scientific reports

Check for updates

OPEN

A comparative analysis of single- and dual-backbone deep learning architectures with explainable AI for cherry leaf disease classification

Hüseyin Tayyip Altay[1], Özge Demir[2], Fatih Ekinci[3], Mehmet Serdar Güzel[4], Eda Kumru[5], Ilgaz Akata[6], Koray Acıcı[7] & Mustafa Sevindik[8]✉

Accurate differentiation of visually similar cherry leaf diseases remains a major challenge in precision agriculture due to overlapping symptom patterns and environmental variability. This study presents a comprehensive deep learning–based framework for multi-class cherry leaf disease classification, integrating systematic architectural comparison, statistical validation, and explainable artificial intelligence (XAI) analysis. Contrary to the common assumption that increased architectural complexity enhances performance, our results show that dual-backbone architectures consistently fail to outperform single-backbone models. A dataset comprising 4,995 cherry leaf images across five categories—brown spot, leaf scorch, healthy leaf, purple leaf spot, and shot hole disease—was used to evaluate multiple convolutional neural network architectures under fully standardized conditions. ResNet50 achieved the highest classification accuracy (98.20%), followed by EfficientNetB2 (98.00%) and DenseNet121 (97.50%), while the best dual-backbone model reached only 97.30% despite increased complexity. Statistical analysis using the Wilcoxon signed-rank test revealed a significant discrepancy between overall accuracy and macro-averaged recall ($p = 0.00195$, $r = 0.89$), demonstrating that accuracy systematically overestimates class-wise detection performance in multi-class scenarios. Grad-CAM–based explainability analysis further revealed that DenseNet-based models produce compact and semantically coherent activation maps aligned with disease-relevant regions, whereas dual-backbone architectures exhibit fragmented attention patterns associated with feature redundancy and gradient interference. These findings indicate that interpretability fidelity does not scale with architectural complexity and that coherent single-backbone feature hierarchies provide a superior balance between performance, interpretability, and generalization. The proposed framework offers both methodological and practical insights for developing reliable and scalable artificial intelligence systems in agricultural disease diagnostics.

Keywords Cherry leaf disease classification, Deep learning, Explainable artificial intelligence, Grad-CAM, Precision agriculture

Diseases affecting cherry leaves of fungal and bacterial origin constitute a major agricultural challenge by directly impairing plant physiology and leading to significant yield losses and quality deterioration[1,2]. Pathologies such as brown spot, leaf scorch, purple leaf spot, and shot hole disease are characterized by discoloration, necrosis, and tissue loss in leaf structures, often exhibiting visually similar symptoms[3,4]. This similarity makes accurate and reliable early-stage disease differentiation particularly challenging. Conventional diagnostic approaches predominantly rely on expert visual inspection, field experience, and laboratory-based analyses[5–7]. However, these methods are time-consuming, costly, and difficult to scale for large agricultural areas[3–7].

[1]Graduate School of Natural and Applied Sciences, Ankara University, Ankara, Turkey. [2]Vocational School, Big Data Analytics Program, Beykoz University, Istanbul, Turkey. [3]Institute of Artificial Intelligence, Ankara University, 06100 Ankara, Turkey. [4]Department of Computer Engineering, Faculty of Engineering, Ankara University, 06830 Ankara, Turkey. [5]Department of Biology, Graduate School of Natural and Applied Sciences, Ankara University, Ankara, Turkey. [6]Faculty of Science, Department of Biology, Ankara University, 06100 Ankara, Turkey. [7]Artificial Intelligence and Data Engineering, Ankara University, Ankara, Turkey. [8]Faculty of Engineering and Natural Sciences, Department of Biology, Osmaniye Korkut Ata University, Osmaniye, Turkey. ✉email: sevindik27@gmail.com

In recent years, artificial intelligence, particularly deep learning–based image analysis, has emerged as a powerful tool for the automated diagnosis of plant diseases[8,9]. Convolutional neural networks enable the automatic extraction of discriminative spatial and textural features from leaf images, resulting in high classification performance[10–12]. Pretrained deep learning models with varying architectural depth and capacity were systematically evaluated[10–12]. Model performance was assessed not only through accuracy-based metrics but also using sensitivity, specificity, F1-score, Matthews correlation coefficient, and AUC to provide a comprehensive evaluation[13–15]. Furthermore, to address evaluation bias arising from the multi-class structure, macro-averaged performance measures were incorporated, and the discrepancy between accuracy and class-based sensitivity was statistically examined using the Wilcoxon signed-rank test together with effect size analysis[16,17].

This paper provides an original contribution to the literature by jointly analyzing single- and multi-backbone architectures for cherry leaf disease classification in terms of both quantitative performance and explainability. The study experimentally and statistically demonstrates that increased architectural complexity does not necessarily yield performance gains, while Grad-CAM–based explainable artificial intelligence analyses are employed to visually interpret model decision mechanisms. The findings offer critical insights for the development of reliable, interpretable, and scalable artificial intelligence systems in agricultural disease diagnostics and hold broad implications for precision agriculture and digital farming applications.

While previous studies have widely applied deep learning models for plant disease classification, no study has systematically compared single-backbone and dual-backbone architectures under fully standardized conditions while simultaneously integrating explainable artificial intelligence to analyze their decision-making mechanisms. In addition, several critical gaps remain in the literature. First, most existing studies focus on evaluating individual model performance without systematically comparing different architectural strategies under controlled experimental conditions. Second, the relationship between model complexity and interpretability is often overlooked, particularly in the context of multi-backbone or hybrid architectures. Third, explainable artificial intelligence (XAI) is frequently applied as a post hoc visualization tool rather than being integrated into the model evaluation framework. Based on these limitations, this study aims to address the following research questions: (i) whether increasing architectural complexity through dual-backbone configurations leads to measurable improvements in classification performance, (ii) how architectural design influences interpretability as assessed by Grad-CAM–based analysis, and (iii) whether classification accuracy reliably reflects class-wise performance in multi-class scenarios. Accordingly, we hypothesize that increasing architectural complexity through dual-backbone designs does not necessarily lead to improved classification performance and may compromise interpretability due to feature redundancy and gradient interference. In contrast, single-backbone architectures are expected to provide a more balanced and reliable trade-off between performance, interpretability, and generalization.

Materials and methods

This section describes the complete methodological framework adopted in the study, encompassing dataset construction, preprocessing and augmentation strategies, model architecture design, training procedures, statistical evaluation, and explainability analysis. The proposed pipeline was designed to ensure methodological consistency, reproducibility, and fair comparison across different deep learning architectures. All models were trained and evaluated under identical experimental conditions, including fixed data splits, unified hyperparameter settings, and standardized performance metrics, enabling an objective assessment of architectural complexity, classification performance, and decision interpretability.

Dataset description

The dataset used in this study consists of a total of 4,995 cherry leaf images, of which approximately one thousand images were reserved for independent testing, while the remaining samples were allocated for model training[18]. The dataset was constructed to maintain an approximately balanced class distribution, with each disease category contributing a comparable proportion of samples[18]. All images were obtained from publicly accessible sources, and the complete dataset is available through the Kaggle platform[18].

The training and testing subsets comprise five distinct classes representing common cherry leaf conditions: brown spot, leaf scorch, healthy leaf, purple leaf spot, and shot hole disease. Image samples were collected from publicly available datasets, agricultural research repositories, and real orchard environments (Fig. 1). As a result, the dataset exhibits substantial variability in background composition, illumination conditions, disease severity, and leaf orientation, thereby providing a diverse and representative foundation for robust model training and evaluation. Representative examples from each disease category are illustrated in the subsequent figures, arranged sequentially from cherry brown spot to cherry shot hole disease[18].

Model architecture

The suggested cherry classification framework unifies data pretreatment, feature extraction, and classification phases through an organized deep learning pipeline, as shown in Fig. 2. RGB input photos are first scaled to a predetermined resolution. In contrast to conventional deep learning pipelines that primarily focus on optimizing individual model performance, the proposed framework introduces a systematically controlled comparative architecture that enables the direct evaluation of single-backbone and dual-backbone configurations under identical experimental conditions. This approach ensures that performance differences arise from architectural characteristics rather than variations in training protocols or data distribution. Furthermore, the study incorporates explainable artificial intelligence (XAI) as an integral component of the architectural evaluation process. Unlike previous studies where explainability is often treated as a secondary or post hoc analysis, Grad-CAM-based feature attribution is directly used to interpret and compare model behavior. Another key novelty

