POLITECNICO DI TORINO
Repository ISTITUZIONALE

Explainable AI-based two-stage Swin transformer for grapevine leaf-based variety and disease classification

Original
Explainable AI-based two-stage Swin transformer for grapevine leaf-based variety and disease classification / Shah, S.T.H., Shah, S.A.H., Godio, S., Kassem, K., Ahmad Qureshi, S., Bilal Hussain, S., Deriu, M.A.. - In: INFORMATION PROCESSING IN AGRICULTURE. - ISSN 2214-3173. - (2026), pp. -1. [10.1016/j.inpa.2026.07.010]

Availability:
This version is available at: 11583/3013827 since: 2026-08-02T13:39:19Z

Publisher:
Elsevier

Published
DOI:10.1016/j.inpa.2026.07.010

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

22 November 2026

Information Processing in Agriculture xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Processing in Agriculture

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/

Explainable AI-based two-stage Swin transformer for grapevine leaf-based variety and disease classification

Syed Taimoor Hussain Shah [a,*], Syed Adil Hussain Shah [a,b], Silvia Godio [a], Karim Kassem [a,c], Shahzad Ahmad Qureshi [d], Syed Bilal Hussain [e,*], Marco Agostino Deriu [a]

[a] PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy
[b] Department of Research and Development (R&D), GPI SpA, Trento 38123, Italy
[c] Centro Medico Santagostino, Milan, Italy
[d] Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad 45650, Pakistan
[e] School of Agriculture and Food Science, University College Dublin, Dublin 4 D04V1W8, Ireland

ARTICLE INFO

Keywords:
Precision viticulture
Grape leaf classification
Swin Transformer
Disease diagnosis
Explainable AI

ABSTRACT

Accurate grapevine cultivar identification and leaf disease diagnosis are important for precision viticulture but remain challenging due to variation in cultivar morphology, disease symptoms, imaging conditions, and dataset sources. This study developed an application-oriented, explainable, two-stage deep learning framework for grapevine leaf analysis that uses a standard Swin-Tiny backbone without architectural modifications. In Stage 1, representative CNN, transformer, and hybrid models were benchmarked on a 16-class grapevine cultivar dataset to learn cultivar-related features. In Stage 2, Swin-Tiny was fine-tuned on a unified 7-class grape leaf disease dataset to evaluate cultivar-to-disease transfer learning. Expanded benchmarking showed that MobileViT-S, MaxViT-Tiny, ResNet50, ConvNeXtV2-Tiny, and InceptionV3 performed strongly across tasks. On the fixed cultivar split used for detailed analysis, Swin-Tiny achieved 85.00% test balanced accuracy, 83.44% F1-score, and 0.8428 MCC; repeated-split analysis gave a higher mean test balanced accuracy of 96.50 ± 3.11%, confirming split sensitivity. For disease classification, the two-stage Swin-Tiny model achieved 99.07% validation accuracy and 99.53% test accuracy, with corresponding balanced accuracies of 99.01% and 99.71%, respectively. However, direct ImageNet-pretrained disease-only Swin-Tiny fine-tuning reached 100.00% test balanced accuracy, indicating that cultivar-stage pretraining did not numerically improve Swin-Tiny in this setting. Multi-backbone comparisons showed that cultivar-to-disease transfer was architecture-dependent rather than uniformly beneficial. Explainable AI visualisations highlighted morphology-related regions for cultivar recognition and symptom-related regions for disease classification. The implementation is available at doi:https://doi.org/10.5281/zenodo.18937849. Overall, the framework provides a transparent and reproducible benchmark baseline for grapevine leaf analysis.

1. Introduction

Grapevine (*Vitis vinifera*) is a globally important horticultural crop grown for fresh consumption, raisin production, and winemaking. In recent years, vineyard management has faced growing challenges from climate variability, foliar disease pressure, and heterogeneous canopy conditions, all of which make manual scouting difficult to standardise across blocks and seasons [1]. The International Organisation of Vine and Wine (OIV) has reported a global vineyard area of approximately 7.2 million hectares, indicating that even marginal inefficiencies in scouting, diagnosis, and the timing of interventions may result in considerable economic and environmental consequences across production regions. Therefore, there is a great need for scalable monitoring approaches that can convert visual crop information into actionable evidence for intervention planning, resource optimisation, and the protection of yield and fruit quality.

In this context, two visual analysis tasks are especially relevant. The first is grapevine cultivar identification, which supports varietal

* Corresponding authors.
E-mail addresses: taimoor.shah@polito.it (S.T.H. Shah), syedadilhussain.shah@gpi.it (S.A.H. Shah), silvia.godio@polito.it (S. Godio), karim.kassem@polito.it (K. Kassem), drsaqureshi@pieas.edu.pk (S.A. Qureshi), syedbilal.hussain@ucd.ie (S.B. Hussain), marco.deriu@polito.it (M.A. Deriu).

https://doi.org/10.1016/j.inpa.2026.07.010
Received 13 April 2026; Received in revised form 21 July 2026; Accepted 23 July 2026
Available online 27 July 2026
2214-3173/© 2026 The Author(s). Published by Elsevier B.V. on behalf of China Agricultural University. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

traceability, phenotyping, and management decisions on canopy structure, irrigation scheduling, and harvest management. The second is early disease recognition, which is critical for reducing treatment delays and limiting quality degradation and yield loss. Grapevine downy mildew, for instance, can spread rapidly under warm, wet conditions, leading to severe outbreaks, defoliation, and substantial crop losses if not identified in time. Together, cultivar-aware and disease-aware image analysis can strengthen vineyard decision-support systems by enabling more targeted scouting, more timely responses, and more consistent assessment across large production areas [2,3].

In practice, however, visual assessment of vineyards remains challenging because leaf appearance is influenced by cultivar-specific morphology, microclimate, illumination, canopy density, camera viewpoint, growth stage, occlusion, background clutter, and symptom severity [4]. These sources of variability reduce the consistency of human inspection, so that even experienced observers may disagree, while large-scale vineyard scouting becomes labour-intensive and difficult to standardise. This gap has driven the adoption of automated image-based methods that can provide more repeatable and scalable predictions under real-world imaging conditions [5]. In the context of precision agriculture, such models are valuable not merely as classification tools but as components of operational monitoring pipelines that support high-throughput, site-specific vineyard management.

In recent years, computer vision and deep learning have emerged as practical tools to automate plant phenotyping and disease diagnosis from images. Pacal et al. [6] highlight rapid progress across classification, detection, and segmentation tasks for plant health monitoring, largely driven by convolutional neural networks (CNNs), improved data augmentation, and transfer learning from large pretraining corpora. Transfer learning is particularly attractive in agriculture, where labelled data can be limited or imbalanced; pre-trained models can be adapted efficiently to domain-specific leaf imagery with reduced training time and improved generalisation [7,8].

Beyond CNNs, transformer-based vision backbones have gained attention for their ability to model contextual dependencies. The Swin Transformer [9] introduces a hierarchical representation with shifted-window self-attention, enabling strong performance with computational efficiency that scales linearly with image size, an advantage for both classification and dense recognition. At the same time, there is a growing expectation that AI systems used in agricultural monitoring be interpretable and operationally trustworthy. Explainable AI (XAI) methods can help reveal whether model predictions are grounded in meaningful visual evidence, support error analysis, and improve user confidence in deployment settings [10–12]. Accordingly, complementary gradient-based, perturbation-based, and attention-based explanation strategies have become increasingly important in agricultural vision systems [9,13–16].

Existing studies [17–19] have demonstrated promising results for grapevine cultivar recognition and plant disease classification. Yet, these tasks are often addressed independently, with limited attention to whether cultivar-specific morphology can serve as an informative intermediate representation for downstream disease recognition. In addition, many studies focus primarily on predictive accuracy while providing limited interpretability, deployment-oriented support, or reproducibility resources. As a result, the potential of a unified, explainable information-processing framework for grapevine leaf analysis remains underexplored.

The current study proposes a unified, explainable deep learning framework for grapevine leaf analysis that addresses both cultivar recognition (16 classes) and disease classification (7 classes) through a two-stage transfer learning strategy. In Stage 1, a backbone model is trained to identify grapevine cultivars from leaf images, encouraging the network to learn stable morphological and venation cues. In Stage 2, the same backbone is fine-tuned to recognise disease symptoms, thereby adapting these structural representations for pathology-focused classification. To operationalise this two-stage framework, we evaluate representative CNN- and transformer-based backbones [9] as candidate encoders and select a suitable model for learning transferable grapevine leaf representations. To improve practical usability, we also provide a lightweight prototype platform that accepts a leaf image and returns the predicted class, confidence scores, and visual explanations for end users, while recognising that further field validation is required before operational decision-support use.

The added value of this study lies in integrating cultivar recognition and disease diagnosis into a single explainable two-stage framework, rather than proposing a new Transformer architecture or treating the two tasks as independent classification problems. In this work, Swin-Tiny is adopted as an existing vision backbone without modifying its patch embedding, patch merging, shifted-window self-attention, or overall network structure. The methodological contribution is instead application-oriented: cultivar-specific morphological representations are first learned from a 16-class grapevine leaf dataset and then transferred to a 7-class disease classification task to experimentally test whether leaf-structure knowledge provides useful downstream information for symptom recognition. The revised comparison with direct disease-only fine-tuning shows that this transfer effect should be interpreted cautiously and is not uniformly superior across backbones. In addition, representative CNN- and transformer-based backbones are evaluated under a unified training and evaluation protocol; complementary explainable AI methods are used to distinguish morphology-driven evidence from symptom-driven evidence; and a lightweight image-to-decision platform with reproducibility resources is provided.

The main contributions of this work are summarised as follows:

- We formulate and evaluate grapevine leaf analysis as a two-stage transfer-learning problem in which cultivar-specific morphology is first learned as an intermediate representation and then adapted to disease recognition, while showing through direct disease-only comparisons that the benefit of this transfer is architecture-dependent rather than universal.
- We evaluate representative established and recent CNN, transformer, and hybrid backbones within a unified training protocol, positioning Swin-Tiny as the selected backbone for the explainable Swin-based pipeline while transparently comparing it with newer alternatives.
- We incorporate complementary post-hoc XAI methods, including gradient-based, perturbation-based, and attention-based visualisations, to provide qualitative evidence of model attention and support error analysis.
- We develop a prototype image-to-decision platform that provides predicted labels with confidence scores and visual explanations for end users, intended as a reproducible demonstration rather than a fully validated field-deployment system.
- We support reproducibility through fixed-seed training, saved checkpoints and configurations, and exported metrics and plots for all experiments.

Karim et al. [20] proposed a lightweight, edge-deployable grape leaf disease classifier by modifying MobileNetV3Large and deploying it on a Jetson Nano for real-time inference. Using a 4-class grape leaf disease dataset, they compared six lightweight backbones. They reported that the customised MobileNetV3Large achieved 99.66% training accuracy and 99.42% test accuracy, while also emphasising low inference latency and power usage for field deployment. They also integrated Grad-CAM to visualise diseased regions, including a transparency-threshold variant, and presented real-time explanatory outputs through a graphical user interface, as summarised in Supplementary Table S1.

Prasad et al. [19] proposed a VGG16-based DCNN augmented with additional CNN layers and data augmentation for multiclass grape leaf disease identification. Using a publicly available grape leaf disease dataset, they reported 99.18% training accuracy and 99.04% test accuracy, concluding that augmentation improved generalisation and reduced overfitting. The study reported performance using accuracy,

loss curves, and confusion-matrix-style evaluation, but it did not include an explicitly defined explainability component.

Ishengoma and Lyimo [21] proposed an ensemble strategy that treats multiple pre-trained CNNs (e.g., VGG16, InceptionV3, Xception, ResNet50) as parallel feature extractors, followed by a Random Forest classifier, targeting improved performance under constrained dataset sizes. Their approach was evaluated on a grape leaf disease dataset split into original and modified/augmented subsets; an independent summary of their reported results indicates that the model trained on the modified/augmented subset achieved a peak accuracy of 95.34%, outperforming the model trained on the original subset. The study focused on predictive performance and did not report a formal XAI analysis.

Hu and Zhang [22] introduced GCS-YOLO, a lightweight grape leaf disease *detection* method built on an improved YOLOv8 design. The method replaces key blocks with a lightweight feature extractor (C2f-GR), integrates CBAM attention, and refines the detection head to improve accuracy while reducing parameters and computation. On a grape leaf disease detection benchmark, they reported 96.20% mAP@0.5, highlighting real-time suitability. The work emphasised detection performance and ablation analysis but did not include a formal explainability module.

Xie et al. [23] proposed a real-time grape leaf disease detector termed Faster DR-IACNN, built upon Faster R-CNN with enhanced feature extraction modules. Using their GLDD dataset, they reported an mAP of 81.10% and a detection speed of 15.01 FPS, demonstrating feasibility for real-time disease monitoring in practical settings. The study emphasised detection performance and speed, without incorporating an explicit explainability analysis.

Kunduracioglu and Pacal [24] performed a broad benchmark of fine-tuned CNNs and vision transformers for grape leaf disease classification on the PlantVillage grape subset (reported as 4062 images across 4 classes). They reported accuracy ranging from 97.00% to 100.00%, including 100.00% for VGG16/VGG19 and 99.84% for Xception, with several transformer models also reaching 100.00%. The paper provides confusion matrices and an extensive comparative analysis, but does not explicitly describe the use of formal XAI techniques.

Atesoglu and Bingol [18] proposed an improved hybrid disease diagnosis pipeline that combines CNN representations with handcrafted texture features. Specifically, they extracted deep features using DenseNet201, combined them with LBP and HOG descriptors, applied Neighbourhood Component Analysis (NCA) for feature selection, and classified using an ML classifier (reported best with SVM). On a 4-class grape leaf disease setting (Black Measles/Black Rot/Isariopsis Leaf Spot/Healthy), they reported 99.10% accuracy. The study did not report an explicit explainability component, as summarised in Supplementary Table S1.

Peng et al. [25] proposed a fast diagnosis method based on fused deep features + SVM, where features extracted from different CNNs are fused (e.g., direct concatenation or correlation-based fusion) to improve discriminative power while keeping SVM training time very low. Using a grape leaf disease dataset (reported 4062 images, 4 classes), they reported multiple top-performing fusion outcomes, including an accuracy of 99.57% with F1 of 99.65% under one fusion configuration, and a best-performing configuration reaching an accuracy of 99.77% with F1 of 99.81%. The work focused on performance and computational efficiency, without explicitly employing formal XAI techniques.

Javidan et al. [26] presented a classical computer-vision and machine-learning pipeline in which regions of interest were separated using automatic K-means clustering, features were extracted across multiple colour spaces, and classification was performed using multi-class SVM, with PCA for dimensionality reduction and Relief for feature selection. Although the study is informative as a traditional baseline, it addresses grape leaf disease classification rather than cultivar recognition. On a 4-class disease task, they reported 98.71% accuracy using GLCM features and 98.97% accuracy after PCA-based reduction, while also comparing against CNN and GoogLeNet baselines. No formal XAI method was reported, as detailed in Supplementary Table S1. This distinction is important because cultivar recognition and disease classification involve related but not identical visual cues, and the literature on cultivar-specific morphology remains comparatively smaller.

Terzi et al. [27] proposed a newly designed CNN for automated grape variety identification using ampelographic characteristics and evaluated transfer-learning baselines (GoogleNet and AlexNet). Their dataset included 27,320 images (leaf, bunch, and fruit) across 50 varieties, split into leaf and cluster/fruit groups; the reported best class accuracy reached 94.10% for the leaf group and 97.20% for the cluster/fruit group, and they reported transfer-learning accuracies of 84.39% (GoogleNet) and 92.31% (AlexNet). The work did not explicitly report formal XAI techniques.

Deng et al. [28] proposed ICS-MobileNetV3-Small (ICS-MS) for grape leaf cultivar recognition under complex backgrounds, integrating architectural modules to improve feature representation while keeping the model lightweight. They used a public dataset comprising 11 cultivars, expanded it via systematic augmentation to 6051 images, and reported 96.53% accuracy (with precision/recall/F1 around 96%) while maintaining only 1.17 M parameters. The study provides qualitative feature visualisation comparisons between the baseline and ICS-MS outputs, but it does not explicitly employ standard XAI methods in a method-labelled manner.

