, trained for only 50 epochs, reinforcing the efficiency of single-stage architectures (Rasheed & Zarkoosh, 2024).

In terms of practical applications, ResNet-34 provides the highest accuracy, making it a strong candidate for offline vineyard monitoring, where real-time processing is unnecessary. It excels in structured classification tasks but has longer training times and is more susceptible to class imbalances, limiting its scalability.

YOLOv11-C emerges as the most well-rounded choice for vineyard phenology classification, balancing accuracy (0.962 validation, 0.954 testing), class representation (highest F1-score), and computational efficiency (fastest training time). Its real-time capabilities and ability to handle imbalanced datasets make it ideal for drone-based monitoring, automated vineyard assessments, and field-based deployment. This aligns with previous research demonstrating that YOLO-based models outperform other architectures in terms of speed, precision, and adaptability to dynamic agricultural conditions (Mahmud *et al.,* 2021).

ViT presents strong classification potential but remains computationally expensive, limiting real-time applicability unless significant hardware resources are available. Recent studies suggest that hybrid CNN-Transformer models could address this limitation. For example, Swin-Transformer-YOLOv5 enhances object detection accuracy (mAP 97 %) and improves robustness under variable lighting conditions (Lu *et al.,* 2022). Future research could explore similar hybrid architectures to achieve both high classification accuracy and computational efficiency.

These findings emphasise that YOLOv11-C offers the most scalable and practical solution for real-world vineyard phenology monitoring, particularly in precision viticulture applications requiring real-time analysis. However, further advancements in hybrid CNN-Transformer models could provide enhanced accuracy and efficiency, representing the next step in vineyard phenology classification.

4. Computational efficiency and real-time applicability

A key advantage of YOLO-based models is their computational efficiency, making them highly suitable for real-time deployment. In the present study, YOLOv11-C required the least training time (~56 minutes), significantly lower than ResNet-34 (~79 minutes), and ViT (~93 minutes). This efficiency is particularly relevant for precision viticulture, where models need to be integrated into edge devices or real-time monitoring systems. Several studies have emphasised YOLO's speed advantage, with Mahmud *et al.* (2021) reporting that a YOLOv4-based grape detection model processed images in 12 milliseconds (~83 FPS) while maintaining 95.6 % F1-score, demonstrating its potential for rapid on-the-fly vineyard monitoring.

Moreover, Ag-YOLO, a lightweight agricultural adaptation of YOLO, achieved real-time speeds of 36.5 FPS while achieving a 12-fold reduction in computational complexity compared to standard YOLO models, making it deployable on embedded devices (Qin *et al.,* 2021). These results suggest that YOLO is particularly well-suited for vineyard phenological classification in mobile or aerial drone-based monitoring, where real-time predictions are essential for automated interventions, such as precision spraying, and harvesting. The continuous evolution of YOLO-based models, from YOLOv1 to the advanced YOLOv11, reflects a sustained effort to improve detection and classification capabilities in real-world applications (Ali & Zhang, 2024). These improvements enhance the model's generalisation ability, particularly in complex agricultural environments where phenological stages exhibit natural variability

ViT, while competitive in accuracy, exhibited the longest training time (93 minutes), reflecting its higher computational cost. This aligns with previous research, where ViTs often require more extensive datasets and computational resources than CNNs. Luoni *et al.* (2024) reported that ViT-B16, despite achieving similar accuracy to ResNet, required significantly larger pre-training datasets, indicating a higher dependence on large-scale data for optimal performance.

5. Limitations and future work

While the deep learning models demonstrated strong classification performance, certain limitations must be acknowledged to guide future improvements. One

key challenge is dataset imbalance, particularly the underrepresentation of certain phenological stages, such as Flowering (2_FLO). This imbalance may contribute to misclassifications, as the models struggle to learn sufficient feature representations for minority classes. Expanding the dataset to include more samples from underrepresented stages, as well as transition phases between them, could enhance model performance and improve its ability to distinguish subtle variations in phenology. Additionally, the use of transfer learning with images from different vineyards, seasons, or years could help improve generalisation across diverse viticultural conditions.

Another limitation lies in the scalability of the models across different vineyard environments. The dataset used in this study was collected from two experimental vineyards, providing valuable insights but limiting the scope of the analysis. To evaluate model robustness, future research should include a broader range of grape varieties, training systems, and geographical regions. Environmental factors, such as variable lighting, canopy occlusions, and seasonal variations, also introduce classification challenges that should be addressed by incorporating more diverse datasets and advanced data augmentation techniques.

Enhancing model robustness through hybrid architectures represents another promising avenue. Combining CNN-based feature extraction with Transformer self-attention mechanisms could improve the models' contextual understanding while maintaining computational efficiency. Recent advances in hybrid architectures, such as Swin-Transformer-YOLOv5, have demonstrated improved performance in object detection tasks under complex environmental conditions (Lu *et al.,* 2022), suggesting their potential application in vineyard phenology classification.

Furthermore, expanding the classification framework to include finer-grained sub-stages of vine development could provide a more detailed understanding of phenological progression. However, achieving reliable classification at this level would require a significantly larger dataset to prevent overfitting and ensure accurate representation of all developmental stages.