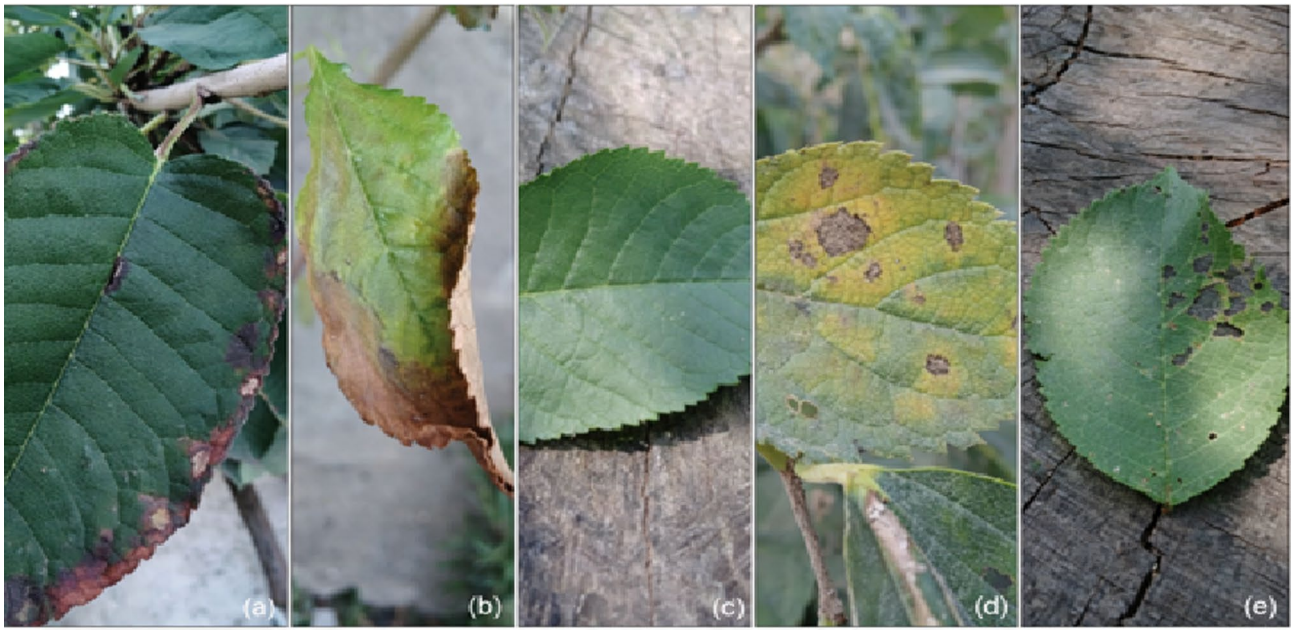

Fig. 1. (**a**) Brown spot–type symptoms predominantly along the leaf margins; (**b**) Leaf scorch–type symptoms characterized by marginal necrosis and tissue drying; (**c**) Healthy (normal) cherry leaf with intact tissue and uniform coloration; (**d**) Purple leaf spot–type symptoms with localized necrotic areas and surrounding discoloration; (**e**) Shot hole–type symptoms showing tissue loss resulting in visible perforations[18].

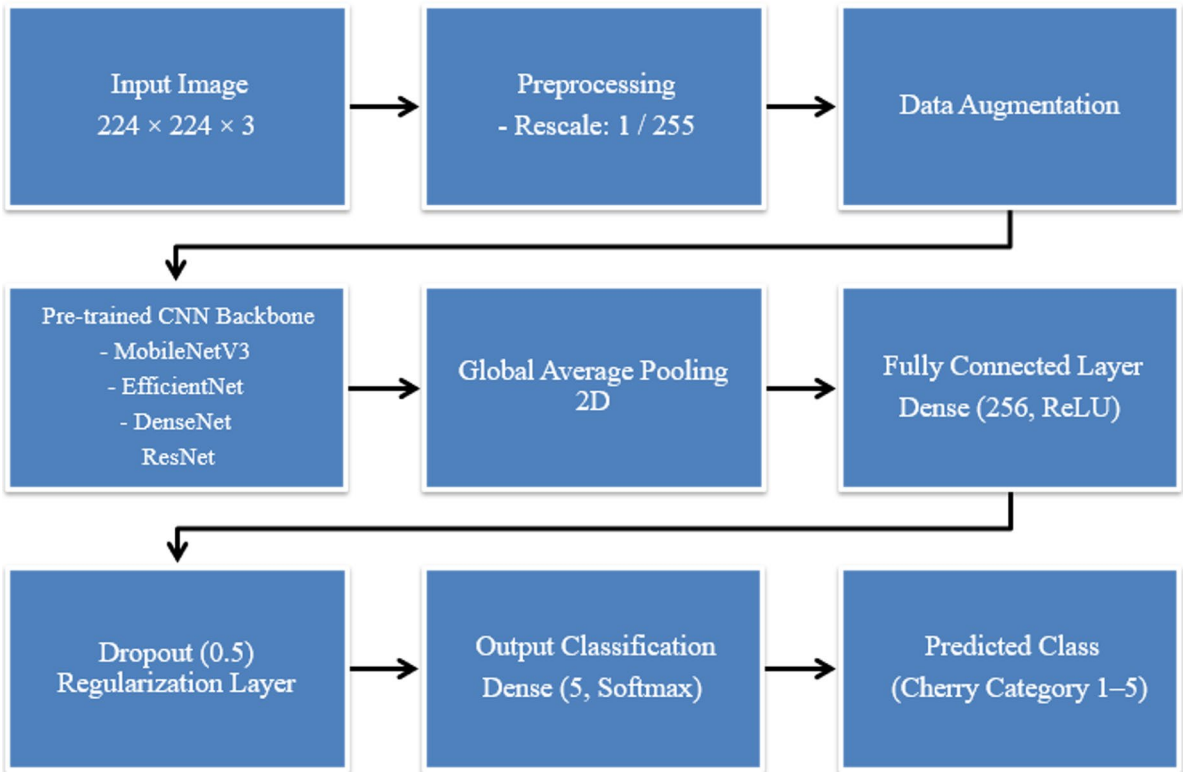

Fig. 2. Schematic representation of the proposed deep learning pipeline for cherry classification.

of the proposed framework lies in the unified optimization and evaluation pipeline, where all models are trained using identical preprocessing steps, augmentation strategies, and hyperparameter settings. This eliminates experimental bias and enables a fair, reproducible, and statistically reliable comparison of architectural complexity versus classification performance.

This preprocessing stage permits uniform input formatting across various convolutional neural network designs and guarantees numerical stability throughout training. Data augmentation approaches are only used during the training phase to improve the model's capacity for generalization and lessen overfitting 7. These include zoom transformations, random rotations, and horizontal flipping, which mimic real-world cherry appearance differences like orientation, scale, and slight viewpoint adjustments. A data generator is used to implement augmentation online, allowing the model to see a wide range of image alterations without expanding the dataset 18.

As seen in Fig. 2, the images are run through a pre-trained convolutional backbone network after preprocessing and augmentation. This study examines a variety of backbone architectures, such as lightweight models (like MobileNetV3) 19, medium-capacity networks (like EfficientNetB2 and DenseNet121) 20,21, and large-scale structures (like EfficientNetB5 and InceptionResNetV2) 22,23. During the first training stage, these backbones are used as fixed feature extractors and initialized with ImageNet weights, enabling the model to take advantage of rich hierarchical visual representations acquired from extensive datasets 24.

A Global Average Pooling (GAP) layer is then used to condense the high-dimensional feature maps generated by the convolutional backbone 25. This process lowers the number of trainable parameters and lessens the chance of overfitting by reducing spatial dimensions while maintaining the most important channel-wise information. A fully linked layer with 256 neurons and ReLU activation receives the pooled information, allowing for nonlinear feature transformation and enhanced class separability 26. In dual-backbone configurations, input images are processed in parallel by two independent pretrained backbone networks with frozen ImageNet weights during the initial training phase. No additional fine-tuning was applied to the convolutional layers, ensuring that all models operated under identical feature extraction conditions. The high-level feature maps obtained from each backbone are individually passed through Global Average Pooling layers and subsequently merged using direct feature concatenation. This approach, referred to as naïve feature-level fusion, does not incorporate any weighting, attention mechanism, or adaptive feature selection, and instead combines features in a straightforward channel-wise manner. The resulting fused feature vector is then fed into fully connected layers for classification.

After the dense layer, a dropout layer with a rate of 0.5 is added to further improve regularization 27. This layer encourages the model to acquire more resilient and distributed representations by randomly deactivating neurons during training. A dense layer with softmax activation that generates probability distributions over the five cherry classes makes up the last classification step 28.

In order to guarantee steady convergence when training on top of frozen convolutional backbones, the Adam optimizer is used for model optimization with an initial learning rate of 3×10^{-1} [29]. The loss function used to quantify the difference between ground truth labels and predicted class probabilities is categorical cross-entropy. Several callback mechanisms, such as EarlyStopping to stop training when validation performance stops improving 30, ReduceLROnPlateau to adaptively lower the learning rate upon performance stagnation 31, and ModelCheckpoint to maintain the best-performing model weights 32, are used to regulate the training process and avoid overfitting. As shown schematically in Fig. 2, these elements work together to provide a reliable and repeatable training process.