Vélez et al. [29] explored “digital ampelography” using a CNN trained with Keras/TensorFlow to classify grapevine cultivars from leaf images. Using a dataset of 1011 leaf images from two cultivars, they reported cultivar-specific accuracies of 95.80% (Verdejo) and 100.00% (Tempranillo), demonstrating that CNNs can extract discriminative leaf traits for cultivar recognition. The work did not explicitly report the use of formal explainable-AI techniques.

Sugiura et al. [30] evaluated VGG16, ResNet50, and Vision Transformer (ViT) for grape variety identification using smartphone images of leaves and berries from three varieties (Shine Muscat and two similar cultivars). They tested the trained models on separate test sets and reported accuracies of >96.10% for leaves, >99.40% for young berries, and 100.00% for mature berries. They further discussed visualisation of the regions used by the model for classification; however, the paper does not clearly identify a standard XAI method, so this is better described as focus-region visualisation rather than a formal explainability analysis.

Ahmed et al. [17] addressed grapevine leaf variety classification (cultivar identification) by adapting a pre-trained DenseNet201 through fine-tuning and systematically studying the impact of layer freezing on recognition performance. Using a public dataset of 500 RGB leaf images distributed across 5 varieties (100 images per class), they applied multiple data augmentation operations to enlarge the training set and mitigate overfitting. Their best configuration, a fine-tuned DenseNet201 variant (DenseNet-30), achieved 97.22% overall accuracy and was reported to outperform prior work on the same dataset. The paper reports standard quantitative evaluation and includes confusion-matrix-style analysis, but it does not explicitly apply established explainable-AI techniques.

Ji et al. [31] provide the closest task-level comparison to the present disease classifier. Their UnitedModel integrated multiple CNNs to distinguish black rot, esca, isariopsis leaf spot, and healthy grape leaves on a PlantVillage hold-out set, reporting 98.57% test accuracy. This result is directly relevant to grape-leaf disease classification, although it is not a head-to-head benchmark because the present work uses seven classes assembled from two sources and a different split protocol. In a viticulture-specific but task-different setting, Liu et al. [32] developed the lightweight YOLOX-RA model for grape-bunch detection under dense growth and occlusion, reporting 88.75% mAP, 84.88 FPS, and a 17.53 MB model; it therefore provides deployment and unstructured-vineyard context rather than a classification-accuracy comparator.

Several other recent studies provide partial methodological comparisons. Fu et al. [33] used an improved Vision Transformer for seven-

class crop-pest recognition, demonstrating the relevance of self-attention to spatially distributed visual evidence. Ngugi et al. [34] combined semantic segmentation with lesion-wise CNN classification and reported that recognising individual lesions improved GoogLeNet accuracy by more than 15% relative to whole-leaf recognition; their lesion class achieved an mIoU of 0.6257. Ametefe et al. [35] combined Canny edges, colour-intensity information, augmentation, and transfer learning, reporting 99.03% accuracy with DenseNet201 and 98.23% with EfficientNetB3. Qiao et al. [36] incorporated a transformer module, small-object detection head, and CBAM into YOLOv5s for cucumber downy mildew spore detection, obtaining 95.40% mAP@0.5, 89.10% precision, and 90.30% recall. Huang et al. [37] fused ConvNeXt and Vision Transformer representations in EConv-ViT and reported 99.20% accuracy on laboratory apple-leaf images but 79.30% on natural-environment images, illustrating the laboratory-to-field generalisation gap.

Explainability-focused original studies also provide relevant context. Polimena et al. [38] used an explainable balanced Random Forest to estimate ammonia and chlorophyll from rocket-leaf images and evaluated whether the learned features were agronomically reasonable. Shah et al. [39] proposed the ResTS residual teacher/student architecture for plant-disease classification and symptom-region visualisation, reporting an F1-score of 0.991 on PlantVillage. The present work differs by applying complementary post-hoc explanation families to an unmodified Swin-Tiny backbone and adding limited occlusion and deletion/insertion checks, rather than embedding a single visualisation mechanism into the classifier.

These studies are considered according to task comparability. Grape-leaf image classification is treated as the closest direct comparison; cross-crop leaf classification, lesion localisation, and interpretable plant-disease classification are treated as partial methodological comparisons; and grape-bunch detection, crop-pest recognition, microscopic spore detection, and leaf-quality regression are used only as contextual evidence. This distinction avoids ranking accuracy, F1-score, mAP, and regression error as if they were equivalent outcomes.

Taken together, recent studies show that grape leaf analysis has progressed substantially in both disease recognition and cultivar identification. Deep learning methods frequently report very high performance, especially on controlled or benchmark datasets, while hybrid pipelines remain competitive in smaller-data settings. However, only a limited number of studies explicitly integrate interpretability, and even fewer attempt to connect cultivar-level morphological learning with downstream disease recognition in a unified transfer-learning framework. This gap is especially relevant for practical vineyard monitoring, where transparent and reusable representations may improve both prediction quality and user trust.

A common challenge across many grape leaf disease and cultivar studies is real-world generalisation. Many reported high accuracies are achieved on datasets collected under controlled conditions (such as uniform backgrounds and consistent lighting), which simplify classification but do not capture vineyard variability, including shadows, occlusions, cluttered backgrounds, and mixed symptom stages. This is especially important for work that uses PlantVillage-style imagery, which is often described as collected in controlled environments with neutral backgrounds, making it useful for initial benchmarking but requiring further adjustments for field application.

Second, many studies rely heavily on augmentation to compensate for small datasets, which can artificially boost performance without accurately reflecting changes in the field distribution. Even when “field” datasets are used, authors still note that natural environments introduce factors such as boundary blur, dense adhesion, irregular morphology, and varying illumination, all of which significantly reduce the recognition accuracy of many existing algorithms, underscoring that robustness remains an ongoing challenge.

Third, while explainability is increasingly recognised as important for agricultural adoption, most studies either do not include explicit XAI or provide only limited visualisation. Recent surveys on XAI for plant health monitoring highlight that interpretability and transparency remain barriers to wider adoption, and existing work often focuses on model performance while underreporting explainability, stakeholder validation, or explanation faithfulness.

Finally, deployment-focused considerations are often incomplete. A few studies demonstrate the feasibility of real-time detection on mobile or edge devices. Still, many papers do not report runtime, calibration, robustness checks, or failure-case analysis, even though real-time requirements are explicitly noted as missing in earlier grape leaf detection research.

In summary, the literature demonstrates strong progress in grape leaf disease recognition and cultivar classification, but still reveals three important gaps: limited integration of interpretability, insufficient analysis of real-world robustness, and a lack of unified frameworks that transfer cultivar-specific morphological knowledge to disease diagnosis. These gaps directly motivate the proposed two-stage explainable framework developed in this study. Against this background, the novelty of the present work lies in a controlled cultivar-to-disease transfer analysis with disease-only, random-initialisation, and frozen-feature controls; multi-backbone benchmarking; complementary post-hoc explanations; source- and split-sensitivity checks; and a reproducible prototype, rather than in proposing a new backbone architecture.

The remainder of the paper is organised as follows. Section 2 presents the datasets, preprocessing and augmentation pipeline, model candidates, training protocol, experimental environment, and evaluation metrics. Section 3 reports and discusses cultivar- and disease-classification results, transfer controls, robustness, comparisons with prior work, explainability, and limitations. Section 4 presents the conclusions and future research directions.

2. Materials and methods

This section describes the data sources and the construction of two unified benchmarks (cultivar and disease), followed by the training protocol, model candidates, selection strategy, and the final Swin Transformer fine-tuning procedure used to obtain the reported results (Fig. 1).

2.1. Datasets

2.1.1. Cultivar (grape leaf varieties) dataset construction (16 classes)

We created a unified 16-class cultivar dataset by combining two publicly available grapevine leaf datasets: (i) a controlled-acquisition dataset containing 5 varieties (Ak, Ala Idris, Büzgülü, Dimnit, Nazli) [40,41] with 500 RGB images (100 per class) acquired using a special self-illuminating system, and (ii) a second dataset [42] containing 11 *Vitis vinifera* varieties with 1009 images (natural leaf photos). The first dataset is explicitly described as having images with 512 × 512 resolution and controlled acquisition conditions, as detailed in Tables 1 and 2.

To standardise evaluation while preserving sufficient training data for the relatively small and imbalanced cultivar dataset, we applied a fixed per-class holdout rule: 5 images per class for validation and 5 images per class for testing, with all remaining images used for training. This ensured that every cultivar was represented in both validation and test sets, even when class sizes differed across sources. We acknowledge that this split yields small class-level validation and test support sets; therefore, class-wise results should be interpreted with caution, as a single misclassification can noticeably affect per-class metrics. For this reason, model comparison was based primarily on aggregate metrics such as balanced accuracy, macro-F1, and MCC rather than on isolated class-level conclusions.

In addition to this fixed split, we later performed a repeated-split stability analysis to assess how sensitive cultivar-recognition performance was to the particular assignment of images to training, validation,

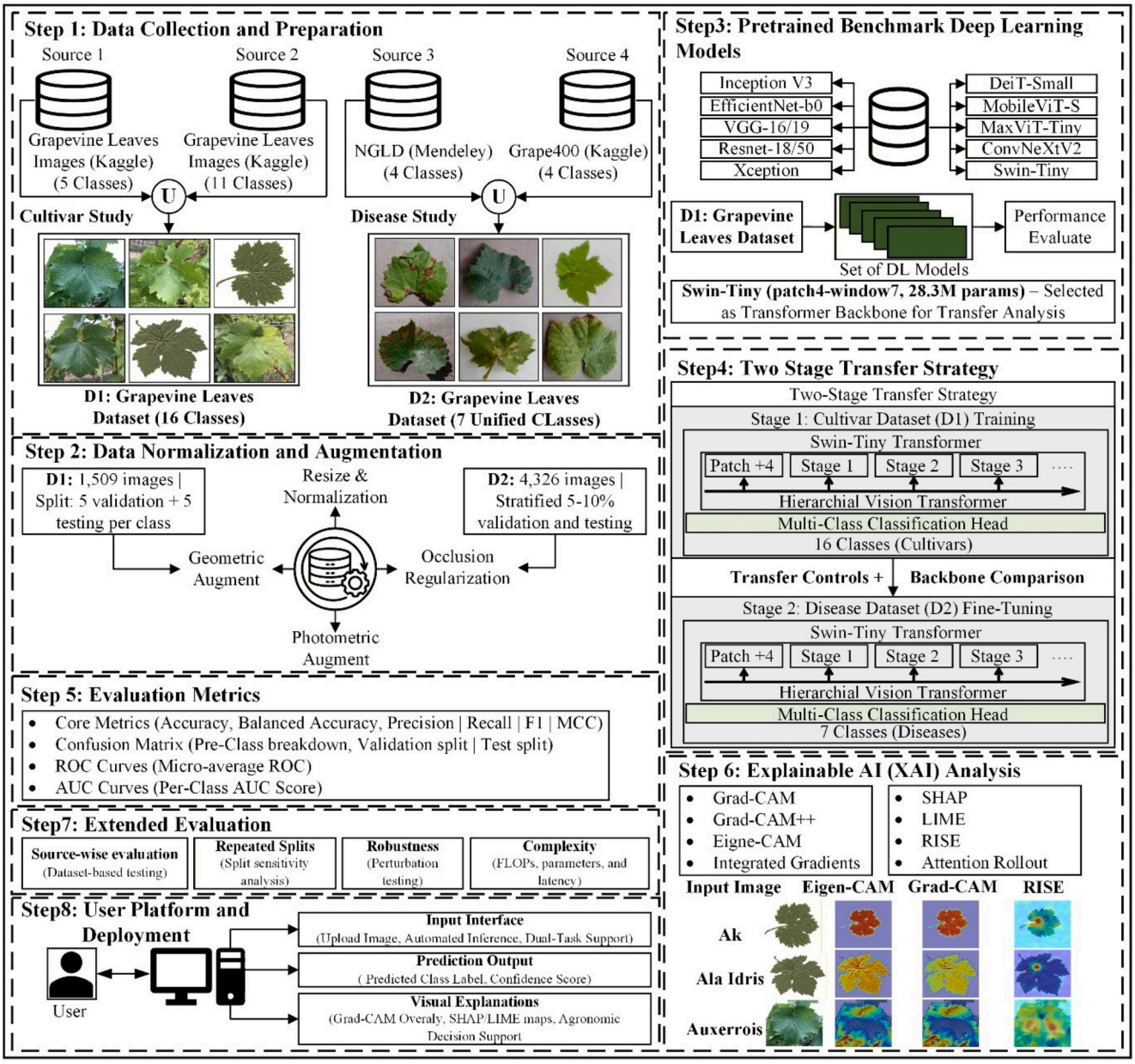


Fig. 1. End-to-end pipeline of the proposed unified deep learning framework, starting from cultivar and disease dataset construction, followed by preprocessing and augmentation, pretrained backbone benchmarking, Swin-Tiny selection as the transformer backbone for two-stage transfer analysis, transfer validation controls, backbone comparison, extended evaluation, post-hoc XAI analysis, and deployment through a lightweight image-to-decision platform.

Table 1
Cultivar dataset sources and composition (combined to 16 classes).

Source dataset	Original labels (as provided)	Total images (verified)	Train	Validation	Test	Image properties (reported)	Notes for unification
Grapevine Leaves Image Dataset (Kaggle) [40,41]	Ak; Ala_Idris; Büzgülü (Buzgulu); Dimnit; Nazli	500	–	–	–	RGB leaf images (as provided on Kaggle)	Label names normalised to match folder names (e.g., Büzgülü → Buzgulu).
Grapevine Leaves (Kaggle; 11 *Vitis vinifera* varieties) [42]	Auxerrois; Cabernet Franc; Cabernet Sauvignon; Chardonnay; Merlot; Muller Thurgau; Pinot Noir; Riesling; Sauvignon Blanc; Syrah; Tempranillo	1009	–	–	–	Leaf photos of 11 *Vitis vinifera* varieties (as provided on Kaggle)	Used to expand cultivar coverage; labels were standardised according to the unified class names.
Combined leaf-only set (this work)	16 cultivars (union of both sources)	1509	1349	80	80	Unified label space	Split performed after unification; validation and test fixed to 5 images per cultivar (80 each).

Table 2
Class-wise distribution of the grape leaf-only cultivar dataset across training, validation, and test splits (1509 images over 16 cultivars).

Cultivar (Leaf only)	Train	Validation	Test	Total
Ak	90	5	5	100
Ala_Idris	90	5	5	100
Auxerrois	78	5	5	88
Buzgulu	90	5	5	100
Cabernet Franc	50	5	5	60
Cabernet Sauvignon	101	5	5	111
Chardonnay	94	5	5	104
Dimnit	90	5	5	100
Merlot	51	5	5	61
Muller Thurgau	111	5	5	121
Nazli	90	5	5	100
Pinot Noir	49	5	5	59
Riesling	106	5	5	116
Sauvignon Blanc	90	5	5	100
Syrah	109	5	5	119
Tempranillo	60	5	5	70
Total	1349	80	80	1509

and test sets. The repeated-split experiment used the same combined 16-class cultivar dataset, preprocessing and augmentation pipeline, model-training configuration, and evaluation metrics. However, it used five alternative stratified random train/validation/test partitions while preserving five validation and five test images per cultivar, where possible. Therefore, the fixed single-split results and repeated-split results are comparable in terms of dataset, training recipe, and metrics, but they do not evaluate the same test images.

2.1.2. Disease dataset construction (7 classes)

For disease recognition, we combined two datasets and then unified their taxonomies into a single 7-class disease benchmark. The first dataset (NGLD) [43] contains 2726 high-quality table-grape leaf images captured by mobile phones at 1024 × 1024 resolution (150 DPI) and organised into Healthy and three disease categories: Downy Mildew, Powdery Mildew, and Bacterial Rot, as mentioned in Tables 3 and 4.