Additionally, while our classification approach aggregates several E-L stages into broader categories, further work should investigate the feasibility of training models to predict more granular E-L stage transitions (*e.g.*, every 2–3 E-L stages), increasing applicability for precise phenological tracking.

Lastly, the practical application of these models in automated vineyard monitoring remains an essential area for future research. The feasibility of real-time implementation using edge computing or drone-based imaging should be explored, as YOLO-based architectures have shown particular promise for real-time agricultural applications (Qin *et al.,* 2021). Developing a streamlined deployment pipeline for phenology classification in precision viticulture could enhance vineyard management practices by enabling real-time decision-making for irrigation, fertilisation, and pest control strategies.

By addressing these limitations and exploring these advancements, future studies can further refine deep learning models for vineyard phenology classification, ensuring both high accuracy and practical applicability in viticulture.

CONCLUSION

This study demonstrates the potential of deep learning models for the automated classification of grapevine phenological stages under real field conditions. The evaluation of three architectures (ResNet-34, YOLOv11-C, and ViT) highlights their ability to discriminate phenological stages with high accuracy and F1-score. Among them, YOLOv11-C emerges as the most suitable model for vineyard monitoring due to its balance between classification performance, computational efficiency, and real-time applicability. Its ability to handle class imbalances while maintaining high accuracy makes it a robust choice for practical vineyard management.

ResNet-34 achieved the highest accuracy, making it a strong candidate for offline phenology analysis, especially when computational time is not a constraint. However, its lower F1-score suggests susceptibility to class imbalance problems, which may limit its generalisation in datasets with under-represented phenological stages. ViT demonstrated competitive classification performance but remains computationally demanding, making it less suitable for real-time applications unless powerful computing resources are available.

One of the main challenges identified in this study is the imbalance of the datasets, especially in the flowering stage (2_FLO), which showed higher misclassification rates. Future work should focus on expanding the datasets to include images from multiple vineyards, grape varieties, and environmental conditions to improve the generalisation capabilities of the models. In addition, the use of hybrid architectures—such as CNN-Transformer combinations— could improve classification performance by integrating the feature extraction power of CNNs with the global contextual awareness of Transformers.

The practical implementation of deep learning-based phenology classification in precision viticulture requires further exploration of real-time deployment strategies, including edge computing and drone-based imaging systems. Given the efficiency of YOLOv11-C, future studies could optimise this model for embedded systems, facilitating its use in automated vineyard monitoring, spraying, and harvest planning.

This study reinforces the growing role of AI in viticulture and offers new opportunities for data-driven decision-making in vineyard management. By addressing the identified limitations and exploring advanced architectures and real-time applications, deep learning models can significantly contribute to improving phenological monitoring and increasing the efficiency of vineyard operations. Additionally, future research should aim at training AI models capable of predicting specific E-L stages rather than broader phenological phases.

Providing precise phenological stage estimates (*e.g.*, E-L 23 for full bloom) would significantly enhance the agronomic value of these tools, enabling more targeted interventions in vineyard management.

ACKNOWLEDGEMENTS

The authors would like to acknowledge South African WINE, through the project "Establishment of the technical and scientific bases for AI applications in wine production. Study case on viticulture, yield, and phenology AI models" for funding the research. We would also like to thank Research Funding FPI Grant 591/2021 from Universidad de La Rioja, Gobierno de La Rioja, Spain, and the Scholarship Program of the University of Talca, Chile (Internationalization of Master's Degree Programs, R.U. Nº 136/2019) for supporting student M. Ignacia Gonzalez.

REFERENCES

Aguiar, A. S., Magalhães, S. A., dos Santos, F. N., Castro, L., Pinho, T., Valente, J., Martins, R., & Boaventura-Cunha, J. (2021). Grape Bunch Detection at Different Growth Stages Using Deep Learning Quantized Models. *Agronomy, 11*(9), 1890. https://doi.org/10.3390/agronomy11091890

Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers, 13*(12), 336. https://doi.org/10.3390/computers13120336

Altimiras, F., Pavéz, L., Pourreza, A., Yañez, O., González-Rodríguez, L., & Leiva-Araos, A. (2024). Transcriptome data analysis applied to grapevine growth stage identification. *Agronomy, 14*, 613. https://doi.org/10.3390/agronomy14030613

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Coombe, B. G. (1995). Growth stages of the grapevine: Adoption of a system for identifying grapevine growth stages. *Australian Journal of Grape and Wine Research, 1*, 100-110 https://doi.org/10.1111/j.1755-0238.1995.tb00086.x

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

Hernández, I., Gutiérrez, S., Ceballos, S., Íñiguez, R., Barrio, I., & Tardáguila, J. (2021). Artificial intelligence and novel sensing technologies for assessing downy mildew in grapevine. *Horticulturae, 7*(5), 103. https://doi.org/10.3390/horticulturae7050103

Hugging Face. (2024). *ViT-Base-Patch16-384 model card*. Retrieved from https://huggingface.co/google/vit-base-patch16-384