In this study, the classification performance of single-backbone, dual-backbone, and multi-backbone deep learning architectures was systematically evaluated on the proposed dataset. Multiple performance metrics based on the confusion matrix were used to enable comparable, repeatable, and objective analysis of different architectural structures. The selection of backbone architectures and their combinations was designed to enable both systematic comparison and exploratory evaluation of complementary feature extraction strategies. Specifically, high-performing and widely adopted architectures such as ResNet50, DenseNet121, and EfficientNet variants were included to represent different design paradigms, including residual learning, dense connectivity, and compound scaling. For dual-backbone configurations, architectures were paired based on their structural complementarity rather than random selection. For instance, ResNet50 and DenseNet121 were combined to integrate residual feature propagation with dense feature reuse, potentially enhancing the diversity of learned representations. Similarly, combinations involving lightweight and mid-scale models (e.g., MobileNetV2 with ResNet50) were used to investigate the trade-off between computational efficiency and representational capacity. The dual-backbone structure follows a parallel feature extraction scheme, where input images are processed independently by each backbone network. The resulting high-level feature maps are then transformed using Global Average Pooling layers and merged through feature-level concatenation. The concatenated feature vector is subsequently passed through fully connected layers for joint representation learning and final classification. This design enables the model to leverage complementary information from multiple feature hierarchies while maintaining a unified classification pipeline.

During the evaluation process, Accuracy, Precision, Recall, Specificity, F1-Score, Matthews Correlation Coefficient (MCC), Geometric Mean (G-Mean), and Area Under the Receiver Operating Characteristic Curve (AUC) were calculated. These metrics allow for a comprehensive evaluation of model performance not only in terms of overall classification accuracy but also in terms of class-based error distribution, decision balance, and discrimination capacity 33–36.

All performance metrics were derived based on the components of the confusion matrix, which defines the relationship between true class labels and model predictions 33–36. For the binary classification problem, the confusion matrix consists of the following four basic components:

- True Positive (TP): The model correctly classifies samples with a true positive class as positive,
- True Negative (TN): The model correctly classifies samples with a true negative class as negative,
- False Positive (FP): The model incorrectly classifies samples with a true negative class as positive,
- False Negative (FN): The model incorrectly classifies samples with a true positive class as negative.

The performance metrics calculated using these basic components are defined as follows: Accuracy, the ratio of all correct classifications to the total number of samples (1); precision, the ratio of truly positive samples among those predicted as positive (2); sensitivity, the ratio of correctly identifying true positive samples (3); Specificity refers to the rate at which true negative examples are correctly identified (4). The F1-Score is defined as the harmonic mean of the precision and sensitivity measures, quantitatively representing the balance between these two measures 33–36.

The Matthews Correlation Coefficient (MCC) is used as a measure that considers correct and incorrect classifications simultaneously, takes values between −1 and +1, and provides a reliable assessment, especially in datasets with class imbalance 33–36. The Geometric Mean (G-Mean) represents the geometric mean of the sensitivity and specificity values, enabling the evaluation of the balance of performance between classes. The AUC measure is calculated as a threshold-independent performance indicator reflecting the model's capacity to distinguish between positive and negative classes across different decision thresholds 33–36.

All the mentioned measures were calculated under the same evaluation protocol for single-backbone networks, dual-backbone combined structures, and the proposed multi-backbone architecture. This approach allows for a methodologically consistent and fair comparison of the impact of architectural complexity on classification performance. The mathematical expressions for the performance criteria used in this study are given below in the form of Eqs. [33–36].

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

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

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

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{4}$$

Statistical analysis

In multi-class classification problems, the overall accuracy measure is known to have the potential to overestimate model performance due to class imbalance and inter-class error asymmetries 37. Therefore, a macro-mean sensitivity (macro recall) measure, which ensures equal weighting of each class, has also been included in the performance evaluation process 38. To statistically examine the systematic difference between accuracy and macro recall values, the Wilcoxon signed-rank test was applied to paired measurements obtained from each model 39. The Wilcoxon test is proposed as a non-parametric method that does not require the assumption of normality when evaluating the difference between two related samples 40,41. In all statistical evaluations, the significance level was accepted as $\alpha = 0.05$ for both sides.

However, considering that interpretations based solely on p-values may be limited, the effect size was calculated to reveal the practical significance of the observed difference. Effect size was obtained using the standardized Z statistic of the Wilcoxon test with the formula:

and interpreted based on the recommended threshold values for small, medium, and large effect sizes 42. This statistical approach allowed for a quantitative assessment of whether the accuracy measure systematically overestimates model performance and enabled the statistical integration of class-based sensitivity into performance analysis 42

$$r = \frac{Z}{\sqrt{N}} \tag{5}$$

Results

The quantitative findings obtained in accordance with the experimental setup and evaluation protocol applied in the study reflect the classification performance of single-spinal, double-spinal, and multi-spinal deep learning architectures on the proposed dataset. All models were evaluated under the same training-test separation and the same hyperparameter settings; performance metrics were based on accuracy, precision, recall, specificity, F1-score, Matthews Correlation Coefficient (MCC), Geometric Mean (G-Mean), and area under the ROC curve (AUC). These metrics were derived from the components of the confusion matrix, and comparisons between architectures were conducted while maintaining methodological consistency in Table 1.

When examining the performance metrics presented in Table X, it is observed that among single-backbone architectures, the ResNet50 model achieved the highest overall performance with 98.20% accuracy, 0.9807 F1-Score, 0.9769 MCC, and 0.9985 AUC. The EfficientNetB2 model exhibited a very similar and stable performance to ResNet50 with 98.00% accuracy and 0.9984 AUC. The DenseNet121 architecture produced strong results in terms of class balance with 97.50% accuracy and 0.9813 G-Mean value. Among lightweight architectures, MobileNetV2 and MobileNetV3Small offered competitive performance with approximately 97% accuracy levels,

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC	G-Mean	AUC
DenseNet121	0.9750	0.9670	0.9688	0.9939	0.9670	0.9680	0.9813	0.9980
ResNet50	0.9820	0.9807	0.9809	0.9954	0.9807	0.9769	0.9881	0.9985
MobileNetV3Small	0.9710	0.9589	0.9571	0.9932	0.9572	0.9628	0.9750	0.9974
NASNetMobile	0.9019	0.9008	0.8706	0.9751	0.8766	0.8766	0.9213	0.9930
MobileNetV2	0.9700	0.9620	0.9612	0.9926	0.9613	0.9615	0.9768	0.9981
EfficientNetB0	0.9620	0.9520	0.9464	0.9909	0.9451	0.9518	0.9684	0.9978
EfficientNetB2	0.9800	0.9700	0.9710	0.9953	0.9697	0.9744	0.9831	0.9984
Dual_ResNet50_DenseNet121	0.9610	0.9507	0.9497	0.9905	0.9483	0.9501	0.9699	0.9975
Dual_ResNet50_MobileNetV2	0.9730	0.9673	0.9689	0.9933	0.9677	0.9653	0.9810	0.9978
Dual_DenseNet121_EfficientNetB0	0.9750	0.9655	0.9668	0.9940	0.9657	0.9679	0.9803	0.9968

Table 1. Model Performance Metrics.

while the NASNetMobile model showed the lowest performance among the evaluated architectures with 90.19% accuracy and 0.8706 sensitivity.

When evaluating dual-spinal architectures, the Dual_ResNet50_MobileNetV2 model was observed to have the highest performance among combined structures with 97.30% accuracy and an AUC value of 0.9978. The Dual_DenseNet121_EfficientNetB0 model achieved 97.50% accuracy but did not surpass the performance of the corresponding single-spinal architectures. Similarly, the Dual_ResNet50_DenseNet121 model lagged behind both main architectures with 96.10% accuracy. Specificity values ranging from 0.9751 to 0.9954 for all models indicate a generally high success rate in correctly classifying negative samples. Furthermore, AUC values above 0.9930 for all architectures reveal a high level of inter-class discrimination.

In addition to quantitative performance metrics, Grad-CAM-based visual explainability analyses revealed significant differences between architectural structures. DenseNet-based architectures, in particular, have exhibited more consistent and stable explainability patterns thanks to their densely connected feature transmission mechanisms. In the DenseNet architecture, each layer is directly connected to all subsequent layers, ensuring that low-level spatial information and high-level semantic representations are preserved within the same feature space. This structure reduces the gradient fading problem, which is critical for gradient-based attribution methods, allowing for more stable gradient flow in Grad-CAM maps.

This architectural feature enables DenseNet-based models to identify disease-related regions in leaf images with more compact, semantically consistent, and low-noise activation maps. Distinctive patterns such as lesion textures, color gradient changes, and structural distortions are preserved across different layers of the network as a natural consequence of feature reuse, directly contributing to the final classification decision. Consequently, Grad-CAM images generated by DenseNet architectures show clearer regional boundaries and higher inter-sample consistency compared to alternative backbone designs.

Furthermore, the implicit regularization effect based on feature repetition in the DenseNet architecture prevents the formation of contradictory attention patterns by limiting overparameterization. This contributes to the stable preservation of class-separator regions even under different lighting conditions, scale changes, and background complexities. As a result, DenseNet-based architectures not only exhibit high quantitative performance but also offer more reliable and interpretable decision structures for the predictable classification of plant diseases. All of the examined models' validation accuracy and validation loss curves over the training epochs are shown in Fig. 3, which sheds light on their convergence behavior and capacity for generalization.

Rapid convergence, little oscillation, and steady loss reduction are characteristics of single-backbone architectures like ResNet50, EfficientNetB2, and DenseNet121, which provide smooth and well-aligned validation accuracy and loss trajectories. Despite their great representational capacity, this behavior shows strong generalization and efficient optimization, with no obvious indications of overfitting.