The second dataset (Grape400) [44,45] contains 1600 RGB images grouped into four categories: Black-measles, Black-rot, Leaf-blight, and Healthy-leaf. We merged the two datasets by (i) harmonising the directory structure and labels, and (ii) consolidating Healthy images from both sources into a single Healthy class. This yields 7 final labels: Healthy, Downy Mildew, Powdery Mildew, Bacterial Rot, Black Measles, Black Rot, and Leaf Blight. The final train/validation/test split was stratified by class while accounting for the stronger class imbalance in the disease dataset. For the lowest-frequency class, Bacterial Rot, only 5% of samples were allocated to validation and 5% to testing to preserve enough images for training. For all other disease classes, approximately 10% of images were allocated to validation and 10% to testing, with the remaining images used for training. This split differs from the cultivar split because the disease dataset has a larger total sample size but stronger class imbalance, especially for Bacterial Rot. The split was therefore designed to maintain class representation in validation and test sets while avoiding excessive reduction of the minority-class training data. Nevertheless, the very limited validation and test support for Bacterial Rot means that its class-specific scores should be interpreted with caution.

Table 4
Class-wise distribution of the unified grape leaf dataset across the training, validation, and test splits (total 4326 images over 7 classes).

Class	Train	Validation	Test	Total
Bacterial Rot	90	5	5	100
Black Measles	320	40	40	400
Black Rot	320	40	40	400
Downy Mildew	772	97	97	966
Healthy Leaves	1324	165	165	1654
Leaf Blight	320	40	40	400
Powdery Mildew	324	41	41	406
Total	3470	428	428	4326

2.1.3. Source-domain considerations and split rationale

Because both benchmarks were assembled from multiple public sources, source-related variation may influence model learning and evaluation. Images from different datasets can differ in background, illumination, acquisition device, resolution, leaf placement, and capture style. These factors may introduce source-domain cues that are not purely biological. In the cultivar benchmark, the dataset combines a controlled-acquisition grapevine leaf dataset with a second dataset containing natural leaf photographs. In the disease benchmark, images were combined from NGLD and Grape400, which differ in their acquisition sources and disease-label compositions. Although labels were harmonised and all images were resized and normalised using the same preprocessing pipeline, some source-specific visual characteristics may remain.

To reduce this risk, we applied a unified preprocessing and augmentation pipeline that included geometric, illumination, blur, noise, compression, and occlusion transformations. These augmentations were intended to discourage reliance on acquisition-specific artefacts and encourage the model to learn leaf morphology and disease-relevant visual cues. However, augmentation does not fully replace direct source-domain evaluation. Therefore, the reported results should be interpreted as performance on the constructed unified benchmarks rather than as definitive evidence of source-independent generalisation. In the revised experiments, we therefore include source-wise disease evaluation and a repeated-cultivar split analysis to partially assess source-domain effects and split sensitivity. However, full cross-source validation and external field testing remain future work because the disease sources have limited label overlap and some classes are source-specific.

2.2. Preprocessing and augmentation

All images were resized to the network input size (224 × 224) and normalised using ImageNet statistics (mean and standard deviation) to match the pretraining distribution, as described in Supplementary Table S2. To improve robustness to simulated imaging variability (pose,

Table 3
Disease dataset sources and composition (combined into 7 classes).

Source dataset	Original labels (as provided)	Total images (verified)	Image properties (reported)	Notes for unification
NGLD (Mendeley) [43]	Healthy Leaves; Downey/Downy Mildew; Powdery Mildew; Bacterial Rot	2726 (Healthy 1254, Downy 966, Powdery 406, Bacterial Rot 100)	1024 × 1024; 150 DPI; mobile-phone capture (as reported in Mendeley)	Kept 3 diseases + Healthy; mapped labels to unified class names (e.g., "Downey/Downy Mildew" → Downy Mildew).
Grape400 (Kaggle "2-Diseases") [44,45]	HealthyGrapes; BlackMeasles; BlackRot; LeafBlight	1600 (400 per class)	RGB images; 4-class set (as reported by the dataset source)	Kept 3 diseases (Black Measles, Black Rot, Leaf Blight); merged HealthyGrapes → Healthy Leaves.
Combined (this work)	7 unified classes (Healthy +6 diseases)	4326	Unified label space	Final split: Bacterial Rot = 90/5/5 (train/val/test = 5%/5%). All other classes use ~80/10/10 (val/test ≈ 10% each).

scale, illumination, and occlusions), we applied a heavy, structured augmentation pipeline during training, so the model learns to ignore spurious capture artefacts and focus on biologically meaningful leaf cues. The augmentation pipeline begins with a wide random-resized crop (scale 0.45–1.0, aspect ratio 0.70–1.40) to enforce partial-view and scale invariance, then applies geometric invariances (horizontal flip $p = 0.5$, vertical flip $p = 0.35$, 90° rotations $p = 0.25$). Next, a strong geometric-distortion block ($p = 0.85$) chooses one of: affine (scale 0.85–1.20, translate ≤8%, rotate ±35°, shear ±10°), perspective warp, optical distortion, or elastic transform simulating leaf bending, viewpoint change, and lens warping. A photometric block ($p = 0.90$) then applies a single colour/illumination transform (ColorJitter, HSV shifts, brightness/contrast, RGB shifts, random gamma, or CLAHE) to mimic variations in sunlight and camera exposure. We optionally add blur (Gaussian / motion/median, p ≈ 0.35) and a corruption block (Gaussian or ISO-like noise, or JPEG compression, p ≈ 0.45) to simulate motion blur, sensor noise and transmission artefacts. To increase robustness to occlusions and missing regions, an occlusion regularisation block (p ≈ 0.65) applies either coarse dropout (with many small/medium holes) or grid dropout. Finally, images are normalised with ImageNet mean/std. and converted to tensors. Validation and test images were only resized and normalised to preserve consistency in evaluation.

Textual flow of the augmentation stage (applied sequentially during training):

1. Resize → RandomResizedCrop (224 × 224, scale 0.45–1.0; ratio 0.70–1.40)
2. HorizontalFlip ($p = 0.5$) → VerticalFlip ($p = 0.35$) → RandomRotate90 ($p = 0.25$)
3. Strong geometric distortions (OneOf, $p = 0.85$): Affine | Perspective | OpticalDistortion | ElasticTransform
4. Photometric changes (OneOf, $p = 0.90$): ColorJitter | HueSaturationValue | RandomBrightnessContrast | RGBShift | RandomGamma | CLAHE
5. Blur (OneOf, p ≈ 0.35): GaussianBlur | MotionBlur | MedianBlur
6. Corruption (OneOf, p ≈ 0.45): GaussNoise | ISONoise | ImageCompression (JPEG)
7. Occlusion regularisation (OneOf, p ≈ 0.65): CoarseDropout | GridDropout
8. Normalise (ImageNet mean/std) → ToTensor

This structured, high-probability "one-of" design preserves diversity while keeping each training sample realistic: geometric changes and photometric shifts enforce invariance to viewpoint and illumination; blur/noise and JPEG simulate capture artefacts; and dropout-style occlusions prepare the model for partial leaf views in the field.

To examine the influence of the augmentation strategy, we performed a brief ablation using the selected Swin-Tiny model on the fixed cultivar split. All ablation settings used the same dataset split, input resolution, ImageNet normalisation, optimiser, learning-rate schedule, batch size, epoch budget, early-stopping rule, and evaluation metrics. Only the training-time augmentation policy was changed. Four settings were compared: minimal preprocessing only, geometric augmentation only, geometric plus photometric augmentation, and the full augmentation pipeline including blur, noise, compression, and occlusion regularisation. Validation and test images were not augmented and were only resized and normalised, as in the main experiments.

2.3. Candidate backbones and model selection strategy

We trained and compared multiple established and recent CNN, transformer, and hybrid backbones using the same training protocol on the 16-class cultivar dataset. The initial comparison included Xception [46], VGG16 [47], VGG19 [47], InceptionV3 [48], EfficientNetB0 [49], ResNet18 and 50 [50], and Swin Transformer (Tiny) [9], as mentioned in Table 5. To address the need for comparison with newer methods, the experiments further included MobileViT-S [51], DeiT-Small [52], ConvNeXt-Tiny [53], ConvNeXtV2-Tiny [54], EfficientNetV2-S [55], and MaxViT-Tiny [56]. In the initial Swin-based framework, Swin-Tiny was selected as the main backbone because it provides a hierarchical shifted-window attention mechanism, strong benchmark performance, compatibility with transformer-specific explanations, and moderate computational complexity. No architectural modification was made to the Swin-Tiny backbone; the proposed contribution lies in how this existing backbone is selected, trained, transferred from cultivar recognition to disease classification, interpreted using complementary XAI methods, and integrated into a decision-support workflow.

Table 5
Compared model families and approximate capacity (parameter counts), where M is Millions.

Backbone	Family	Approx. parameters (M)
Swin-Tiny (swin_tiny_patch4_window7_224)	Vision Transformer	27.53
MobileViT-S (mobilevit_s.cvnets_in1k)	Hybrid CNN-Transformer	4.95
DeiT-Small (deit_small_patch16_224.fb_in1k)	Vision Transformer	21.67
ConvNeXt-Tiny (convnext_tiny.fb_in1k)	Modern CNN	27.83
ConvNeXtV2-Tiny (convnextv2_tiny.fcmae_ft_in1k)	Modern CNN	27.88
MaxViT-Tiny (maxvit_tiny_tf_224.in1k)	Hybrid CNN-Transformer	30.41
EfficientNetV2-S (tf_efficientnetv2_s.in1k)	CNN	20.20
Xception (xception41)	CNN	24.95
InceptionV3	CNN	21.82
ResNet18	CNN	11.18
ResNet50	CNN	23.54
EfficientNetB0	CNN	4.03
VGG16	CNN	134.33
VGG19	CNN	139.64

2.4. Rationale for using Swin transformer

The Swin Transformer was selected because its architectural properties are well aligned with the visual characteristics of grapevine leaf analysis. Grapevine cultivar recognition depends on both local and global morphological cues, including venation texture, petiole insertion, leaf contour, lobing, serration, and overall leaf geometry. Conventional CNNs are effective at extracting local texture patterns, but their spatial context is accumulated gradually through stacked convolutional layers. In contrast, Swin Transformer combines local window-based self-attention with shifted-window connections, allowing the model to capture fine-grained local patterns while progressively integrating broader spatial dependencies across the leaf surface. This is particularly relevant for cultivar recognition, where discriminative information may be distributed across distant anatomical regions rather than concentrated in a single local patch.

Swin Transformer also provides hierarchical feature representations through patch embedding and patch merging, making it suitable for leaf images with multi-scale visual cues. Early stages can encode low-level features such as edges, vein texture, and colour variation. In contrast, deeper stages can capture higher-level structural patterns such as global shape, venation arrangement, and symptom distribution. This hierarchical design is advantageous for the proposed two-stage framework because the first stage requires learning cultivar-specific morphology. In contrast, the second stage requires adapting these representations to disease-specific visual symptoms such as lesions, discolouration, necrotic patches, and texture disruption.

The use of Swin-Tiny in the proposed framework was motivated by its architectural suitability for hierarchical leaf representation learning, its moderate computational scale, and its compatibility with

transformer-specific explanation methods. After expanded benchmarking with newer architectures, Swin-Tiny was not the highest-performing standalone cultivar classifier; however, it remained a competitive transformer-based backbone and provided a suitable basis for analysing an explainable Swin-based cultivar-to-disease transfer-learning workflow. Therefore, Swin-Tiny is used here as the selected backbone for the proposed framework rather than as a claim of universal superiority over all modern alternatives.

2.5. Training protocol

All models were initialised from ImageNet-pretrained weights and trained in a supervised manner using cross-entropy loss with label smoothing, as described in Table 6. Optimisation used AdamW [57] with weight decay, and the learning rate was scheduled using cosine annealing over the configured epoch budget. Early stopping was applied based on validation performance to prevent overfitting. We report performance using accuracy, balanced accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), together with confusion matrices and multiclass ROC/AUC curves. Balanced accuracy and MCC were emphasised because they are more informative when class imbalance and small per-class test support are present.

2.6. Final model: Swin transformer training and fine-tuning on the disease dataset

After a model comparison on the cultivar benchmark, we selected the Swin Transformer (Tiny) as the final backbone and trained it on the 16-class cultivar dataset using the pipeline described above. For disease recognition, we then fine-tuned the Swin Transformer on the unified 7-class disease dataset by replacing the classification head to match the new label space and retraining with the disease-specific split rules (smallest class: 5%/5% val/test; other classes: 10%/10%). This two-stage strategy was designed to evaluate whether cultivar-pretrained leaf representations can be adapted to symptom-driven disease classification. To directly test this assumption, transfer validation controls were included by comparing the proposed two-stage Swin-Tiny pipeline against direct ImageNet-pretrained disease fine-tuning, random initialisation, and frozen cultivar-feature transfer. In addition, multi-backbone, disease-stage comparisons were performed to evaluate whether the effect of cultivar-to-disease transfer was architecture-dependent rather than specific to Swin-Tiny. Accordingly, the two-stage setting was treated as a transfer-analysis framework rather than as an assumed performance-improvement strategy; its value was assessed only after comparison with direct disease-only fine-tuning and other transfer controls.

2.7. Experimental environment and evaluation metrics

All experiments were conducted on a workstation running Ubuntu 25.10 (64-bit) with an Intel® Core™ i9-10900KF CPU, 128 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. The system storage capacity was 3.0 TB. The RTX 3090 provides 24 GB of GDDR6X memory, enabling stable mixed-precision training and large mini-batch processing for transformer backbones.

Table 6
Core training configuration used in this work.

Component	Setting
Input resolution	224 × 224
Optimizer	AdamW
Initial learning rate	$3e^{-4}$
Weight decay	$1e^{-4}$
Loss	Cross-entropy + label smoothing (0.10)
LR schedule	Cosine annealing
Batch size	32
Epoch budget	100 (with early stopping)
Early stopping	Patience = 30 (validation criterion)
Mixed precision	Enabled (AMP)

2.7.1. Performance measures and definitions

Let C be the number of classes, and let TPc, FPc, FNc, and TNc denote true positives, false positives, false negatives, and true negatives for class c. We report standard classification measures together with imbalance-aware metrics.

Accuracy is defined as in Eq. (1):

$$Accuracy = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{1}$$

Precision and Recall for class c are defined in Eqs. (2) and (3):

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{2}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{3}$$

F1-score for class c is the harmonic mean of precision and recall as defined in Eq. (4):

$$F1_c = \frac{2Precision_c Recall_c}{Precision_c + Recall_c} \tag{4}$$

Because class frequencies may be imbalanced, we report Balanced Accuracy (BAcc), defined in Eq. (5) as the mean recall across classes:

$$BAcc = \frac{1}{C}\sum_{c=1}^{C} Recall_c = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{5}$$

For probabilistic outputs, we report multiclass ROC curves and AUC using one-vs-rest binarisation, including a micro-average ROC computed by aggregating all class decisions.

3. Results and discussion

3.1. Model comparison and selection

A total of fourteen deep learning architectures were evaluated for grapevine cultivar classification, covering classical CNNs, residual CNNs, efficiency-oriented CNNs, modern CNNs, vision transformers, and recent hybrid CNN-Transformer architectures. The evaluated models included VGG16, VGG19, ResNet18, ResNet50, Xception, InceptionV3, EfficientNetB0, EfficientNetV2-S, ConvNeXt-Tiny, ConvNeXtV2-Tiny, Swin-Tiny, DeiT-Small, MobileViT-S, and MaxViT-Tiny. As summarised in Table 7, all models were trained and evaluated using the same cultivar split, preprocessing, augmentation, optimiser, learning-rate schedule, and evaluation metrics. Fig. 2 presents the training and evaluation curves for the selected Swin-Tiny pipeline. Performance was assessed using balanced accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), with an emphasis on test-set balanced accuracy and MCC, given the relatively small cultivar dataset and the fixed class support of 5 test images per cultivar.

The expanded comparison shows that MobileViT-S achieved the highest balanced accuracy on the cultivar-stage test (91.25%), with a test F1-score of 89.88% and an MCC of 0.9086. MaxViT-Tiny followed with 90.00% test balanced accuracy, while ConvNeXtV2-Tiny achieved 88.75%. Swin-Tiny, ConvNeXt-Tiny, and EfficientNetV2-S each achieved 85.00% test balanced accuracy. Among the original CNN baselines, EfficientNet-B0, Inception v3, and ResNet-50 each reached 83.75% test balanced accuracy, while Xception reached 82.50%.

Swin-Tiny achieved 85.00% test balanced accuracy, 83.44% test F1-score, and 0.8428 MCC. Therefore, after expanded benchmarking, Swin-Tiny is not presented as the strongest standalone cultivar classifier

Table 7
Expanded cultivar-stage backbone comparison on the fixed cultivar split used for detailed Swin-Tiny pipeline analysis, evaluated using balanced accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC).