Íñiguez, R., Gutiérrez, S., Poblete-Echeverría, C., Hernández, I., Barrio, I., & Tardáguila, J. (2024). Deep learning modelling for non-invasive grape bunch detection under diverse occlusion conditions. *Computers and Electronics in Agriculture, 226*, 109421. https://doi.org/10.1016/j.compag.2024.109421

Khanam, R., & Hussain, M. (2024). YOLOv11: An Overview of the Key Architectural Enhancements [Unpublished manuscript]. *Department of Computer Science Huddersfield University, Queensgate, Huddersfield HD1 3DH, UK*. https://doi.org/10.48550/arXiv.2410.17725

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lorenz, D. H., Bleiholder, H., Klose, R., Meier, U., & Weber, E. (1995). Phenological growth stages of the grapevine (Vitis vinifera L. ssp. vinifera)—Codes and descriptions according to the extended BBCH scale. *Australian Journal of Grape and Wine Research, 1*, 100-110. https://doi.org/10.1111/j.1755-0238.1995.tb00085.x

Lu, S., Liu, X., He, Z., Zhang, X., Liu, W., & Karkee, M. (2022). Swin-Transformer-YOLOv5 for real-time wine grape bunch detection. *Remote Sensing*, *14*(22), 5853. https://doi.org/10.3390/rs14225853

Luoni, S. A. B., Ricci, R., Corzo, M. A., Hoxha, G., Melgani, F., & Fernandez, P. (2024). Sunpheno: A Deep Neural Network for Phenological Classification of Sunflower Images. *Plants 2024*, *13*, 1998. https://doi.org/10.3390/plants13141998

Mahmud, M. S., Zahid, A., Das, A. K., Muzammil, M., & Khan, M. U. (2021). A systematic literature review on deep learning applications for precision cattle farming. *Computers and Electronics in Agriculture*, *187*, 106313. https://doi.org/10.1016/j.compag.2021.106313

Mendes, J., Peres, E., Santos, F. N., Silva, N., Silva, R., Sousa, J. J., Cortez, I., & Morais, R. (2022). VineInspector: The vineyard assistant. *Agriculture, 12*(5), 730. https://doi.org/10.3390/agriculture12050730

Mullins, M. G., Bouquet, A., & Williams, L. E. (1992). *Biology of the grapevine.* Cambridge University Press

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Qin, Z., Wang, W., Dammer, K. H., Guo, L., & Cao, Z. (2021). Ag-YOLO: A real-time low-cost detector for precise spraying with case study of palms. *Frontiers in Plant Science, 12*, 753603. https://doi.org/10.3389/fpls.2021.753603

Rasheed, A. F., & Zarkoosh, M. (2024). YOLOv11 Optimization for Efficient Resource Utilization [Unpublished manuscript]. *Department of Computer Science, Cornell University.* http://doi.org/10.48550/arXiv.2412.14790

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. https://doi.org/10.1109/CVPR.2016.91

Reis, S., Fraga, H., Carlos, C., Silvestre, J., Eiras-Dias, J., Rodrigues, P., & Santos, J. A. (2020). Grapevine phenology in four Portuguese wine regions: modeling and predictions. *Applied Sciences*, *10*(11), 3708. https://doi.org/10.3390/app10113708

Rodrigues, L., Magalhães, S. A., da Silva, D. Q., dos Santos, F. N., & Cunha, M. (2023). Computer vision and deep learning as tools for leveraging dynamic phenological classification in vegetable crops. *Agronomy*, *13*(2), 463. https://doi.org/10.3390/agronomy13020463

Schieck, M., Krajsic, P., Loos, F., Hussein, A., & Franczyk, B. (2023). Comparison of deep learning methods for grapevine growth stage recognition. *Computers and Electronics in Agriculture, 211*, 107944. https://doi.org/10.1016/j.compag.2023.107944

Silva, J. A. O. S., Siqueira, V. S. d., Mesquita, M., Vale, L. S. R., Silva, J. L. B. d., Silva, M. V. d., Lemos, J. P. B., Lacerda, L. N., Ferrarezi, R. S., & Oliveira, H. F. E. d. (2024). Artificial Intelligence Applied to Support Agronomic Decisions for the Automatic Aerial Analysis Images Captured by UAV: A Systematic Review. *Agronomy, 14*(11), 2697. https://doi.org/10.3390/agronomy14112697

Verdugo-Vásquez, N., Acevedo-Opazo, C., Valdés-Gómez, H., Ingram, B., García de Cortázar-Atauri, I., & Tisseyre, B. (2018). Temporal stability of within-field variability of total soluble solids of grapevine under semi-arid conditions: A first step towards a spatial model. *OENO One*, *52*(1), 15–30. https://doi.org/10.20870/oeno-one.2018.52.1.1782

Waskom, M. L. (2021). *Seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Yu, F., Zhang, Q., Xiao, J., Ma, Y., Wang, M., Luan, R., & Liu, X. (2023). Progress in the application of CNN-based image classification and recognition in whole crop growth cycles. *Remote Sensing, 15*(12), 2988. https://doi.org/10.3390/rs15122988