On the other hand, a number of dual-backbone models show less consistent validation dynamics, such as delayed convergence and discernible variations in validation loss, especially in later epochs. These trends imply that naïve feature-level fusion and increased architectural complexity may create optimization issues, resulting in partial overfitting or decreased generalization efficiency. Crucially, compared to their well-regularized single-backbone counterparts, certain dual models attain competitive validation accuracy, but their larger loss variance suggests a less consistent decision boundary. The confusion matrices of all ten assessed models are shown in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13, which provides a thorough class-wise examination of each model's prediction performance on the test dataset.

Strong diagonal dominance is generally seen by high-performing single-backbone architectures like ResNet50, EfficientNetB2, and DenseNet121, suggesting accurate classification across all cherry categories with little inter-class confusion. These models are very good at learning discriminative and well-separated feature representations, as evidenced by their consistently low false positive and false negative rates.

On the other hand, off-diagonal components are more prominent in lightweight architectures and some dual-backbone models, indicating localized misclassification patterns between visually identical classes. Notably, physically similar cherry leaf categories are frequently confused by models like NASNetMobile and several dual combinations, indicating shortcomings in fine-grained feature classification. This trend is consistent with their comparatively lower recall and MCC values seen in Table 1, indicating that misclassifications are clustered in particular class pairs rather than being uniformly distributed.

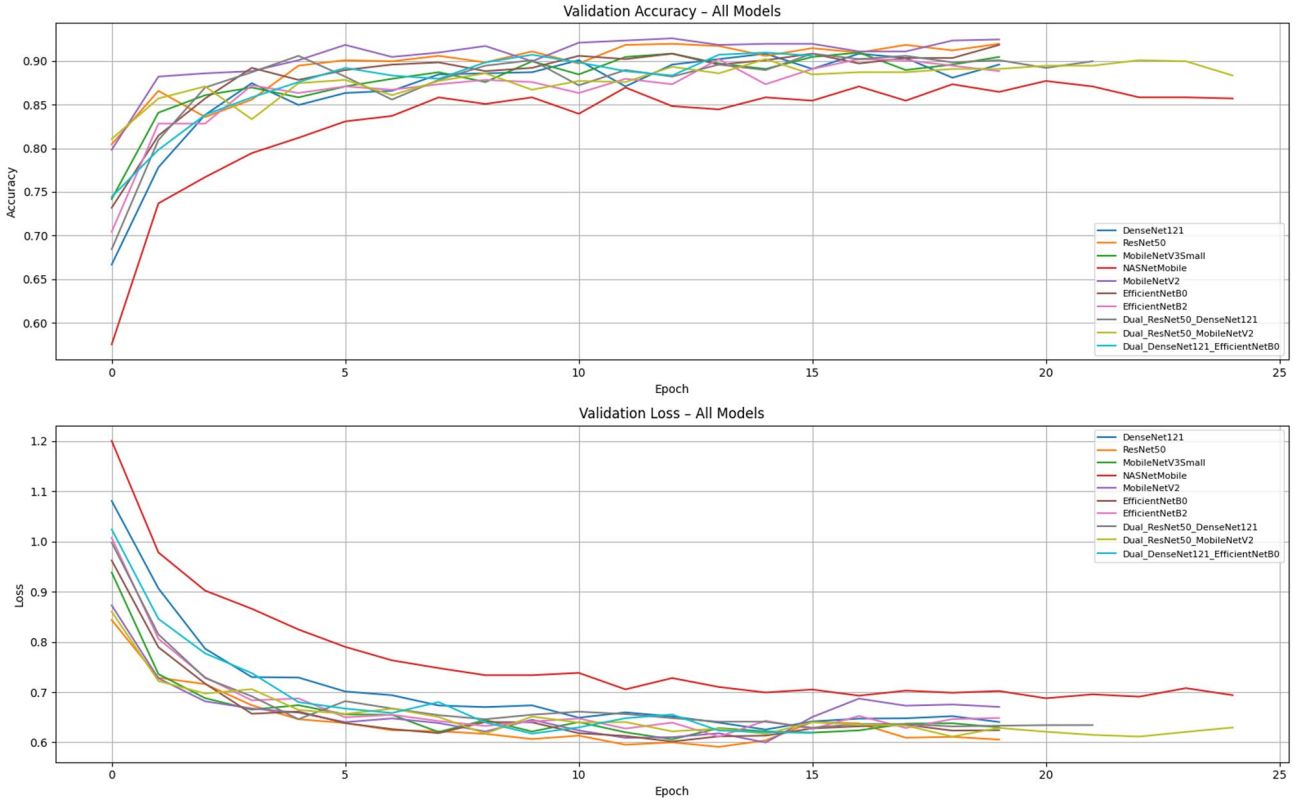

Fig. 3. Validation accuracy/loss curves.

Crucially, the confusion matrices provide additional insight into why dual-backbone models did not consistently perform better than their single-backbone equivalents. The observed error dispersion suggests that redundant or contradictory feature responses may erode class-specific decision boundaries, despite the fact that feature fusion boosts representational capacity. Overall, Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 supports the quantitative performance measurements by showing that balanced and steady class-wise generalization across all categories, in addition to accurate predictions, promote greater overall accuracy. A concise and clear summary of relative generalization performance is given by the radar chart in Fig. 14, which compares the best validation accuracy attained by each single and dual model.

DenseNet-based designs continually occupy the outside areas of the radar space, as the chart plainly shows, demonstrating better validation accuracy and more consistent learning behavior during training epochs. This finding implies that dense connection patterns support efficient gradient propagation and feature reuse, which are especially helpful for fine-grained visual discriminating in the cherry leaf dataset.

On the other hand, the radar chart shows that robust single-backbone models are routinely outperformed by naïve dual-backbone fusion techniques. Some dual models are still limited to the inner areas of the radar plot while having a greater representational capacity, which indicates that the increases in validation accuracy have been minimal or nonexistent. This result suggests that rather than improving discriminative signals, basic feature-level concatenation may add redundancy or weaken them. The conclusion that architectural coherence and feature integration technique are more important for generalization performance than model complexity alone is often supported by Fig. 14. The Receiver Operating Characteristic (ROC) curves for all single and dual models are shown in Fig. 15, offering a thorough understanding of their ability to discriminate across all classes.

The curves unequivocally show that ResNet50 and DenseNet121 attain the largest Area Under the Curve (AUC) values, demonstrating their greater capacity to discriminate between subtly distinct classes in the cherry leaf dataset. Their robust generalization and efficient feature extraction are supported by their dense and residual connection patterns.

In the multi-class classification structure, considering the possibility that the overall accuracy measure may be affected by the class distribution, macro-mean sensitivity (macro recall) values giving equal weight to each class were also calculated. The statistical significance of the differences between overall accuracy and macro-mean sensitivity was examined using the Wilcoxon signed-rank test on paired measurements; effect size values were reported to reveal the practical significance of the observed differences. Wilcoxon signed-rank test revealed a statistically significant difference between accuracy and recall values among the evaluated models ($p = 0.00195$) shown in Table 2.

The consistent aspect of the differences (accuracy > recall for all models) indicates that overall accuracy systematically increases model performance compared to class-based sensitivity. The calculated effect size is large ($r = 0.89$), suggesting that this discrepancy is not only statistically significant but also practically meaningful.These