Model	Balanced Acc % (Val/Test)	Precision % (Val/Test)	Recall % (Val/Test)	F1% (Val/Test)	MCC (Val/Test)
MobileViT-S	87.50 / 91.25	89.50 / 92.80	87.50 / 91.25	86.80 / 89.88	0.8700 / 0.9086
MaxViT-Tiny	87.50 / 90.00	89.20 / 92.85	87.50 / 90.00	86.70 / 89.63	0.8700 / 0.8959
ConvNeXtV2-Tiny	88.75 / 88.75	90.10 / 91.50	88.75 / 88.75	87.90 / 87.44	0.8830 / 0.8828
Inception v3	80.00 / 83.75	81.24 / 86.12	80.00 / 83.75	78.45 / 82.67	0.7923 / 0.8351
ResNet-50	80.00 / 83.75	84.21 / 86.34	80.00 / 83.75	79.34 / 82.78	0.7956 / 0.8362
Swin-Tiny (patch4-window7)	81.25 / 85.00	85.32 / 87.17	81.25 / 85.00	80.21 / 83.44	0.8040 / 0.8428
ConvNeXt-Tiny	88.75 / 85.00	90.00 / 81.67	88.75 / 85.00	88.20 / 82.51	0.8830 / 0.8434
EfficientNetV2-S	86.25 / 85.00	87.80 / 82.07	86.25 / 85.00	85.50 / 81.62	0.8570 / 0.8445
EfficientNet-B0	80.50 / 83.75	84.80 / 82.31	80.50 / 83.75	79.60 / 81.67	0.7980 / 0.8318
Xception	80.50 / 82.50	84.90 / 81.81	80.50 / 82.50	79.70 / 80.01	0.7990 / 0.8188
DeiT-Small	71.25 / 71.25	72.50 / 68.68	71.25 / 71.25	68.90 / 67.19	0.7000 / 0.6998
ResNet-18	75.00 / 72.50	83.56 / 74.67	75.00 / 72.50	72.43 / 70.14	0.7409 / 0.7136
VGG16	67.50 / 70.00	68.42 / 67.98	67.50 / 70.00	63.30 / 66.43	0.6650 / 0.6898
VGG19	52.50 / 55.00	52.01 / 49.51	52.50 / 55.00	46.41 / 49.15	0.5088 / 0.5368

Note: This table reports the single fixed-split results used for the detailed Swin-Tiny pipeline, confusion matrix, and explainability analyses. Because the cultivar test set contains only five images per class, these fixed-split results should be interpreted cautiously. Repeated-split stability results are reported separately in Supplementary Table S8 and provide a more stable estimate of model-ranking performance.

among all evaluated models. Instead, it is retained as the selected backbone for the proposed explainable Swin-based two-stage pipeline because it provides a hierarchical shifted-window attention mechanism, moderate parameter scale, compatibility with transformer-specific explanation methods, and strong disease-stage performance after fine-tuning. This distinction is important: the aim of the framework is not to propose the best standalone cultivar classifier, but to evaluate an explainable cultivar-to-disease transfer-learning workflow using a standard Swin-Tiny backbone while transparently benchmarking it against established and recent alternatives.

Among the non-transformer models, ConvNeXtV2-Tiny achieved the strongest performance, reaching 88.75% test balanced accuracy and 0.8828 MCC, demonstrating that modernised convolutional designs can effectively capture cultivar-specific leaf structure. EfficientNet-B0, Inception v3, and ResNet-50 each achieved 83.75% test balanced accuracy, with MCC values of 0.8318, 0.8351, and 0.8362, respectively. These results confirm that efficient scaling, multi-scale convolutional feature extraction, and residual learning remain useful for recognising subtle differences in leaf shape, venation, and texture. However, they remained below the strongest newly added hybrid and modern backbones under this split.

The efficiency-oriented and lightweight models showed different trade-offs. MobileViT-S achieved the highest test performance while using only 4.95 M parameters, suggesting a favourable accuracy-capacity balance for cultivar recognition. EfficientNetB0 also remained competitive with 83.75% test balanced accuracy using 4.03 M parameters, while EfficientNetV2-S reached 85.00% test balanced accuracy with 20.20 M parameters. These results indicate that model size alone does not determine performance; rather, architectural design and the ability to combine local and broader spatial cues are important for this fine-grained task.

The weakest performance was observed for VGG16 and VGG19, with test balanced accuracies of 70.00% and 55.00%, respectively. Their lower performance is consistent with the limitations of classical deep CNNs that lack residual connections, modern normalisation strategies, and efficient feature-reuse mechanisms. The poorer performance of VGG19 compared with VGG16 suggests that simply increasing depth does not improve performance in limited-data fine-grained cultivar recognition and may increase the risk of overfitting.

Overall, the expanded benchmark demonstrates that modern and hybrid architectures outperform classical CNNs for grapevine cultivar recognition, and that Swin-Tiny provides a competitive transformer-based baseline rather than the top standalone cultivar classifier. For the subsequent disease-stage experiments, Swin-Tiny was retained because the objective of this study is to analyse and explain an explainable Swin-based two-stage transfer-learning framework, not to claim universal superiority of Swin-Tiny over all modern alternatives. The broader comparison in Table 7 provides context for this choice and addresses the need for comparison with newer methods.

The results in Table 7 should be interpreted as performance on the fixed cultivar split used for the main Swin-Tiny pipeline, confusion-matrix analysis, and subsequent explainability experiments. They should not be treated as the sole basis for ranking stable models because the cultivar test set contains only 5 images per class. In this setting, a single test error in one cultivar reduces that class's recall by 20%, corresponding to a 1.25-percentage-point change in overall balanced accuracy across 16 classes. Thus, the 85.00% test-balanced accuracy of Swin-Tiny reflects performance on this particular fixed and relatively difficult split. In contrast, repeated-split results are needed to assess the stability of the model's ranking.

3.2. Grape leaf disease classification

The two-stage Swin-Tiny model was evaluated on the seven-class grape leaf disease classification task covering Bacterial Rot, Black Measles, Black Rot, Downy Mildew, Healthy Leaves, Leaf Blight, and Powdery Mildew. Under the same disease split and evaluation protocol, this model achieved validation and test accuracies of 99.07% and 99.53%, respectively, while the corresponding balanced accuracies were 99.01% and 99.71% (Fig. 3, Tables 8 and 9). Thus, the value of 99.53% refers to overall test accuracy (weighted recall), whereas 99.71% refers to balanced accuracy, calculated as the mean class-wise recall.

These results demonstrate strong benchmark-level disease classification by Swin-Tiny after two-stage fine-tuning. However, they should not be interpreted as evidence that cultivar-stage pretraining was the main reason for the high disease performance. As shown later in the direct transfer-control analysis, the disease-only ImageNet-pretrained Swin-Tiny model performed comparably, achieving a slightly higher test balanced accuracy under the same split. Therefore, the disease-stage results are interpreted as strong performance on the Swin-Tiny benchmark. At the same time, the benefit of cultivar-to-disease transfer is evaluated separately in Section 3.3, “Transfer-control analysis of cultivar-to-disease pretraining”.

Downy Mildew, which had the largest class support at 97 samples, performed with near-perfect scores across both splits. Its precision improved from 98.96% on validation to a perfect 100.00% on the test set, while recall remained stable at 97.94% across both splits, yielding F1-scores of 98.45% and 98.96% on validation and test, respectively (Table 8). The consistent recall across both splits suggests the model occasionally misses a small number of Downy Mildew cases, likely due to visual similarity to other foliar conditions. However, its overall performance remains strong under the present benchmark setting, given the class prevalence in the dataset.

Powdery Mildew proved the most challenging case for the model,

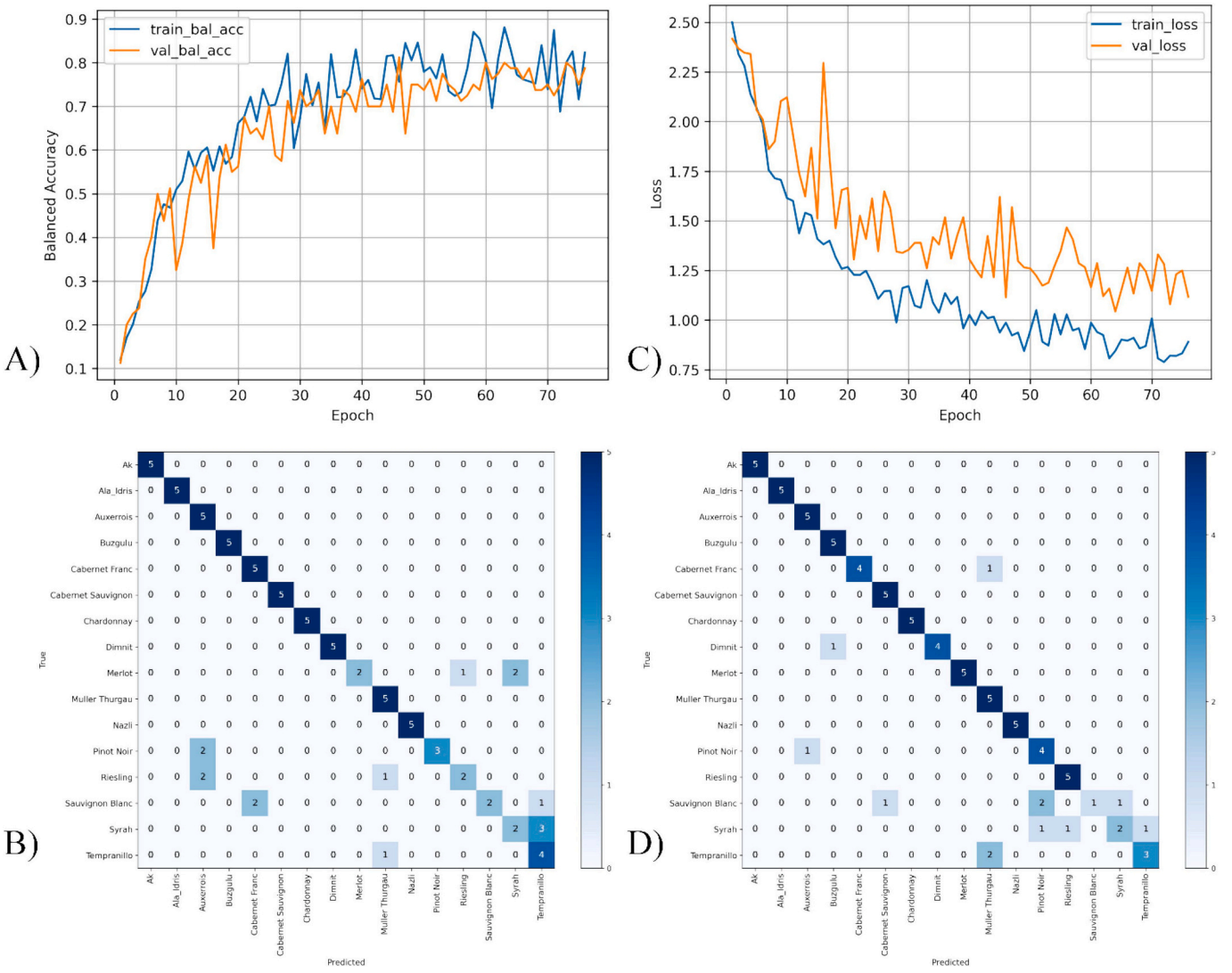


Fig. 2. Cultivar-stage Swin-Tiny training and evaluation results: (A) balanced accuracy curve, (B) loss curve, (C) validation confusion matrix, and (D) test confusion matrix for the 16-class grapevine cultivar classification task.

yielding the lowest scores in both splits. On the validation set, precision, recall, and F1 were all 95.12%, indicating symmetric confusion in the model, with both true Powdery Mildew cases and other conditions occasionally misclassified as Powdery Mildew at equal rates. On the test set, recall improved to 100.00%, meaning all Powdery Mildew samples were correctly identified, while precision remained at 95.35%, indicating a small number of false positives. This resulted in a test F1-score of 97.62%, a meaningful improvement over the validation score of 95.12% (Tables 8 and 9). Powdery Mildew's characteristic white powdery coating on leaf surfaces can sometimes overlap in appearance with early-stage Healthy Leaves or Downy Mildew under certain imaging conditions, which may explain the residual confusion.

Bacterial Rot, despite having by far the smallest support of only 5 samples, showed an interesting pattern across the two splits. On the validation set, precision was 83.33%, recall was 100.00%, yielding an F1-score of 90.91%, indicating that all true Bacterial Rot cases were identified, but one sample from another class was incorrectly labelled as Bacterial Rot. On the test set, the model achieved perfect scores of 100.00% across all three metrics, fully resolving the earlier false positive, as confirmed in Tables 8 and 9. While this test-set perfection is encouraging, it should be interpreted with caution, given the very small support size; five samples are insufficient to draw statistically robust conclusions about per-class performance for this category.

At the aggregate level, the macro-averaged F1-score improved from 97.78% on validation to 99.51% on the test set, while the macro-averaged precision improved from 96.77% to 99.34% and macro-averaged recall from 99.01% to 99.71%. The weighted average F1 remained consistently high at 99.07% and 99.54% on the validation and test sets, respectively, as detailed in Tables 8 and 9. The macro average is particularly meaningful here as it treats all classes equally, regardless of support, and its high value confirms that the model performs well even on minority classes such as Bacterial Rot. The small gap between macro and weighted averages across both splits further suggests that the class imbalance, ranging from just 5 samples for Bacterial Rot to 165 for Healthy Leaves, did not meaningfully bias the model toward majority classes.

Overall, these results show that Swin-Tiny achieved very strong performance for automated grape leaf disease classification under the present benchmark and split protocol. However, the near-perfect validation and test scores should be interpreted with caution, as the dataset is compiled from public sources and does not constitute independent field validation. The model's performance suggests promise for agricultural disease-monitoring applications. Still, robustness and deployment readiness should be confirmed through external vineyard datasets, cross-device testing, and systematic evaluation under variable illumination, occlusion, background clutter, and mixed symptom stages.

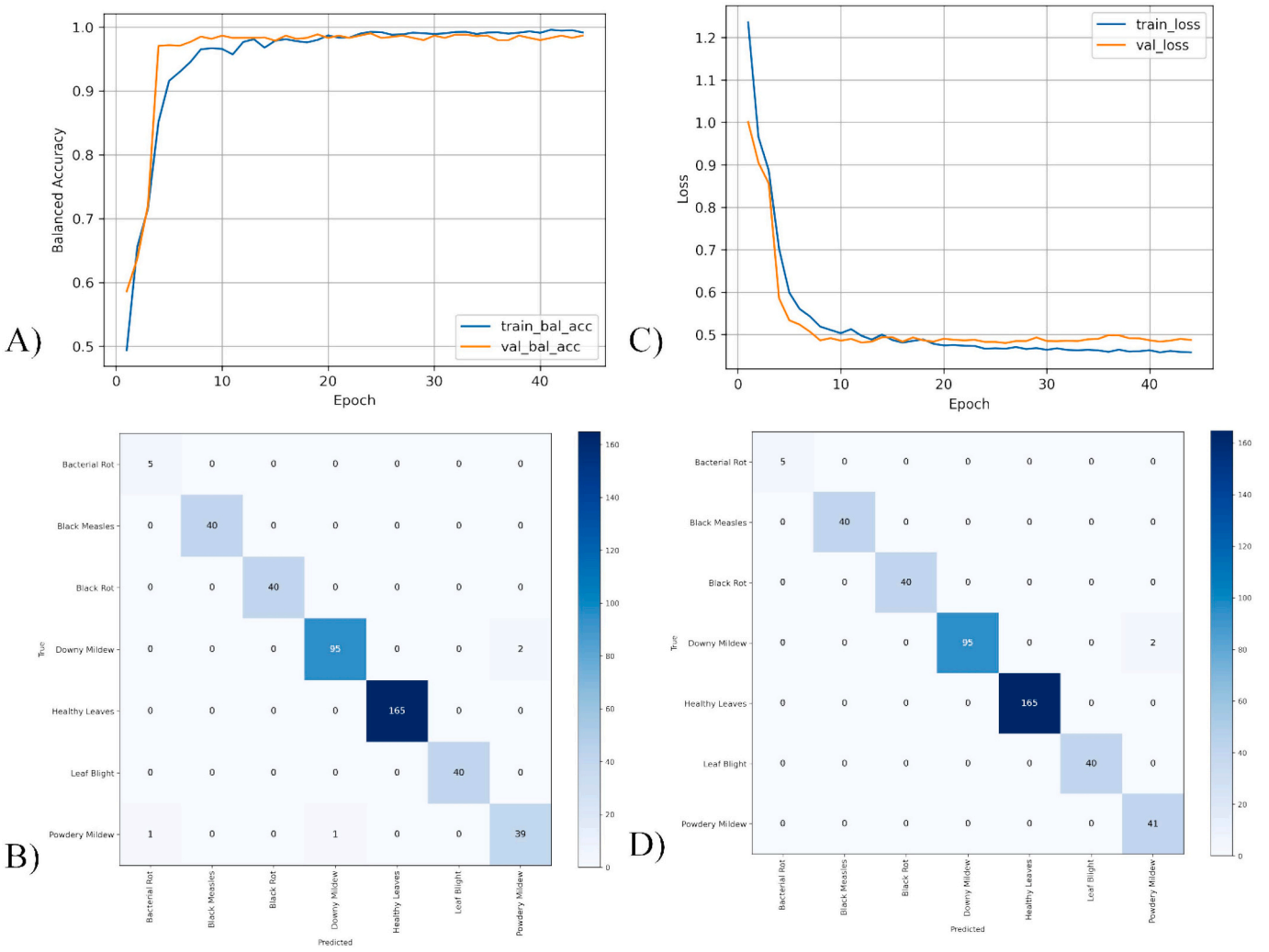


Fig. 3. Disease-stage Swin-Tiny fine-tuning and evaluation results: (A) balanced accuracy curve, (B) loss curve, (C) validation confusion matrix, and (D) test confusion matrix for the 7-class grape leaf disease classification task.

Table 8
Per-class classification performance (%) of the Swin-Tiny model on the grape leaf disease dataset for validation and test sets.

Class	Precision % (Val/Test)	Recall % (Val/Test)	F1-Score % (Val/Test)	Support
Bacterial Rot	83.33 / 100.00	100.00 / 100.00	90.91 / 100.00	5
Black Measles	100.00 / 100.00	100.00 / 100.00	100.00 / 100.00	40
Black Rot	100.00 / 100.00	100.00 / 100.00	100.00 / 100.00	40
Downy Mildew	98.96 / 100.00	97.94 / 97.94	98.45 / 98.96	97
Healthy Leaves	100.00 / 100.00	100.00 / 100.00	100.00 / 100.00	165
Leaf Blight	100.00 / 100.00	100.00 / 100.00	100.00 / 100.00	40
Powdery Mildew	95.12 / 95.35	95.12 / 100.00	95.12 / 97.62	41

Table 9
Aggregate classification performance (%) of the Swin-Tiny model on the grape leaf disease dataset.

Metric	Validation	Test	Support
Accuracy / Weighted recall	99.07	99.53	428
Balanced accuracy / Macro recall	99.01	99.71	428
Macro precision	96.77	99.34	428
Macro F1-score	97.78	99.51	428
Weighted precision	99.10	99.55	428
Weighted F1-score	99.07	99.54	428

3.3. Transfer-control analysis of cultivar-to-disease pretraining

To directly assess whether cultivar-stage pretraining contributes to disease classification, we compared the proposed Swin-Tiny cultivar-to-disease transfer pipeline with three additional training strategies: direct ImageNet-pretrained disease fine-tuning, random initialisation, and frozen cultivar-feature transfer. This comparison is important because high disease classification performance alone does not demonstrate that the improvement is specifically attributable to the cultivar pretraining stage. Therefore, all settings were evaluated on the same unified 7-class disease benchmark using balanced accuracy, macro-F1, weighted-F1, and Matthews Correlation Coefficient.

The proposed Swin-Tiny two-stage cultivar-to-disease model achieved very high disease performance, with 99.71% test balanced accuracy, 99.51% macro-F1, 99.54% weighted-F1, and an MCC of 0.9939 (Table 10). However, direct ImageNet-pretrained Swin-Tiny fine-tuning achieved 100.00% test balanced accuracy, 100.00% macro-F1, and an MCC of 1.0000 on the same test split. This result indicates that, for Swin-

Table 10
Direct Swin-Tiny transfer evidence.

Setting	Test BAcc %	Macro-F1%	Weighted-F1%	MCC
Swin-Tiny two-stage cultivar→disease	99.71	99.51	99.54	0.9939
Swin-Tiny disease-only ImageNet	100.00	100.00	100.00	1.0000
Swin-Tiny random initialisation	96.70	93.78	97.03	0.9604
Swin-Tiny frozen cultivar features	95.13	93.56	95.06	0.9363

Note: The two-stage Swin-Tiny result in this table uses the same disease split and evaluation protocol reported in Table 9. The test balanced accuracy is 99.71%, calculated as macro-averaged recall across the seven disease classes. The reported value of 99.53% elsewhere corresponds to overall accuracy (weighted recall), not to balanced accuracy.

Tiny on the current dataset split, cultivar pretraining did not yield a numerical improvement over direct disease-only fine-tuning. Nevertheless, the two-stage model clearly outperformed both random initialisation and frozen cultivar-feature transfer, which achieved test balanced accuracies of 96.70% and 95.13%, respectively. These findings suggest that pretrained representations and full fine-tuning are important for disease classification. At the same time, the additional benefit of cultivar-stage pretraining should be interpreted cautiously rather than assumed to be universal.

The comparison in Table 10 clarifies the role of the two-stage Swin-Tiny design. The two-stage cultivar-to-disease model achieved near-ceiling performance in disease classification. Still, the direct ImageNet-pretrained disease-only Swin-Tiny model achieved a slightly higher test balanced accuracy under the same split. This indicates that, specifically for Swin-Tiny, the high disease-classification performance cannot be attributed primarily to cultivar-stage pretraining. Instead, the two-stage design is better interpreted as a controlled experimental framework for analysing cultivar-to-disease transfer, fine-tuning behaviour, and explanation patterns. At the same time, the numerical performance benefit of cultivar pretraining should be considered limited in this setting.

To further examine whether the effect of cultivar-stage pretraining is specific to Swin-Tiny or generalises across architectures, we evaluated multiple backbones under both disease-only and two-stage training settings. As reported in Table 11, the effect of cultivar pretraining was architecture-dependent. ResNet50 improved from 98.45% to 100.00% test balanced accuracy, EfficientNetV2-S improved from 99.29% to 100.00%, and ConvNeXtV2-Tiny improved from 99.50% to 100.00%. ConvNeXt-Tiny showed no change, remaining at 99.85% under both settings. In contrast, Xception and MaxViT-Tiny showed small decreases after cultivar pretraining, although their performance remained very high.

These results show that cultivar-to-disease transfer learning is not uniformly superior across all architectures and splits. Instead, it provides a competitive and sometimes beneficial training strategy whose effectiveness depends on the backbone. This supports a more cautious interpretation of the proposed framework: cultivar-stage representation learning can provide useful transferable information for several architectures, but direct disease fine-tuning may perform equally well or better in some cases, especially when the disease dataset is visually separable and ImageNet-pretrained features are already strong. Therefore, the revised experiments directly address the transfer-learning question by showing that the proposed two-stage strategy is competitive and beneficial for some backbones, but not uniformly superior to direct disease-only fine-tuning.

Table 11
Multi-backbone effect of cultivar pretraining.

Backbone	Disease-only BAcc %	Two-stage BAcc %	Delta %
ResNet50	98.45	100.00	+1.55
EfficientNetV2-S	99.29	100.00	+0.71
ConvNeXtV2-Tiny	99.50	100.00	+0.50
ConvNeXt-Tiny	99.85	99.85	0.00
Xception	99.91	99.85	−0.06
MaxViT-Tiny	100.00	99.65	−0.35

Note: Swin-Tiny is reported separately in Table 10 because it was evaluated in more detail using two-stage fine-tuning, direct disease-only fine-tuning, random initialisation, and frozen cultivar-feature transfer. Delta indicates two-stage balanced accuracy minus disease-only balanced accuracy.

3.4. Interpretation of the main findings

This study frames grapevine leaf analysis as an agricultural information-processing problem in which cultivar recognition, disease classification, interpretability, and lightweight deployment are evaluated within a unified workflow. The two-stage cultivar-to-disease design should be interpreted primarily as an experimental framework for analysing whether cultivar-learned morphology transfers to disease recognition, rather than as a confirmed performance-improvement mechanism for Swin-Tiny.

The revised transfer-control experiments show that this distinction is important. Although the two-stage Swin-Tiny model achieved very high disease classification performance, direct ImageNet-pretrained, disease-only Swin-Tiny fine-tuning achieved slightly higher performance on the same split. Therefore, the high Swin-Tiny disease performance cannot be attributed mainly to cultivar-stage pretraining. Instead, it likely reflects a combination of strong ImageNet-pretrained visual representations, the visual separability of disease symptoms in the constructed benchmark, the effectiveness of full fine-tuning, and the representational capacity of the Swin-Tiny backbone. The contribution of the two-stage design is therefore analytical and application-oriented: it enables controlled examination of cultivar-to-disease transfer, comparison with direct disease-only training, and interpretation of shifts in attention from morphology-related regions to symptom-related regions.

Turning to the grape leaf disease classification task, the two-stage Swin-Tiny model reached 99.07% validation accuracy and 99.53% test accuracy, with corresponding balanced accuracies of 99.01% and 99.71%. These near-ceiling results indicate strong benchmark performance under the current split. Still, the transfer-control results show they should not be interpreted as proof that cultivar-stage pretraining improved Swin-Tiny's disease recognition. Direct disease-only Swin-Tiny fine-tuning performed comparably and even slightly better, reaching 100.00% test balanced accuracy. This suggests that the disease classes in the current benchmark are highly visually separable and that ImageNet-pretrained features are already sufficient for near-perfect disease classification under this split.

This architectural suitability explains why the Swin Transformer was preferred over the CNN alternatives in the proposed framework. Grapevine leaf analysis is not purely a local texture-recognition problem; instead, cultivar and disease decisions often depend on relationships among spatially separated structures, such as the petiole region, primary and secondary veins, leaf margins, lobes, and distributed symptom areas. The shifted-window mechanism allows neighbouring windows to exchange information across layers, reducing the locality limitation of fixed-window attention while retaining computational efficiency. Therefore, Swin-Tiny offers a practical compromise between CNN-like locality, transformer-like contextual modelling, and computational feasibility. This balance is especially important in the present two-stage setting, where the model must first learn general leaf morphology and then transfer that representation to disease-specific classification.

The competitive performance of Inception v3 and ResNet-50, both reaching a test balanced accuracy of 83.75%, highlights that deep convolutional architectures with strong multi-scale and residual feature-extraction capabilities remain viable alternatives when transformer-

based models are computationally prohibitive. Inception v3's parallel convolutional branches at multiple kernel sizes enable it to capture both fine-grained texture and coarser morphological patterns simultaneously. At the same time, ResNet-50's residual learning framework enables effective gradient propagation across its 50 layers without degradation. The modest val-to-test decline observed in both models, however, suggests that their inductive biases toward local feature aggregation may limit their ability to fully encode the holistic leaf-level patterns that differentiate closely related grape varieties. Xception followed a similar trajectory, performing well on the validation set but declining more sharply on the test set, which may reflect a greater susceptibility to overfitting due to its exclusively depthwise separable convolution design when fine-tuned on a limited dataset.

EfficientNet-B0 demonstrated stable performance among the CNN-based models, improving from 80.50% validation balanced accuracy to 83.75% test balanced accuracy, with consistent test F1 and MCC values. This stability is a hallmark of EfficientNet's compound scaling strategy, which balances network depth, width, and input resolution in a principled manner to achieve efficient generalisation. While EfficientNet-B0 did not surpass Swin-Tiny, its consistency makes it a practical choice in resource-constrained environments where inference efficiency is prioritised alongside accuracy.

The comparatively weak performance of ResNet-18 underscores the role of model capacity in fine-grained classification. With only 18 layers and a significantly smaller parameter count than ResNet-50, the model lacks the representational depth needed to disentangle the subtle inter-class visual differences inherent in grape variety identification. The 11.25 percentage-point gap in test accuracy between ResNet-18 (72.50%) and ResNet-50 (83.75%) is substantial and directly attributable to this capacity limitation rather than to architectural design flaws, as ResNet-50, with the same residual framework but greater depth, performs among the top models. This finding is consistent with the broader literature on transfer learning for plant phenotyping, where deeper pre-trained backbones consistently outperform their shallower counterparts on fine-grained tasks.

The poor performance of the VGG family warrants careful consideration. VGG16 achieved a test accuracy of only 70.00%, while VGG19 performed significantly worse at 55.00%, which is counterintuitive given that greater depth is typically associated with improved representational capacity. This reversal is most plausibly explained by the compounding effect of overfitting in the absence of architectural regularisation mechanisms. VGG architectures were designed without residual connections, bottleneck layers, or depthwise separable convolutions, all of which serve to regularise feature learning and improve gradient flow in deeper networks. When fine-tuned on a small dataset, VGG19's additional layers likely amplify the overfitting already exhibited by VGG16, effectively memorising training-set-specific patterns that fail to generalise to the test set. These results strongly suggest that VGG-family models are ill-suited for grape variety classification without substantial data augmentation, dropout regularisation, or dataset expansion.

Turning to the grape leaf disease classification task, the two-stage Swin-Tiny model reached 99.07% validation accuracy and 99.53% test accuracy, with corresponding balanced accuracies of 99.01% and 99.71%. These near-ceiling results indicate strong benchmark performance under the present split. Still, the transfer-control results should not be interpreted as proof that cultivar-stage pretraining improved Swin-Tiny's disease recognition. Direct disease-only Swin-Tiny fine-tuning performed comparably and even slightly better, reaching 100.00% test balanced accuracy. This suggests that the disease classes in the current benchmark are highly visually separable and that ImageNet-pretrained features are already sufficient for near-perfect disease classification under this split.

The two classes that exhibited less-than-perfect performance, Downy Mildew and Powdery Mildew, are worth examining from an agronomic and image-analysis perspective. Downy Mildew's consistent recall of 97.94% across both splits, despite perfect precision on the test set, indicates that approximately 2% of true Downy Mildew cases are systematically misclassified. This is likely attributable to early-stage disease presentations, in which characteristic sporulation has not yet developed, causing affected leaves to appear visually similar to those in early Healthy or Bacterial Rot cases. Similarly, Powdery Mildew's precision below 1.0 on both splits suggests occasional false-positive classifications, in which other conditions are mistakenly attributed to Powdery Mildew, a confusion likely arising from the visual similarity between light-coloured fungal deposits and certain reflectance artefacts or early-stage Healthy leaf surfaces under variable imaging conditions. Despite these observations, both classes achieved F1-scores above 95.00% on the test set, indicating strong benchmark performance; however, practical diagnostic reliability should be assessed using independent field data before operational use.

The extremely small support for Bacterial Rot (5 samples) introduces important statistical caveats. While the model achieved perfect scores on the test set for this class, conclusions about its true generalisation capability for Bacterial Rot must be drawn cautiously. With only five test samples, a single misclassification would reduce recall to 0.80, and the current perfect score may partially reflect chance alignment between the model's learned features and this particular sample of five images.

From a broader methodological perspective, the results suggest several implications for agricultural AI research under benchmark conditions. First, Swin-Tiny performed strongly across the evaluated tasks, indicating that shifted-window transformer backbones can be effective for grapevine leaf image analysis. However, this should not be interpreted as a general claim that vision transformers will outperform CNNs across all vineyard conditions or datasets. Second, the very high disease-classification performance compared with the more moderate cultivar-recognition results suggests that disease symptoms in the present dataset are more visually separable than cultivar morphology. Third, the sensitivity of model performance to dataset size, source composition, and minority-class support, as evidenced by the Bacterial Rot sample-size concern and the source-domain discussion, reinforces the need for larger, more diverse, and externally validated grapevine datasets.