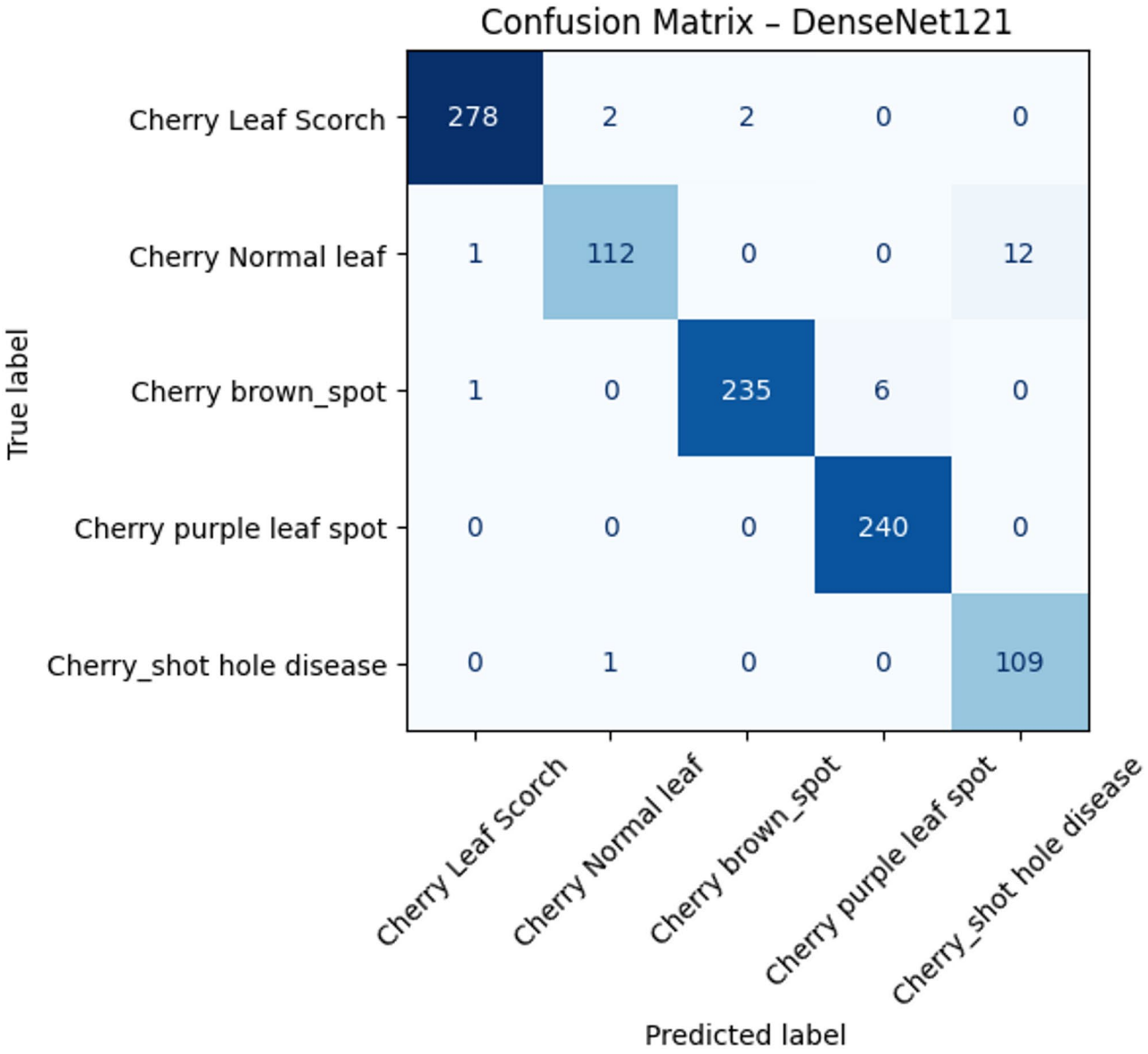

Fig. 4. Confusion Matrix-DenseNet121.

findings highlight that relying solely on accuracy can mask shortcomings in class-based detection performance, particularly in multi-class classification environments with visually similar categories. Consequently, recall-focused evaluation provides a more reliable and discriminatory assessment of true model effectiveness.

On the other hand, as compared to their single-backbone counterparts, a number of dual-backbone models exhibit relatively small increases or decreases in AUC. This implies that naïve feature-level fusion may result in feature redundancy or attention conflicts, which could reduce the merged representation's ability to discriminate. Overall, Fig. 15 shows that although ensemble-like designs have promise, obtaining steady gains in classification performance requires careful consideration of backbone compatibility and feature integration techniques.

GradCAM++was preferred over EigenCAM due to its native compatibility with TensorFlow-based architectures and its improved localization stability. Grad-CAM–based explainability analysis revealed distinct attention behaviors across single-backbone, dual-backbone, and ensemble architectures shown in Figs. 16, 17, 18, 19, 20, 21 and 22.

Single-backbone models, particularly DenseNet-based architectures, produced compact and semantically coherent activation regions that consistently overlapped with disease-relevant leaf structures such as lesion boundaries, texture irregularities, and color discontinuities. This indicates that a unified feature hierarchy supports stable gradient propagation and class-discriminative localization, thereby enhancing both interpretability and generalization performance.

In contrast, dual-backbone architectures exhibited more diffuse and fragmented attention maps. Although dual models may capture complementary representations, naïve feature-level fusion often introduces redundancy and interference between heterogeneous feature spaces. As a result, Grad-CAM activations in dual-backbone

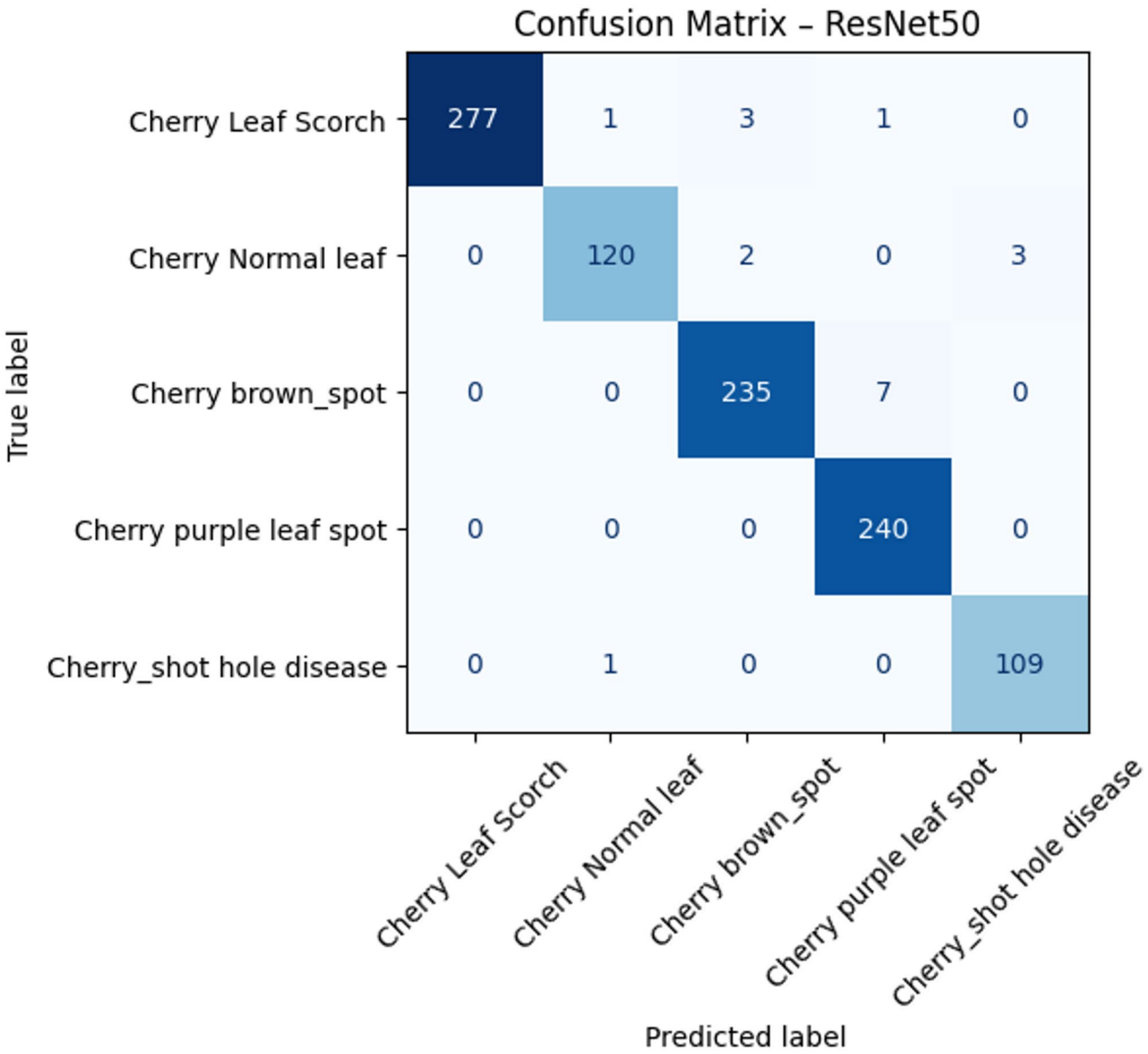

Fig. 5. Confusion Matrix-ResNet50.

models tend to emphasize broader and less class-specific regions, including background or non-pathological areas, suggesting feature over-compression and reduced saliency of discriminative patterns. These effects are further amplified by gradient dilution across parallel feature streams, leading to diminished localization precision and interpretability consistency. Although Grad-CAM analysis in this study is primarily qualitative, the observed consistency between attention patterns and performance metrics suggests a meaningful relationship between interpretability and classification reliability. For example, models such as DenseNet121, which exhibit more compact and localized activation regions, also achieve higher class-wise performance metrics, whereas dual-backbone models with fragmented attention patterns show reduced discrimination capability. A fully quantitative evaluation of attention coherence, such as measuring overlap with annotated disease regions or computing entropy-based attention dispersion, could further strengthen this relationship. However, such analysis requires pixel-level annotations, which are not available in the current dataset. Therefore, future work should incorporate quantitatively grounded explainability metrics to systematically evaluate the link between attention quality and model performance.

Ensemble architectures presented a different explainability trade-off. While frequently achieving competitive or superior quantitative performance, ensemble-based Grad-CAM visualizations displayed smoother and averaged attention distributions due to decision-level aggregation rather than a shared latent representation. This behavior may improve robustness and reduce overfitting but can obscure fine-grained spatial cues, limiting transparency at the individual feature level. Overall, the Grad-CAM analysis demonstrates that explainability fidelity does not scale linearly with architectural complexity. Instead, models with a single, well-aligned feature hierarchy such as DenseNet121 offer more reliable and interpretable decision mechanisms.