The use of multiple public datasets also introduces a potential source-domain effect. Although the datasets were harmonised into common label spaces and processed using the same resizing, normalisation, and augmentation pipeline, images originating from different sources may still contain dataset-specific cues related to background, lighting, camera characteristics, resolution, and leaf positioning. Consequently, part of the measured performance may reflect the model's adaptation to the constructed benchmark distribution, rather than to biological cultivar or disease cues. This is particularly relevant when some classes are strongly associated with specific source datasets. Therefore, the results should be interpreted as evidence of strong performance under the present unified benchmark and split protocol. At the same time, source-independent and field-level generalisation require further validation through source-wise testing, cross-source evaluation, and external vineyard datasets.

In summary, within the constructed benchmark and split protocol, Swin-Tiny provided a competitive transformer-based backbone for the proposed explainable two-stage framework, while expanded benchmarking showed that several recent architectures achieved higher standalone cultivar-recognition performance. These findings suggest that shifted-window transformer backbones are promising for grapevine leaf image analysis. Still, broader conclusions about field-level generalisation and operational precision agriculture deployment require additional validation across independent, multi-source, and in-vineyard datasets. However, the repeated-split analysis showed that this fixed-split result was not a stable estimate of average Swin-Tiny cultivar-recognition performance. Across five stratified random splits, Swin-Tiny achieved 96.50 $\pm$ 3.11% test balanced accuracy, indicating that a difficult fixed holdout subset likely caused the lower 85.00% value. Therefore, conclusions about cultivar-stage model ranking should rely

mainly on the repeated-split analysis, while the fixed split is retained for detailed pipeline reporting and explanation analysis.

Evidence from the broader literature supports this cautious interpretation. Ji et al. [31] reported 98.57% test accuracy for four-class grape-leaf disease classification on a PlantVillage hold-out set, while Ngugi et al. [34] showed that changing the unit of analysis from whole leaves to individual lesions could improve recognition by more than 15%. Huang et al. [37] reported a pronounced decline from 99.20% accuracy in laboratory imagery to 79.30% in natural-environment imagery. Together, these findings show that apparent performance depends strongly on task formulation, dataset composition, and acquisition conditions. Accordingly, the present study reports balanced accuracy, MCC, repeated splits, source-wise evaluation, perturbation tests, and explicit limitations rather than treating a single accuracy value as sufficient evidence of field readiness.

3.5. Cultivar recognition: Comparison with related studies

Direct numerical comparison across cultivar-recognition studies is inherently limited because datasets differ substantially in the number of cultivars (2 to 50), image modalities (leaf-only vs leaf+bunch+fruit), capture conditions (smartphone/field vs controlled), and splitting protocols (often not standardised or not fully specified), as detailed in Table 12. Many papers, therefore, report only test accuracy, which can be inflated by class imbalance, small test sets, or visually "easy" cultivar separations. To mitigate these issues, our study reports a unified internal benchmark across established and recent backbones, including the selected Swin-Tiny pipeline using balanced accuracy and MCC in addition to precision/recall/F1, which are better aligned with realistic cultivar recognition where varieties may be unevenly represented and visually similar. In this context, our Swin-Tiny test performance (balanced accuracy 0.8500; precision 87.17%; recall 85.00%; F1 83.44%; MCC 0.8428) should be interpreted as a metric-rich, transformer-based reference under a unified evaluation pipeline, rather than a claim of superiority over works that use different datasets. Importantly, the value of our comparison is that we provide a consistent internal benchmark across fourteen models and a transparent transformer baseline for cultivar recognition, addressing a gap in prior cultivar studies where transformer evaluation and robust, imbalance-aware reporting (e.g., MCC/balanced accuracy) are often missing or incomplete.

Liu et al. [32] reported 88.75% mAP, 84.88 FPS, and a 17.53 MB model for grape-bunch detection under dense occlusion. Because object-detection mAP is not numerically comparable with cultivar-classification accuracy, that study is not ranked against the entries in Table 12 and is used only to contextualise efficiency and unstructured vineyard imaging. No study identified in the reviewed literature evaluates leaf-based grape cultivar classification under a matching protocol, so no directly comparable numerical cultivar result is available.

3.6. Grape leaf disease recognition: Comparison with related studies

To keep the comparison methodologically valid, Table 13 is restricted to image-level disease-classification studies. Object-detection and microscopic spore-detection studies are discussed in the literature review but excluded from the numerical classification table because mAP and image-level accuracy/F1 measure different outcomes. Even among classification studies, datasets differ in crop, class count, acquisition conditions, and split protocol. The present Swin-Tiny model is evaluated on a seven-class, two-source grape-leaf benchmark and is reported using per-class metrics, macro and weighted averages, balanced accuracy, and MCC. Its 99.53% test accuracy and 99.71% balanced accuracy therefore indicate strong performance under the present protocol, but do not establish cross-paper superiority.

The closest direct task-level comparator is Ji et al. [31], whose four-class UnitedModel achieved 98.57% test accuracy on a PlantVillage

Table 12
Cultivar recognition: best reported test results in prior studies vs our Swin-Tiny (test metrics; unavailable metrics shown as "–").

Study	Task	Type	Model	Dataset (as reported)	Acc	Balanced Acc	Precision	Recall	F1	MCC	XAI
[27]	Variety classification	Deep learning	Proposed CNN (+ TL baselines)	27,320 images; 50 varieties	94.10% (leaf)	–	–	–	–	–	Not used
[28]	Variety classification	Deep learning	Improved MobileNetV3-Small	11 cultivars; 6051 images	96.53%	–	~96%	~96%	~96%	–	Not used (explicitly)
[29]	Variety classification	Deep learning	CNN (Keras/TensorFlow)	1011 images; 2 cultivars	95.80% / 100.00%	–	–	–	–	–	Not used
[30]	Variety classification	Deep learning	VGG16 / ResNet50 / ViT	3 varieties (smartphone)	>96.10% (leaf)	–	–	–	–	–	Region visualisation (method not specified)
[17]	Variety classification	Deep learning	Fine-tuned DenseNet201	500 images; 5 varieties	97.22%	–	–	–	–	–	Not used
Ours (Table 7)	Variety classification	Deep learning	Swin-Tiny (patch4-window7)	Grape variety dataset (this work)	–	0.8500	0.8717	0.8500	0.8344	0.8428	Reported in the manuscript (XAI section)

Note: Comparison of grape cultivar (variety) recognition studies using test-set results, highlighting that our Swin-Tiny model is reported with balanced accuracy, precision, recall, F1, and MCC, while many prior works provide only test accuracy.

ARTICLE IN PRESS

Table 13
Image-level disease-classification studies most relevant to the present Swin-Tiny disease model (test metrics; unavailable values shown as "-").

Study	Task	Type	Model	Dataset (as reported)	Test Acc	Test F1	Comparison scope	Test Precision	Test Recall	Test Balanced Acc	XAI
Ji et al. [31]	Disease classification	Deep learning	UnitedModel (multiple CNNs)	PlantVillage grape subset; 4 classes	98.57%	–	Direct task	–	–	–	Not used
Karim et al. [20]	Disease classification	Deep learning	Modified MobileNetV3Large	4-class grape disease dataset	99.42%	99.42%	Near-direct	–	–	–	Grad-CAM
Prasad et al. [19]	Disease classification	Deep learning	VGG16-based DCNN	Public grape disease dataset	99.04%	–	Near-direct	–	–	–	Not used
Ishengoma and Lyimo [21]	Disease classification	Hybrid (DL → ML)	CNN features + RF	PlantVillage grape subset	>98.50%	–	Near-direct	–	–	–	Not used
Kunduracioglu and Pacal [24]	Disease classification	Deep learning	CNNs + ViTs benchmark	PlantVillage grape (4062 images; 4 classes)	up to 100.00%	–	Near-direct	–	–	–	Not used (explicitly)
Atesoglu and Bingol [18]	Disease classification	Hybrid (DL + features+ML)	DenseNet201 + LBP/ HOG + SVM	4-class grape diseases	99.10%	–	Near-direct	–	–	–	Not used
Peng et al. [25]	Disease classification	Hybrid (DL → ML)	Fused deep features + SVM	4062 images; 4 classes	99.57%	99.65%	Near-direct	–	–	–	Not used
Ametefe et al. [35]	Leaf disease classification	Hybrid deep learning	DenseNet201 / EfficientNetB3 + edge and colour fusion	Apple leaves; four conditions	99.03% (DenseNet201)	–	Cross-crop method	–	–	–	Not used
Huang et al. [37]	Leaf disease classification	Hybrid CNN–Transformer	EConv-ViT	Apple leaves; laboratory + natural environments	99.20% (laboratory); 79.30% (natural)	–	Cross-crop method	–	–	–	Not used
Shah et al. [39]	Plant disease classification	Interpretable deep learning	ResTS	PlantVillage; 54,306 images; 14 crop species	–	99.10%	Cross-crop XAI	–	–	–	Built-in symptom-region visualisation
Ours (Table 8)	Disease classification	Deep learning	Swin-Tiny (patch4-window7)	Grape leaf disease dataset (this work)	99.53%	99.54% (weighted F1)	Reference	99.55% (weighted)	99.53% (weighted recall)	99.71%	Reported in the manuscript (XAI section)

Note: This table contains image-level classification studies only. Ji et al. [31] and the grape-specific classification studies are the closest task comparators; Ametefe et al. [35], Huang et al. [37], and Shah et al. [39] provide cross-crop methodological context. Detection studies are excluded because mAP is not directly comparable with classification accuracy, F1-score, or balanced accuracy.

grape-leaf hold-out set. The present model reports 99.53% accuracy on seven classes, but the difference should not be interpreted as a direct performance gain because the datasets, labels, and split protocols are not identical. The method-level comparison is more informative: Ngugi et al. [34] explicitly localised and classified individual lesions; Ametefe et al. [35] fused edge and colour cues with transfer learning; Huang et al. [37] combined ConvNeXt and ViT representations and exposed a substantial laboratory-to-natural performance drop; and Shah et al. [39] embedded interpretability into a residual teacher/student classifier. In contrast, the present contribution combines controlled cultivar-to-disease transfer experiments, disease-only and random-initialisation controls, multi-backbone benchmarking, imbalance-aware metrics, source- and split-sensitivity checks, and multiple post-hoc XAI methods. Fu et al. [33] and Qiao et al. [36] remain useful transformer/attention references, but their pest-recognition and spore-detection results are treated as contextual rather than numerical comparators.

3.7. Post-hoc explainability analysis

3.7.1. Overview of post-hoc explanation methods

To interpret Swin Transformer decisions beyond aggregate metrics, we employed complementary explainable AI (XAI) methods that provide post hoc evidence of which regions drive each prediction. Grad-CAM [58] produces a class-discriminative heatmap by weighting feature maps with the gradients of the target-class score, highlighting spatial regions that most strongly increase the predicted class score. Grad-CAM++ [59] extends Grad-CAM by using a refined weighting scheme based on higher-order partial derivatives, often improving localisation when multiple discriminative regions exist. Eigen-CAM [60] generates explanations from the principal components of convolutional features, avoiding explicit gradient weighting and often yielding robust, class-relevant maps. Integrated Gradients attributes a prediction to input pixels by integrating gradients along a path from a baseline input to the actual image, satisfying key axioms such as sensitivity and implementation invariance. SHAP (SHapley Additive exPlanations) [13] provides additive feature attributions grounded in Shapley values, along with a unifying framework for explanation consistency. LIME [15] explains an individual prediction by fitting an interpretable surrogate model locally around the sample via perturbations. RISE [14] estimates pixel importance by probing the model with randomly masked images (black-box perturbations) and aggregating the resulting changes in class confidence. For transformer-specific interpretability, we used attention rollout, which approximates attention flow across layers to estimate how information from input tokens contributes to the final decision. Finally, we computed token-gradient importance (input-gradient attribution on intermediate token representations) and window-attention visualisations to inspect how shifted-window attention distributes relevance across leaf structures. These methods are used here as complementary post-hoc visual explanations. The analysis is primarily qualitative and is intended to identify visually plausible evidence regions rather than to provide a quantitative or causal validation of explanation faithfulness.

Among related explainability studies, Shah et al. [39] integrated symptom-region visualisation into the ResTS plant-disease classifier, whereas Polimena et al. [38] used explanation to assess the agronomic plausibility of image-based regression features. The present study follows the same trustworthiness principle but uses a broader post-hoc suite CAM-based, perturbation-based, additive-attribution, and transformer-oriented methods together with limited occlusion and deletion/insertion checks. This comparison supports the novelty of the explanation workflow without implying that visual plausibility alone constitutes causal or expert-validated interpretability.

The explainability results for grapevine cultivar recognition are provided in the supplementary material (Supplementary Section S8; Supplementary Figs. S1–S3). The explainability results for grape leaf disease recognition are presented below.

3.7.2. Explainability results for grape leaf disease recognition

For the disease benchmark, the post hoc explanations qualitatively indicate a shift in the highlighted regions compared with cultivar recognition. Instead of emphasising global morphology (leaf contour and venation topology), the disease-tuned Swin Transformer focuses on localised symptom evidence, typically in discrete regions characterised by texture disruption, discolouration, necrotic patches, or lesion-like patterns. This behaviour is desirable because disease classes are defined primarily by the appearance of pathology rather than cultivar-specific leaf geometry, as depicted in Figs. 4 and 5. Across methods, we observe that strong predictions are accompanied by tight localisation within symptom regions, whereas weaker predictions tend to exhibit more diffuse attention across the leaf surface.

In the Bacterial Rot example (true = pred, confidence ≈0.91), the CAM-family maps (Grad-CAM, Grad-CAM++, Eigen-CAM) consistently highlight a compact region near the upper-right portion of the leaf where a visually distinct spot/lesion is present. Grad-CAM++ yields a more concentrated hotspot than Grad-CAM, which aligns with its design to improve localisation when multiple or small discriminative regions exist.

Eigen-CAM produces a similarly coherent region without relying on gradients, suggesting that the learned principal components of the feature space contain a stable disease signature in that area. The perturbation-based RISE map also peaks around the same lesion region, indicating that masking this area most reduces the predicted class probability, which is an important cross-check because RISE is a black box and does not depend on internal gradients.

The agreement among gradient-based CAMs, the gradient-free Eigen-CAM, and the perturbation-based RISE provides visual evidence that lesion regions are influential in the prediction. However, quantitative faithfulness tests would be required to confirm their causal importance.

For Leaf Blight, explanations tend to be more spatially distributed because blight often manifests as larger or multiple affected regions rather than a single small spot. In the provided Leaf Blight samples, the maps span broader leaf portions (often near margins and across textured/irregular areas), indicating that the classifier aggregates evidence from several candidate symptom zones. In such cases, Grad-CAM++ typically assigns higher intensity to multiple affected regions, whereas Eigen-CAM provides smoother coverage that better matches diffuse symptom patterns.

The SHAP overlay for Leaf Blight further supports this interpretation: it shows stronger attribution on the leaf body where texture and tonal changes occur, with comparatively lower emphasis on uniform background areas. Transformer-specific explanations provide complementary insight into how the model integrates evidence. The attention rollout visualisation for Black Measles (true = pred, confidence ≈0.91) reveals that attention mass is routed toward localised suspect regions rather than uniformly across the leaf, indicating that symptom-relevant tokens dominate the model's token mixing. Attention rollout is particularly valuable here because raw attention weights can be misleading; rollout approximates attention flow through layers to better reflect how information is propagated to the final representation. The corresponding token-gradient importance map likewise focuses on the same disease-relevant region, supporting the view that gradients with respect to intermediate tokens identify a similar subset of influential areas. When we inspect window-attention maps, we occasionally observe vertical banding or grid-like artefacts (a known side effect of projecting window-based attention onto pixel space in shifted-window transformers); nonetheless, the most salient regions still coincide with visually plausible symptoms, and we therefore interpret these maps qualitatively and in combination with RISE/CAM/SHAP for robustness.