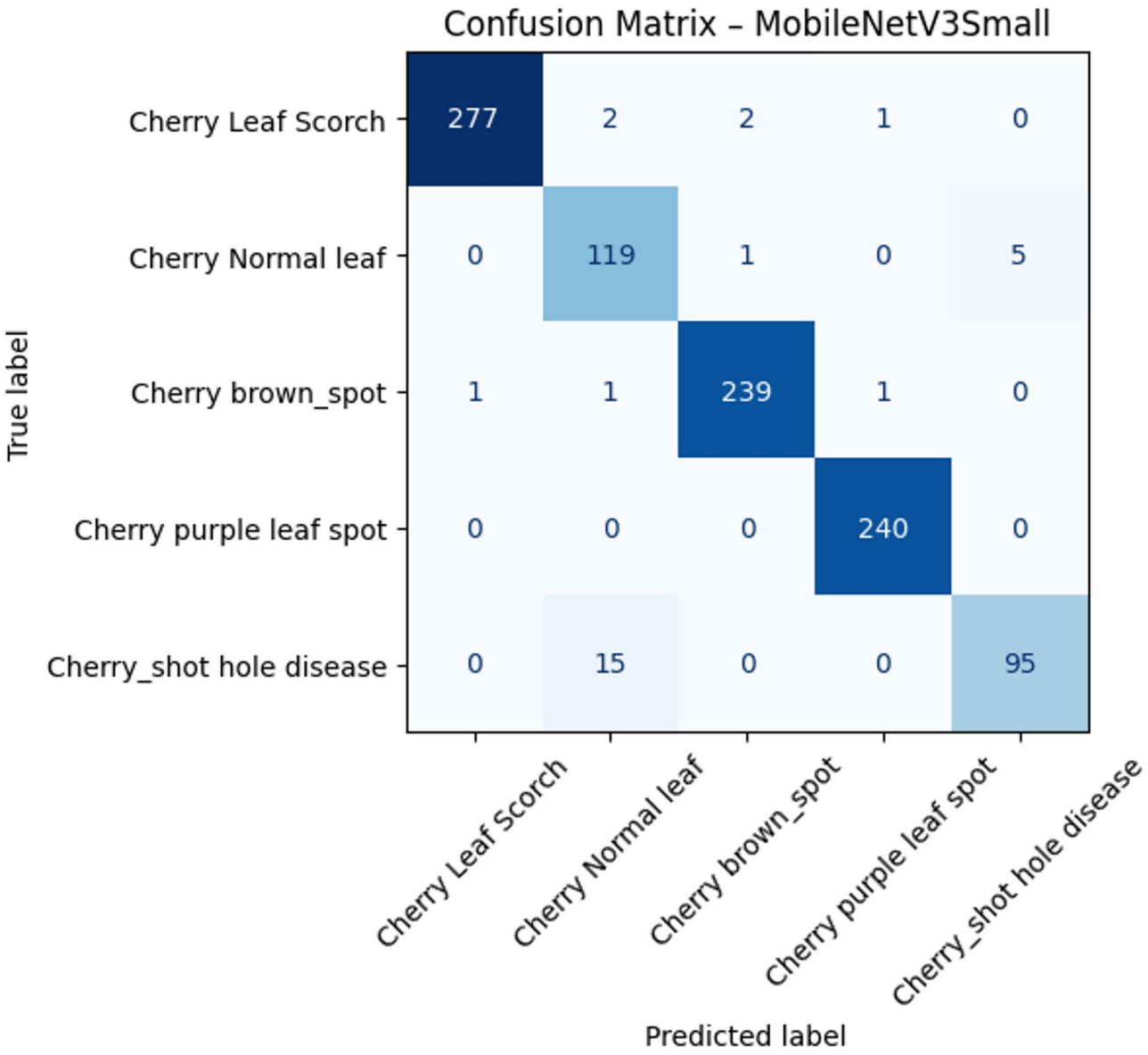

Fig. 6. Confusion Matrix-MobileNetV3Small.

Additionally, visual differences observed in Grad-CAM overlays across backbone architectures were primarily attributed to architecture-specific preprocessing pipelines rather than inconsistencies in the explainability method itself. To ensure faithful qualitative comparison, all Grad-CAM visualizations were inverse-normalized to a common intensity range prior to rendering. This standardization preserved spatial attribution integrity and enabled consistent cross-model evaluation of attention distributions. Collectively, these findings highlight the importance of architectural coherence and informed feature integration strategies when designing explainable deep learning systems for fine-grained plant disease classification.

Discussion

The quantitative comparison of single-backbone and dual-backbone architectures reveals clear performance stratification across model families. Among single-backbone networks, ResNet50 achieved the highest classification accuracy at 98.20% [43], marginally outperforming EfficientNetB2 (98.00%, −0.20%) [44] and DenseNet121 (97.50%, −0.70%) [45]. These three architectures consistently demonstrated strong class balance, as reflected by their high F1-scores (0.97–0.98), MCC values (0.97–0.98), and AUC scores exceeding 0.998, indicating near-optimal inter-class discrimination. In contrast, lightweight architectures exhibited a moderate decline in performance; MobileNetV2 and MobileNetV3Small achieved accuracies of approximately 97.0% [46,47], corresponding to a 1.2% reduction relative to ResNet50, while still maintaining acceptable generalization. A more pronounced degradation was observed for NASNetMobile, which reached only 90.19% accuracy 48, reflecting a substantial 8.01% decrease compared to the best-performing model, and indicating limited capacity for fine-grained differentiation among visually similar disease categories.

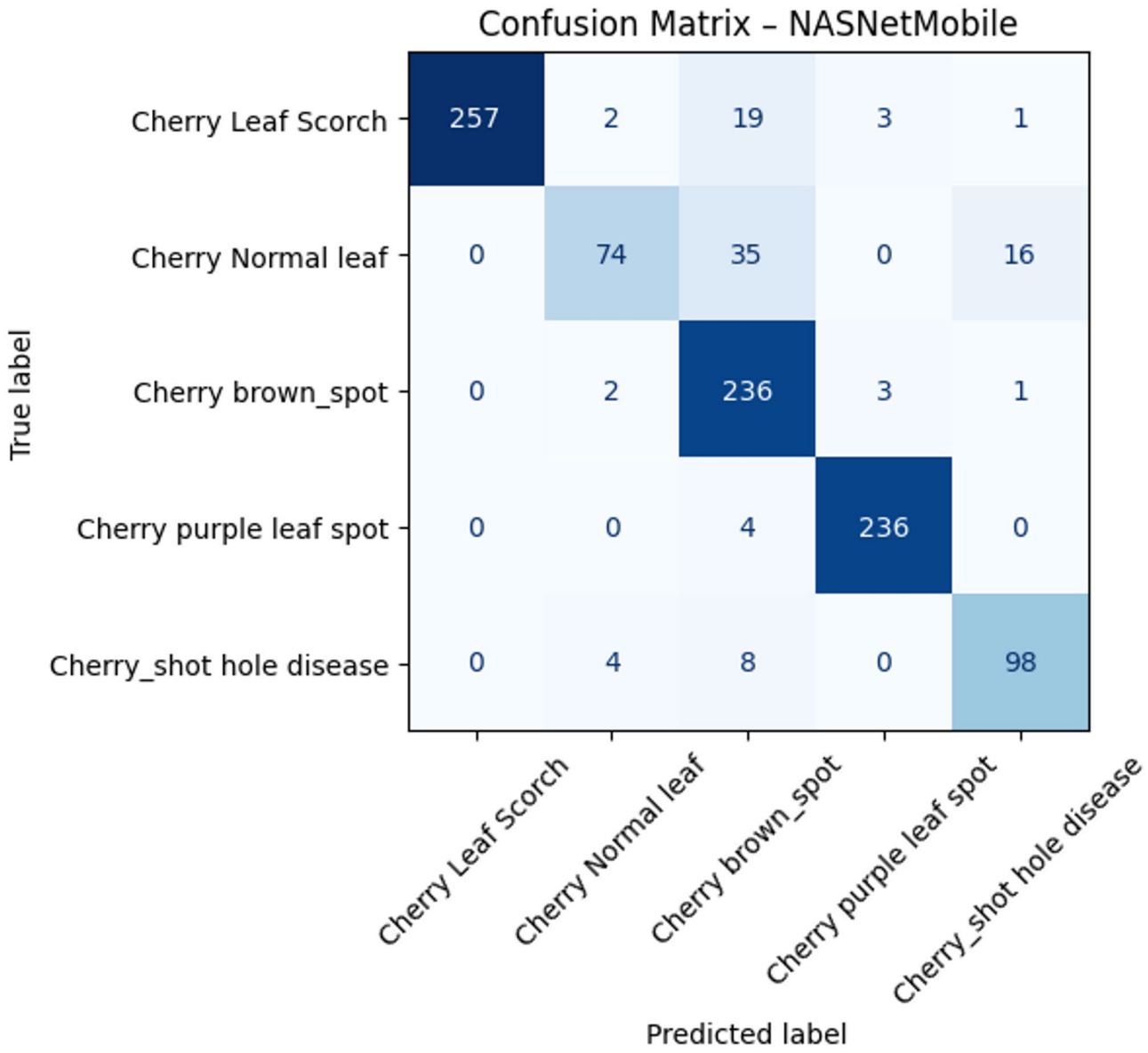

Fig. 7. Confusion Matrix-NasNetMobile.