Overall, the disease XAI analysis provides qualitative post hoc evidence that the fine-tuned Swin Transformer often highlights pathology-consistent regions, including compact lesions for bacterial rot, broader, irregular areas for leaf blight, and localised symptom patches for black measles. Cross-method agreement among CAM-family methods, Eigen-

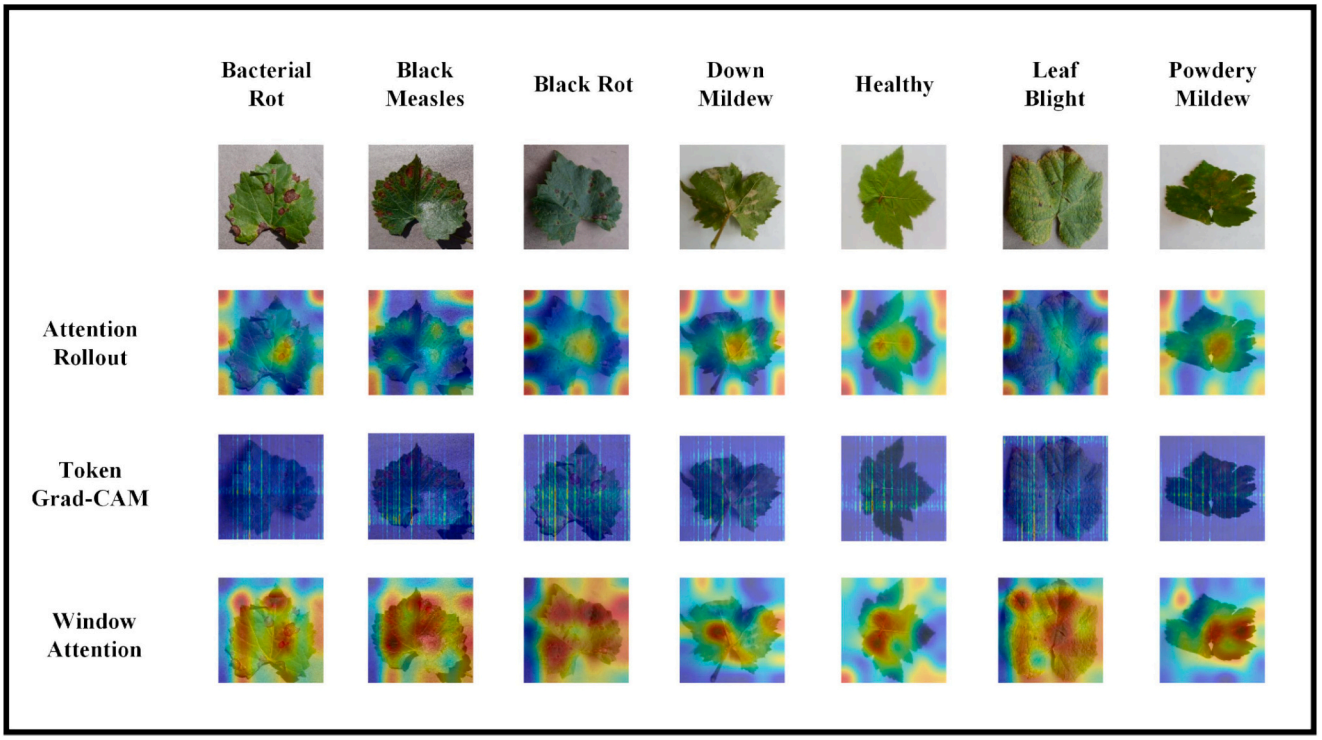


Fig. 4. Native transformer post-hoc explanations for the Swin Transformer trained on grape leaf disease images, showing Attention Rollout and transformer-specific attention/gradient maps across representative disease classes to visualise regions that may contribute to disease predictions.

CAM, RISE, and transformer rollouts lends plausibility to these visual interpretations. Although the added faithfulness checks provide limited quantitative support for some explanations, no manually annotated lesion-region comparison or expert scoring was performed. Therefore, these explanations should still be interpreted as post hoc supportive evidence rather than as rigorously validated or causal proof of the model's reasoning.

3.8. Limitations

A main limitation of this work is that, although we combined multiple public sources, the final models are still trained primarily on datasets collected under controlled or semi-controlled imaging conditions, such as relatively uniform backgrounds and consistent capture setups. These conditions can overestimate real-world performance and may lead to a pronounced lab-to-field generalisation gap when the framework is deployed in vineyards with cluttered backgrounds, shadows, occlusions, varying illumination, and mixed symptom stages. Consequently, the results reported on benchmark datasets may not fully translate to real operational vineyard environments, where domain shift remains a major challenge for reliable deployment.

A second limitation concerns the scale and diversity of the dataset. Although the proposed two-stage framework was evaluated on two labelled grape leaf datasets spanning cultivar and disease categories, broader external validation across additional vineyards, acquisition devices, seasons, growth stages, and symptom severities is still needed. In particular, the cultivar-to-disease transfer strategy should be further tested across a broader range of field conditions to confirm its robustness, stability, and generalizability beyond the currently available datasets.

A further limitation concerns the interpretation of cultivar-to-disease transfer. Although the two-stage design provides a useful framework for analysing whether cultivar-learned morphology can be adapted to disease recognition, the transfer-control experiments show that this strategy did not improve Swin-Tiny over direct ImageNet-pretrained disease-only fine-tuning under the present split. Moreover, the multi-backbone comparison indicates that the effect of cultivar pretraining is architecture-dependent, with improvements for some backbones, no change for others, and slight decreases in some cases. Therefore, the two-stage strategy should be regarded as an analytical and comparative framework rather than as a generally superior training strategy.

A third limitation concerns interpretability. Although we used multiple XAI families, including CAM-based, perturbation-based, and transformer-oriented explanations, and added limited faithfulness checks using top-attribution occlusion and deletion/insertion metrics, the analysis remains only partially validated. Saliency maps are not guaranteed to be fully faithful to the model's true decision process, can vary across methods and architectures, and may introduce artefacts, especially when projecting window-based transformer representations back to pixel space. Moreover, no manually annotated lesion-region comparison or expert-in-the-loop validation was performed. Therefore, the visual explanations presented in this study should be interpreted as supportive post hoc evidence rather than definitive causal evidence or fully validated explanations. Future work should include broader quantitative validation through systematic occlusion-sensitivity tests, deletion/insertion analyses across larger sample sets, consistency checks across explanation methods, comparisons with manually annotated lesion regions, and expert reviews with agronomists or plant pathologists.

A further limitation concerns deployment profiling. Although we report FLOPs, inference time, throughput, and peak GPU memory for the selected Swin-Tiny model on the experimental workstation, these metrics do not guarantee real-time performance on field devices. Deployment in vineyards would require additional profiling on the intended hardware platforms, including mobile devices, embedded GPUs, CPU-only systems, and edge accelerators. It should also account for preprocessing time, image upload time, explanation generation, and user interface latency.

Liu et al. [32] demonstrated that an application-specific grape detector can achieve 84.88 FPS with a 17.53 MB model, providing a useful

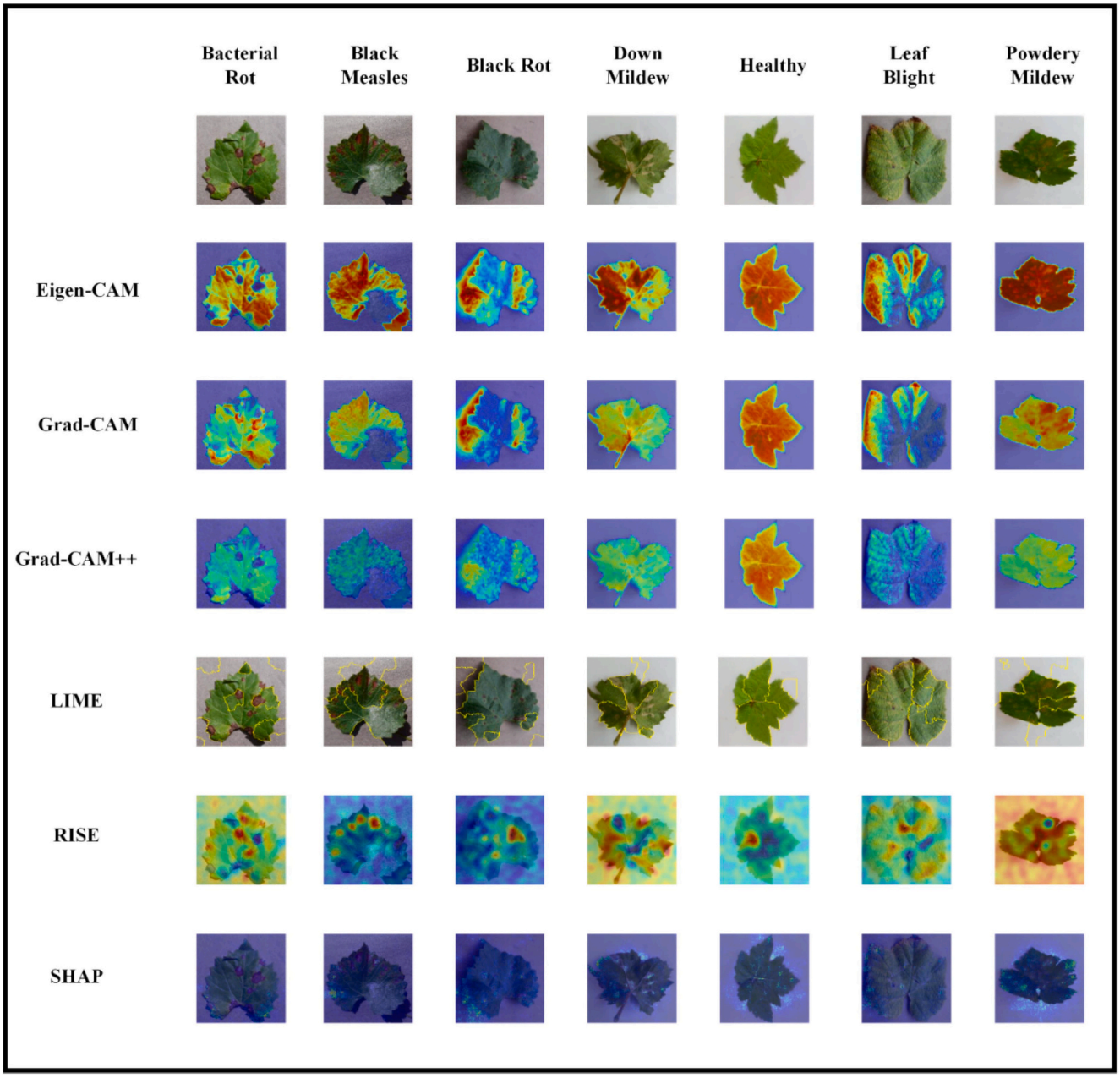


Fig. 5. Post-hoc explainability results for the Swin Transformer grape leaf disease classifier, comparing EigenCAM/Grad-CAM/Grad-CAM++, LIME, RISE, and SHAP on representative disease classes to visualise regions that may be associated with the model's predictions.

reference for edge-oriented viticulture. The present Swin-Tiny prototype, however, was profiled on a workstation and addresses a different classification task; therefore, no direct speed or model-size ranking is claimed. Deployment readiness must instead be established through profiling on mobile and embedded hardware, including end-to-end preprocessing, inference, explanation generation, interface latency, and usability under vineyard workflows.

4. Conclusions

This study presented an explainable framework for grapevine leaf analysis that integrates cultivar recognition, disease classification, interpretability, practical complexity profiling, and prototype-level image-to-decision inference within a unified agricultural information-processing pipeline. A representative set of established and recent CNN, transformer, and hybrid backbones was evaluated to contextualise the selected Swin-Tiny pipeline within the landscape of modern alternatives. The expanded cultivar benchmark showed that several recent architectures achieved higher standalone cultivar-recognition performance than Swin-Tiny; therefore, Swin-Tiny is interpreted as a competitive transformer-based backbone selected for explainable analysis rather than as the universally best cultivar classifier. On the disease benchmark, the two-stage Swin-Tiny model reached 99.07% validation accuracy and 99.53% test accuracy, with corresponding balanced accuracies of 99.01% and 99.71%. However, direct ImageNet-pretrained disease-only Swin-Tiny fine-tuning achieved 100.00% test balanced accuracy under the same split, showing that cultivar-stage pretraining did not provide a numerical improvement for Swin-Tiny in this setting. Overall, the study demonstrates that cultivar-to-disease transfer can be evaluated within a unified explainable framework, but its performance benefit is backbone-dependent and should not be considered uniformly superior to direct disease fine-tuning.

Future work should assess "in-the-wild" robustness through domain adaptation, broader field-condition datasets, cross-device testing, and external validation across vineyards with cluttered backgrounds, varying illumination, occlusion, and mixed symptom stages. The current platform should therefore be viewed as a prototype for reproducible image-to-decision inference, not as a fully validated operational deployment system. We also plan to extend the framework from single-label classification to severity grading and multi-symptom/multi-label diagnosis, and to improve the reliability of explanations by quantitatively validating XAI faithfulness using occlusion sensitivity, deletion/insertion analysis, cross-method consistency checks, comparison with manually annotated lesion regions, and user studies with agronomists or plant pathologists.

CRediT authorship contribution statement

Syed Taimoor Hussain Shah: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Syed Adil Hussain Shah:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Silvia Godio:** Writing – review & editing, Writing – original draft, Visualization, Validation, Data curation. **Karim Kassem:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology. **Shahzad Ahmad Qureshi:** Writing – review & editing, Visualization, Supervision, Methodology. **Syed Bilal Hussain:** Writing – review & editing, Validation, Supervision, Project administration. **Marco Agostino Deriu:** Writing – review & editing, Supervision, Project administration, Methodology.

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.

Acknowledgements

The authors gratefully acknowledge University College Dublin (UCD), Dublin, Ireland, and Politecnico di Torino, Torino, Italy, for supporting this joint research project and providing the academic and technical environment that enabled the development of the proposed framework, experimental pipeline, and the associated image-to-decision platform. We also thank colleagues and collaborators from both institutions for their constructive feedback during the study design, implementation, and manuscript preparation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.inpa.2026.07.010.