Dual-backbone architectures did not yield systematic performance improvements over their single-backbone counterparts. The best-performing dual configuration, Dual ResNet50 and MobileNetV2, achieved 97.30% accuracy 49–51, which remained 0.90% lower than ResNet50 and 0.70% lower than EfficientNetB2. Other dual models exhibited larger performance gaps; Dual DenseNet121 and EfficientNetB0 [52,53] and Dual ResNet50 and DenseNet121 [54,55] achieved accuracies of 97.50% and 96.10%, corresponding to declines of 0.70% and 2.10%, respectively. Despite their increased representational capacity, these architectures failed to translate architectural complexity into measurable performance gains, suggesting that naïve feature-level fusion may introduce redundancy and suppress discriminative features 56. The limited performance of dual-backbone models can be attributed to conflicting gradient updates between parallel feature streams, which may destabilize optimization. In addition, direct feature concatenation may introduce redundancy rather than complementary information, weakening class separation, as also reflected in increased off-diagonal errors in confusion matrices. The higher parameter count further increases the risk of overfitting, suggesting that architectural complexity without adaptive fusion does not guarantee improved generalization.

Beyond quantitative performance, explainability analysis using Grad-CAM provided critical insights into the decision-making behavior of the evaluated architectures 57. DenseNet-based models consistently produced compact, well-localized, and semantically meaningful activation maps aligned with pathological leaf regions, such as lesion boundaries and necrotic textures 58. In contrast, dual-backbone architectures exhibited more diffuse and fragmented attention patterns, often extending into background or non-pathological regions 59. This behavior reflects gradient dilution and feature interference arising from parallel feature streams, which reduces localization precision and interpretability. Ensemble models, while achieving competitive accuracy and AUC

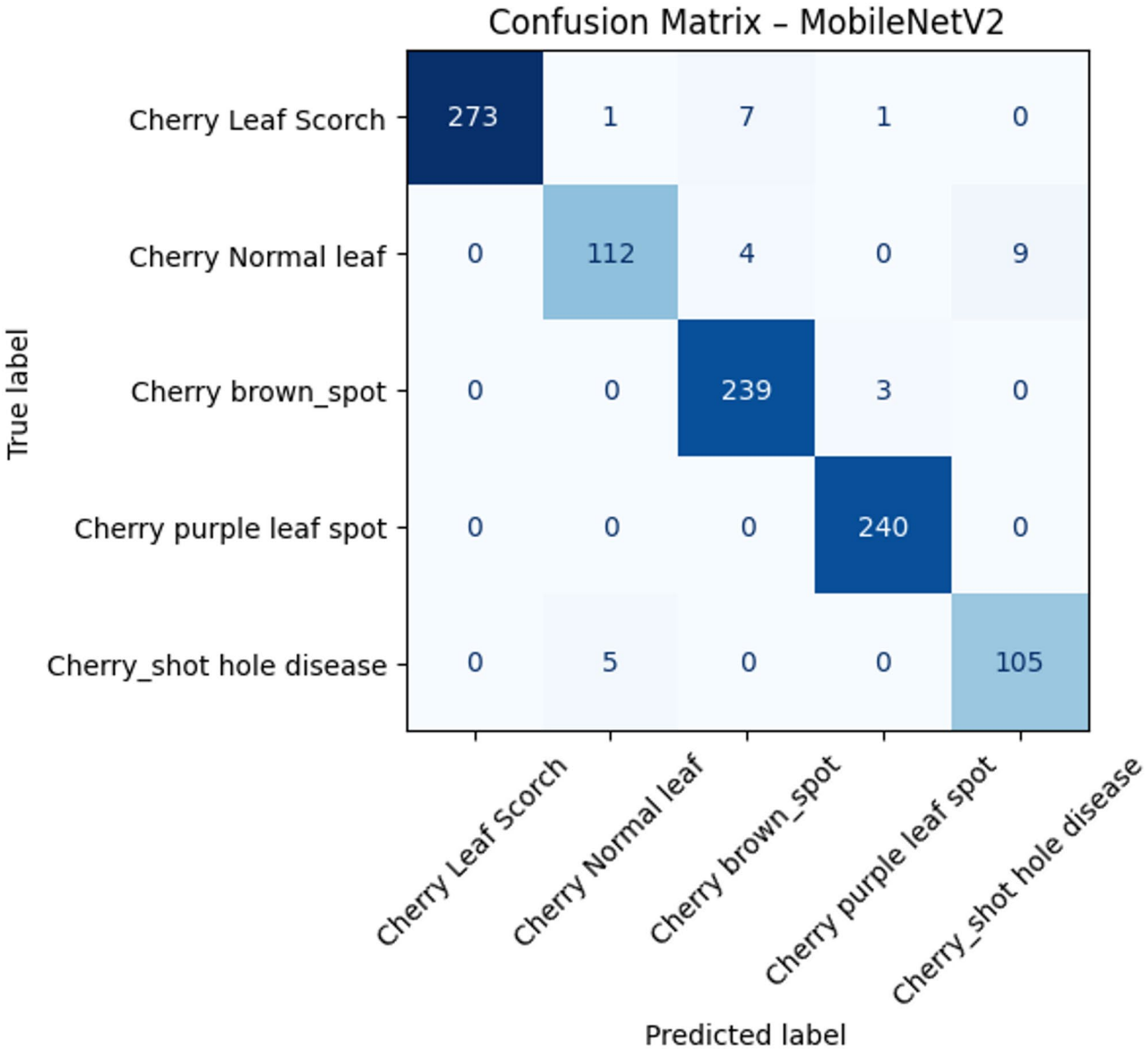

Fig. 8. Confusion Matrix-MobileNetV2.

values, generated smoother and averaged attention maps due to decision-level aggregation, limiting fine-grained spatial interpretability 60. These findings indicate that XAI fidelity does not scale linearly with architectural complexity, and that a coherent single-backbone feature hierarchy offers superior interpretability alongside high classification performance 61. The Grad-CAM findings are consistent with the quantitative performance metrics. For instance, the compact and well-localized activation patterns observed in DenseNet-based models correspond to their high MCC and recall values, indicating more balanced class-wise discrimination. In contrast, the fragmented and diffuse attention maps in dual-backbone architectures align with increased off-diagonal errors in the confusion matrices, particularly between visually similar classes such as brown spot and leaf scorch. This suggests that less focused attention reduces fine-grained classification capability and contributes to performance degradation. Beyond qualitative visualization, explainable artificial intelligence also serves as a diagnostic tool for evaluating model reliability. Models that generate focused and disease-relevant activation regions tend to exhibit more stable and generalizable classification behavior, as their predictions are guided by biologically meaningful features rather than background noise. The alignment between Grad-CAM activation regions and pathological leaf structures supports the validity of the learned feature representations and reduces the likelihood of spurious correlations. Consequently, models with more coherent attention patterns achieve better balance across class-wise metrics such as recall and MCC, highlighting a direct link between interpretability and classification performance.

When compared with existing image-based plant disease detection studies, the results of this work are consistent with the growing body of literature demonstrating the effectiveness of deep learning for leaf disease classification, typically reporting accuracy levels in the high-range depending on dataset complexity and disease