References

[1] Vine, wine and climate change. Éditions Quae; 2025.
[2] Ames JE. Advancing sustainable grape downy mildew management: Biopesticide trials and risk assessment system. Virginia Tech; 2025.
[3] Diseases of the Grapevine: Downy Mildew. Viticulture & Enology n.d. https://aggie-horticulture.tamu.edu/vitwine/viticulture/viticulture-resources/httpaggie-horticulture-tamu-eduvitwineviticultureviticulture-resourceshttpaggie-horticulture-tamu-eduvitwineviticultureviticulture-resourcesimage-gallery-of-grapevine-diseasesdiseases-of-the-g/diseases-of-the-grapevine-downy-mildew/ (accessed March 5, 2026).
[4] Kim D, Won J, Park H, Choi J-G, Park S, Lee G. Toward the 3rd generation of smart farming: materials, devices, and systems for E-plant technologies. Adv. Funct. Mater. 2026;36:e12264. https://doi.org/10.1002/adfm.202512264.
[5] Beery SM. Where the wild things are: Computer vision for global-scale biodiversity monitoring. . Ph.D. California Institute of Technology; 2023. https://doi.org/10.7907/m4mt-2q51.
[6] Pacal I, Kunduracioglu I, Alma MH, Deveci M, Kadry S, Nedoma J, et al. A systematic review of deep learning techniques for plant diseases. Artif. Intell. Rev. 2024;57:304. https://doi.org/10.1007/s10462-024-10944-7.
[7] Chen J, Chen J, Zhang D, Sun Y, Nanehkaran YA. Using deep transfer learning for image-based plant disease identification. Comput. Electron. Agric. 2020;173:105393. https://doi.org/10.1016/j.compag.2020.105393.
[8] Shah SBH, Naseer F, Shah SAH, Razzaq K, Javed T, Asghar Q, et al. B2-GraftingNet: a hybrid deep-machine learning framework with explainable AI for automated grape leaf disease detection. ICCK J Image Anal Process 2026;2:27–52. https://doi.org/10.62762/JIAP.2026.937901.
[9] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. 2021. https://doi.org/10.48550/arXiv.2103.14030.
[10] Khalid MI, Ahmad I, Saleem S, Hussain A, Hussain SA, Waghan AA, et al. Interpretable deep learning for diabetic retinopathy grading using regression activation maps. ICCK J Image Anal Process 2025;1:196–209. https://doi.org/10.62762/JIAP.2025.346328.
[11] Liu Z, Zhou X, Liu Y. Student dropout prediction using ensemble learning with SHAP-based explainable AI analysis. J Soc Syst Policy Anal 2025;2:111–32. https://doi.org/10.62762/JSSPA.2025.321501.
[12] Raza A, Ali A, Kumar A, Fatima N, Ali M. Bridging predictive modeling and clinical interpretability: an explainable AI approach to Parkinson's disease detection. Biomed Inform Smart Healthc 2026;2:20–37. https://doi.org/10.62762/BISH.2026.470997.
[13] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Information Processing Sys 2017 2017:4766–75.
[14] Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models. 2018. https://doi.org/10.48550/arXiv.1806.07421.
[15] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. NAACL-HLT 2016–2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session 2016:97–101. doi:10.18653/v1/n16-3020.
[16] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017. https://doi.org/10.48550/arXiv.1703.01365.
[17] Ahmed HA, Hama HM, Jalal SI, Ahmed MH. Deep learning in grapevine Leaves varieties classification based on dense convolutional network. JOIG 2023;11:98–103. https://doi.org/10.18178/joig.11.1.98-103.
[18] Atesoglu F, Bingol H. The detection and classification of grape leaf diseases with an improved hybrid model based on feature engineering and AI. AgriEngineering 2025;7. https://doi.org/10.3390/agriengineering7070228.
[19] Prasad KV, Vaidya H, Rajashekhar C, Karekal KS, Sali R, Nisar KS. Multiclass classification of diseased grape leaf identification using deep convolutional neural network(DCNN) classifier. Sci. Rep. 2024;14:9002. https://doi.org/10.1038/s41598-024-59562-x.
[20] Karim MJ, Goni MOF, Nahiduzzaman M, Ahsan M, Haider J, Kowalski M. Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and grad-CAM. Sci. Rep. 2024;14:16022. https://doi.org/10.1038/s41598-024-66989-9.
[21] Ishengoma FS, Lyimo NN. Ensemble model for grape leaf disease detection using CNN feature extractors and random forest classifier. Heliyon 2024;10:e33377. https://doi.org/10.1016/j.heliyon.2024.e33377.
[22] Hu Q, Zhang Y. GCS-YOLO: a lightweight detection algorithm for grape leaf diseases based on improved YOLOv8. Appl. Sci. 2025;15. https://doi.org/10.3390/app15073910.
[23] Xie X, Ma Y, Liu B, He J, Li S, Wang H. A deep-learning-based real-time detector for grape leaf diseases using improved convolutional neural networks. Front. Plant Sci. 2020;11. https://doi.org/10.3389/fpls.2020.00751.
[24] Kunduracioglu I, Pacal I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. J Plant Dis Prot 2024;131:1061–80. https://doi.org/10.1007/s41348-024-00896-z.
[25] Peng Y, Zhao S, Liu J. Fused-deep-features based grape leaf disease diagnosis. Agronomy 2021;11. https://doi.org/10.3390/agronomy11112234.
[26] Javidan SM, Banakar A, Vakilian KA, Ampatzidis Y. Diagnosis of grape leaf diseases using automatic *K*-means clustering and machine learning. Smart Agric Technol 2023;3:100081. https://doi.org/10.1016/j.atech.2022.100081.
[27] Terzi I, Ozguven MM, Yagci A. Automatic detection of grape varieties with the newly proposed CNN model using ampelographic characteristics. Sci. Hortic. 2024;334:113340. https://doi.org/10.1016/j.scienta.2024.113340.
[28] Deng L, Du Z, Liu X, Wu Z, Lin X, Wen B. Grape leaf cultivar identification in complex backgrounds with an improved MobileNetV3-small model. Plants 2025;14. https://doi.org/10.3390/plants14233581.
[29] Vélez S, Rubio JA, Vacas R, Barajas E. Digital ampelography: deep learning (CNN) using Keras to identify grapevine cultivars. Acta Hortic. 2024;1390:311–20. https://doi.org/10.17660/ActaHortic.2024.1390.38.
[30] Sugiura R, Ishihara M, Lim J, Iwasaki C, Oishi Y, Yakushiji H, et al. Variety Classification by Image Recognition of Grape Leaf and Berry. 2024. p. 1–6.
[31] Ji M, Zhang L, Wu Q. Automatic grape leaf diseases identification via UnitedModel based on multiple convolutional neural networks. Info Processing in Agric 2020;7:418–26. https://doi.org/10.1016/j.inpa.2019.10.003.
[32] Liu B, Zhang Y, Wang J, Luo L, Lu Q, Wei H. An improved lightweight network based on deep learning for grape recognition in unstructured environments. Inf Process Agric 2024;11:202–16. https://doi.org/10.1016/j.inpa.2023.02.003.
[33] Fu X, Ma Q, Yang F, Zhang C, Zhao X, Chang F. Crop pest image recognition based on the improved ViT method. Inf Process Agric 2024;11:249–59. https://doi.org/10.1016/j.inpa.2023.02.007.

[34] Ngugi LC, Abdelwahab M, Abo-Zahhad M. A new approach to learning and recognizing leaf diseases from individual lesions using convolutional neural networks. Inf Process Agric 2023;10:11–27. https://doi.org/10.1016/j.inpa.2021.10.004.
[35] Ametefe DS, Sarnin SS, Ali DM, Caliskan A, Tatar Caliskan I, Aliu AA. Enhancing leaf disease detection accuracy through synergistic integration of deep transfer learning and multimodal techniques. Inf Process Agric 2025;12:279–99. https://doi.org/10.1016/j.inpa.2024.09.006.
[36] Qiao C, Li K, Zhu X, Jing J, Gao W, Zhang L. Detection of cucumber downy mildew spores based on improved YOLOv5s. Inf Process Agric 2025;12:179–94. https://doi.org/10.1016/j.inpa.2024.05.002.
[37] Huang X, Xu D, Chen Y, Zhang Q, Feng P, Ma Y. EConv-ViT: a strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and transformer. Inf Process Agric 2025;12:466–77. https://doi.org/10.1016/j.inpa.2025.03.001.
[38] Polimena S, Pio G, Cefola M, Palumbo M, Ceci M, Attolico G. A novel random forest-based approach for the non-destructive and explainable estimation of ammonia and chlorophyll in fresh-cut rocket leaves. Inf Process Agric 2025;12:221–31. https://doi.org/10.1016/j.inpa.2024.09.002.
[39] Shah D, Trivedi V, Sheth V, Shah A, Chauhan U. ResTS: residual deep interpretable architecture for plant disease detection. Info Processing in Agric 2022;9:212–23. https://doi.org/10.1016/j.inpa.2021.06.001.
[40] Grapevine Leaves Image Dataset n.d. https://www.kaggle.com/datasets/muratkokludataset/grapevine-leaves-image-dataset. [Accessed 5 March 2026].
[41] Koklu M, Unlersen MF, Ozkan IA, Aslan MF, Sabanci K. A CNN-SVM study based on selected deep features for grapevine leaves classification. Measurement 2022;188:110425. https://doi.org/10.1016/j.measurement.2021.110425.
[42] Grapevine Leaves n.d.. https://www.kaggle.com/datasets/maximvlah/grapevine-leaves. [Accessed 5 March 2026].
[43] Dharrao M, Dharrao D, Sonawane R, Zade N. Niphad Grape Leaf Disease Dataset (NGLD). 2025. https://doi.org/10.17632/8nnd2ypcv3.4. 4.
[44] Grape400 Dataset n.d.. https://www.kaggle.com/datasets/nirmalsankalana/grape400-dataset. accessed March 5, 2026.
[45] Khan M. Healthy and Disease affected Leaves of Grape Plant. 2020. https://doi.org/10.6084/m9.figshare.13083890.v1.
[46] Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions, 2017, p. 1251–8.
[47] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015. https://doi.org/10.48550/arXiv.1409.1556.
[48] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2015. https://doi.org/10.48550/arXiv.1512.00567.
[49] Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. 2020. https://doi.org/10.48550/arXiv.1905.11946.
[50] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015. https://doi.org/10.48550/arXiv.1512.03385.
[51] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXivOrg 2021. https://arxiv.org/abs/2110.02178v2 (accessed May 18, 2026).
[52] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. arXivOrg 2020. https://arxiv.org/abs/2012.12877v2 (accessed May 18, 2026).
[53] Wang F, Li H, Chao W, Zhuo Z, Ji Y, Peng C, et al. E-ConvNeXt: a lightweight and efficient ConvNeXt variant with cross-stage partial connections. arXivOrg 2025. https://arxiv.org/abs/2508.20955v1 (accessed May 18, 2026).
[54] Yamagishi Y, Hanaoka S. Ensemble of ConvNeXt V2 and MaxViT for long-tailed CXR classification with view-based aggregation. arXivOrg 2024. https://arxiv.org/abs/2410.10710v2 (accessed May 18, 2026).
[55] Tan M, V. Le Q. [2104.00298] EfficientNetV2: smaller models and faster training. arXivOrg 2021. https://arxiv.org/abs/2104.00298 (accessed May 18, 2026).
[56] Yurdakul M, Uyar K, Tasdemir S. MaxGlaViT: a novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images. arXivOrg 2025. https://arxiv.org/abs/2502.17154v1 (accessed May 18, 2026).
[57] Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019. https://doi.org/10.48550/arXiv.1711.05101.
[58] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017. p. 618–26. https://doi.org/10.1109/ICCV.2017.74.
[59] Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: improved visual explanations for deep convolutional networks. arXivOrg. 2017. https://doi.org/10.1109/WACV.2018.00097.
[60] Muhammad MB, Yeasin M. Eigen-CAM: class activation map using principal components. 2020 International Joint Conference on Neural Networks (IJCNN). 2020. p. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206626.



Syed Taimoor Hussain Shah is a PostDoc Researcher within the GALATEA Project at Politecnico di Torino, Italy. He has completed his PhD under the PARENT project, which is funded by the European Union's Horizon 2020 initiative. He is based at the Politecnico di Torino in Turin, Italy. Shah earned a Master of Science in Computer Science from the Pakistan Institute of Engineering and Applied Sciences and a Bachelor of Science in Computer Science from Bahauddin Zakariya University. His current focus involves contributing to various projects aimed at developing computational explainable AI engines. These engines are designed to predict the evolution of pathology based on clinical and biological patient variables and imaging data, with a focus on adults and neonates. Shah's research interests encompass machine learning, deep learning, and pattern recognition with an emphasis on computer vision. He has contributed to the academic landscape through published works, including original articles and conference papers in esteemed peer-reviewed journals and conference proceedings, such as Frontiers, MDPI, IEEE, Springer, and CEUR Workshop Proceedings. Shah's specific research endeavours extend to areas such as facial feature extraction, characterisation, and identification in both adults and infants. Additionally, his interests include hyperspectral imaging, pattern classification, and recognition, as well as other facets of biomedical and automated computer vision, machine learning, and deep learning systems. Affiliation: PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy. ORCID: https://orcid.org/0000-0002-6010-6777. E-mail: taimoor.shah@polito.it.



Syed Adil Hussain Shah is an AI researcher at GPI SpA in Trento, Italy. In May 2026, he completed his PhD through the Marie Curie Fellowship at Politecnico di Torino (Turin, Italy). Currently, he is developing novel IoT-based solutions for the early diagnosis of motor/cognitive impairments in newborns. His field of expertise is machine vision, biomedical imaging, and intelligent system development. In 2022, he completed his master's in Computer Science with a specialisation in biomedical image analysis at COMSATS University of Islamabad, Wah campus, Pakistan. In this era, he specialised in Artificial Intelligence and related areas, namely Computer Vision, Advanced Neural Networks, Pattern Recognition, Machine Learning, and Deep Learning. Moreover, his master's thesis focused on medical image processing (Title: An ensemble deep neural network method for the classification of Pneumonia using X-ray images). The main objective of this study is to focus on deep multi-parallel fusion layers to analyse multi-aspect features in X-ray images and classify pneumonia cases at an earlier stage. In 2019, he earned his bachelor's degree in Computer Science from Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan. Affiliation: Department of Research and Development (R&D), GPI SpA, Trento 38123, Italy; PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy. ORCID: https://orcid.org/0009-0008-3916-1692. E-mail: syedadilhussain.shah@gpi.it.



Silvia Godio obtained her B.Sc in Biomedical Engineering in 2020 and her M.Sc in Biomedical Engineering in 2022, both from Politecnico di Torino. From November 2024, Silvia has been a PhD candidate in the Bioengineering and Medical-Surgical Sciences PhD program, jointly offered by Politecnico di Torino and Università di Torino. Her PhD research primarily focuses on developing a telemedicine platform that integrates AI-driven medical support and ensures compliance with critical regulations, including the AI Act, ISO 63024, and the MDR, in collaboration with SaluberMD S.r.l. Affiliation: PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy. ORCID: N/A. E-mail: silvia.godio@polito.it.



Karim Kassem is a PhD student affiliated with the PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy, and Centro Medico Santagostino, Milan, Italy. His research focuses on biomedical engineering, with particular interests in computational modelling, medical technologies, and translational applications in healthcare. Affiliation: PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy; Centro Medico Santagostino, Milan, Italy. ORCID: N/A. E-mail: karim.kassem@polito.it.



Shahzad Ahmad Qureshi received Ph.D. from PIEAS, Pakistan, in 2008. He was a Research Associate at the University of Warwick, United Kingdom, during his PhD studies. He received his MSc Engineering (Nuclear) from Quaid-i-Azam University (QAU), Islamabad. He received his MSc in Computer Science from AK University, AJK, and a Postgraduate diploma in Computer Systems from CTC, PAEC, Pakistan. He received his BSc Engineering degree in Metallurgy and Materials from UET Lahore, Pakistan. Most recently, he is an Associate Professor in the Department of Computer and Information Sciences at PIEAS. His research interests include Medical Image Analysis, Deep Learning, Biophotonics, Bioinformatics, Evolutionary Computing and Computer Vision on applications for solving problems related to Mammography, Mitosis Detection in Breast Cancer, Retinopathy, He-La Cervical Cancer Cells Investigation, 3D-DSA, Acoustic Signal Processing, Radon gas concentration-based Earthquake Prediction, Satellite Imaging, and Brain Tumour Segmentation. Affiliation: Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad 45650, Pakistan. ORCID: https://orcid.org/0000-0001-8213-1431. E-mail: drsaqureshi@pieas.edu.pk.



Syed Bilal Hussain is an accomplished academic and researcher, currently serving as an Assistant Professor at the School of Agriculture and Food Science, University College Dublin (UCD), since September 2023. He completed his PhD at the College of Horticulture & Forestry Sciences, Huazhong Agricultural University, P.R. China. His groundbreaking research during this period centred on the identification and characterisation of citrus vacuolar H^+-pyrophosphatase genes, shedding light on their crucial role in regulating the accumulation of sucrose and citrate. In September 2021, Dr Hussain started postdoctoral research fellowship at the Citrus Research and Education Centre (CREC), Institute of Food and Agricultural Sciences, University of Florida, USA. His postdoctoral research focused on plant phenology and the effects of leaf position on carbon fixation and translocation. Additionally, he made significant contributions to the Hurricane Ian citrus recovery project. Throughout his academic career, Dr Hussain has been a prolific researcher, with several articles and book chapters. He has also been an active participant in both national and international conferences, presenting his research findings to diverse audiences. His dedication to horticultural science and his extensive experience in teaching, research, and publication underscore his valuable contributions to the field. Affiliation: School of Agriculture and Food Science, University College Dublin, Dublin 4, Ireland. ORCID: https://orcid.org/0000-0002-5446-7939. E-mail: syedbilal.hussain@ucd.ie.



Marco Agostino Deriu is Professor of Industrial Bioengineering at Politecnico di Torino. He has received the European Doctorate in Biomedical Engineering in 2009 at Politecnico di Torino, Italy. His research focuses on computational and mechanistic modelling and artificial intelligence-driven data analysis, applied to investigate the mechanisms underlying biological, physiological, and pathological functions. In this context, he employs multiscale modelling to investigate molecular mechanisms of cancer and neurodegenerative diseases, structure-function relationships in physiology and pathology, drug mechanism of action, drug delivery systems, protein folding, nanoparticle features and biophysical properties. Part of his research is also devoted to developing modelling methodologies to interpret and treat biophysical, biological, and clinical data, and to build predictive models in physiology and physiopathology. He teaches "Multiscale Modelling in Biomechanics, "Biomechanical Design", and "Rational Drug Design: Principles and Applications" at Politecnico di Torino. He is the author of several publications in international peer-reviewed Journals, book chapters and proceedings in the field of computational modelling applied to bioengineering, biophysics, molecular biology, and Medicine. Affiliation: PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin 10129, Italy. ORCID: https://orcid.org/0000-0003-1918-1772. E-mail: marco.deriu@polito.it.