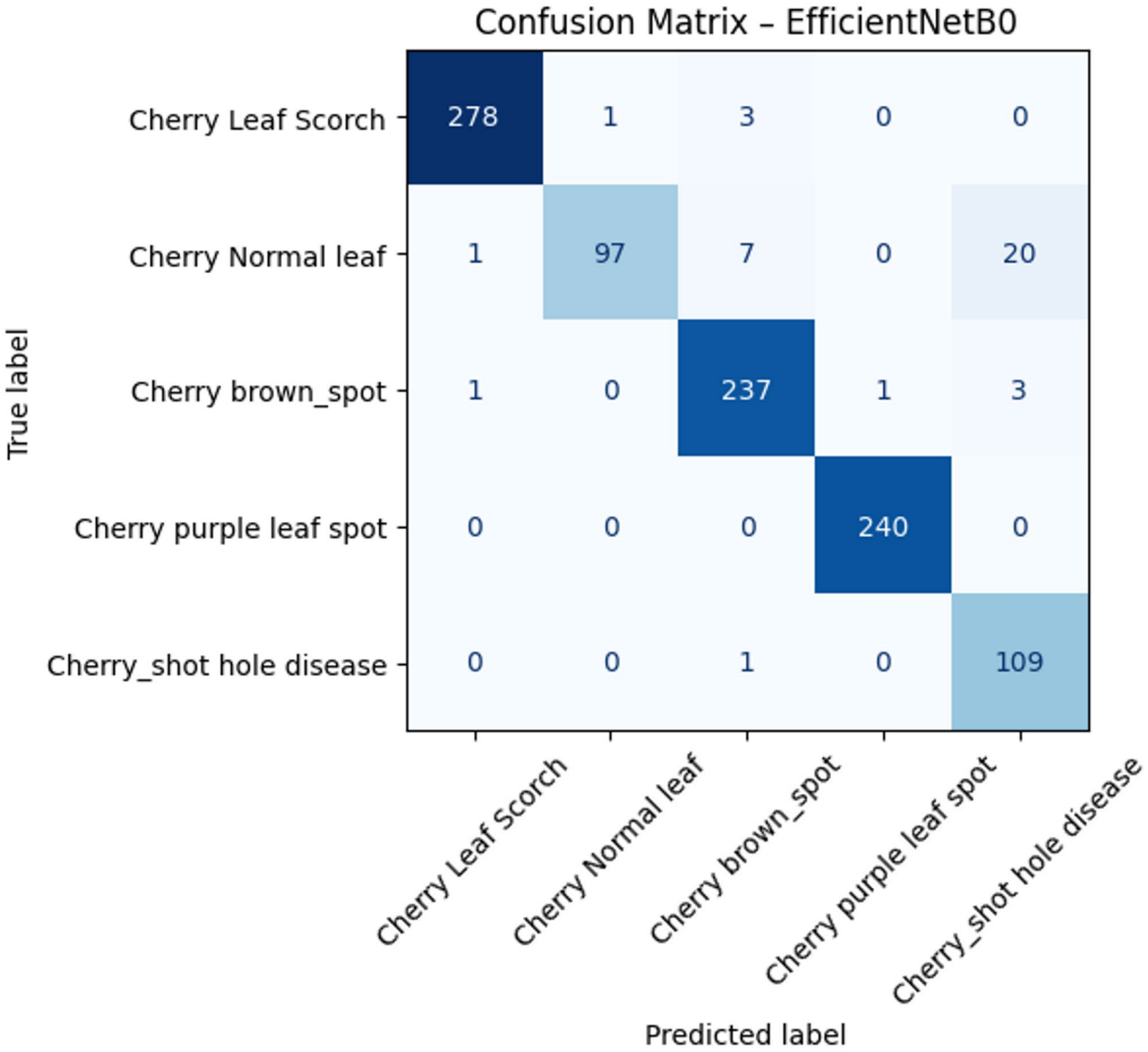

Fig. 9. Confusion Matrix-EfficientNetB0.

similarity 62. However, many prior studies rely on limited architectures, single performance metrics, or lack systematic explainability analysis 63. In contrast, this study provides a comprehensive comparison across multiple backbone configurations, integrates statistical validation to address metric bias, and incorporates XAI-based evaluation to assess decision reliability 59–63. Furthermore, unlike traditional image processing approaches based on handcrafted features or color–texture descriptors, the deep learning frameworks employed here automatically learn hierarchical representations that are robust to variations in illumination, orientation, and disease severity 64. Collectively, these contributions position the proposed analysis as a more reliable and interpretable reference framework for AI-driven plant disease diagnosis, with direct implications for scalable agricultural monitoring and precision farming systems 65,66.

Although the dataset used in this study was derived from a publicly available Kaggle source, it incorporates images collected under diverse environmental conditions, including variations in lighting, background complexity, leaf orientation, and disease severity. This variability provides a partially realistic representation of field conditions and supports the robustness of the trained models. However, it is acknowledged that fully real-time or field-acquired datasets may introduce additional challenges such as motion blur, occlusions, sensor noise, and uncontrolled environmental factors. Therefore, future studies should focus on validating the proposed framework using real-time data collected from in-field imaging systems, mobile devices, or drone-based platforms. Such validation would further strengthen the practical applicability and deployment potential of the model in precision agriculture scenarios.

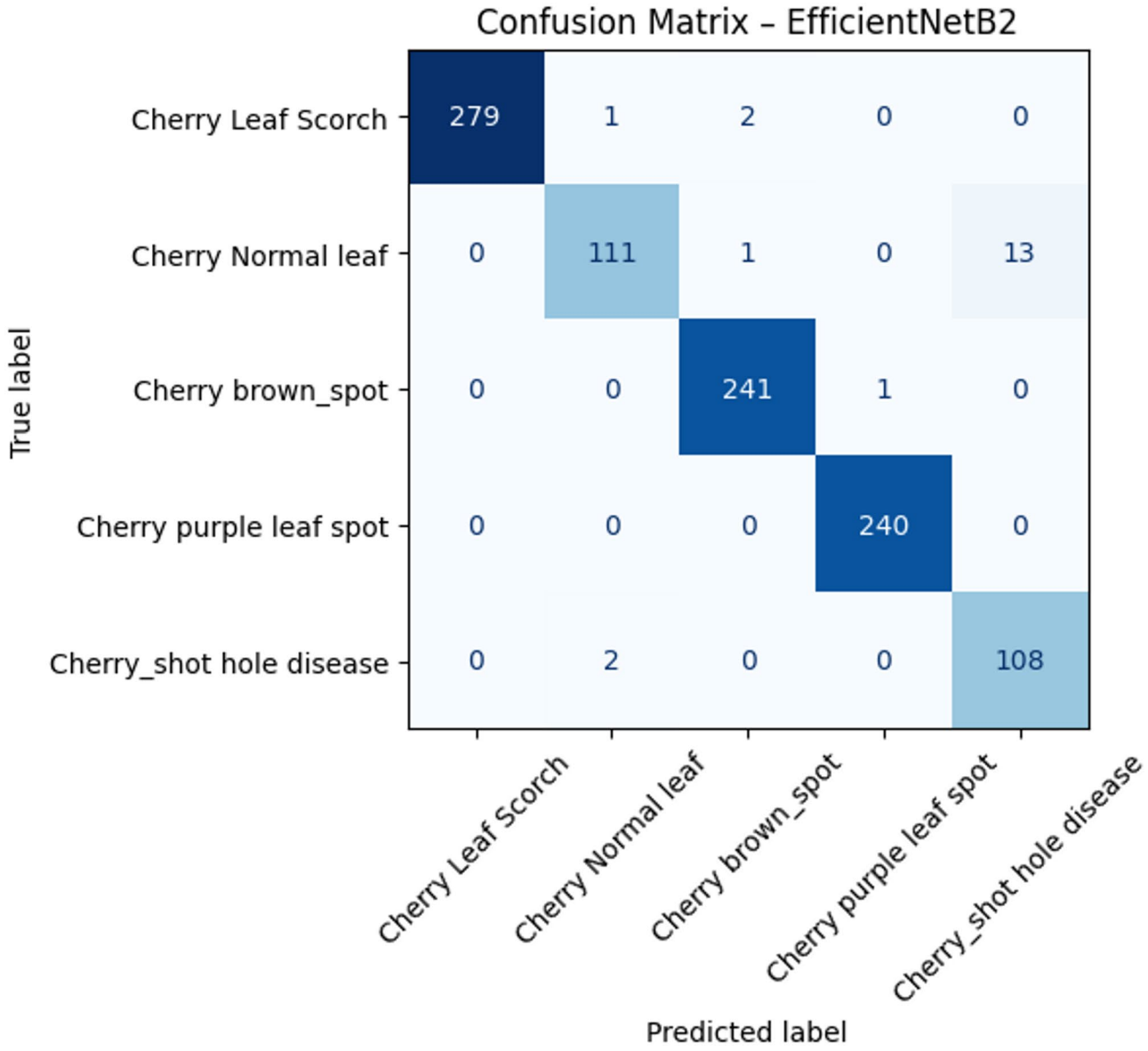

Fig. 10. Confusion Matrix-EfficientNetB2.

Conclusion

This study provides a comprehensive evaluation of deep learning architectures for cherry leaf disease classification by jointly analyzing quantitative performance metrics and explainable artificial intelligence outputs. Among all evaluated models, single-backbone architectures demonstrated the most consistent and robust performance. ResNet50 achieved the highest accuracy (98.20%), followed closely by EfficientNetB2 (98.00%) and DenseNet121 (97.50%), with performance differences confined within a narrow 0.2–0.7% range. These models also exhibited superior class balance, reflected by F1-scores and MCC values exceeding 0.97, and AUC scores above 0.998, indicating strong inter-class discrimination. In contrast, lightweight models such as MobileNetV2 and MobileNetV3Small showed a moderate performance reduction of approximately 1.0–1.2%, while NASNetMobile experienced a substantial decline of nearly 8.0%, underscoring its limited suitability for fine-grained disease differentiation. Dual-backbone architectures failed to surpass single-backbone counterparts, with the best-performing dual model (Dual_ResNet50_MobileNetV2) achieving 97.30% accuracy, corresponding to a 0.9% decrease relative to ResNet50. These findings, supported by Grad-CAM analyses showing more coherent and localized attention maps for single-backbone models, confirm that architectural coherence and stable gradient flow outweigh increased model complexity in fine-grained plant disease classification tasks.

Future work should investigate alternative learning paradigms and architectural designs to further improve performance and interpretability. Recent Ultralytics models such as YOLO11 and YOLO26, along with attention-based architectures, indicate a shift toward more efficient and adaptive feature modeling. Future studies should explore whether such architectures and advanced fusion strategies can provide better performance–interpretability trade-offs in plant disease classification. In particular, Vision Transformer (ViT)

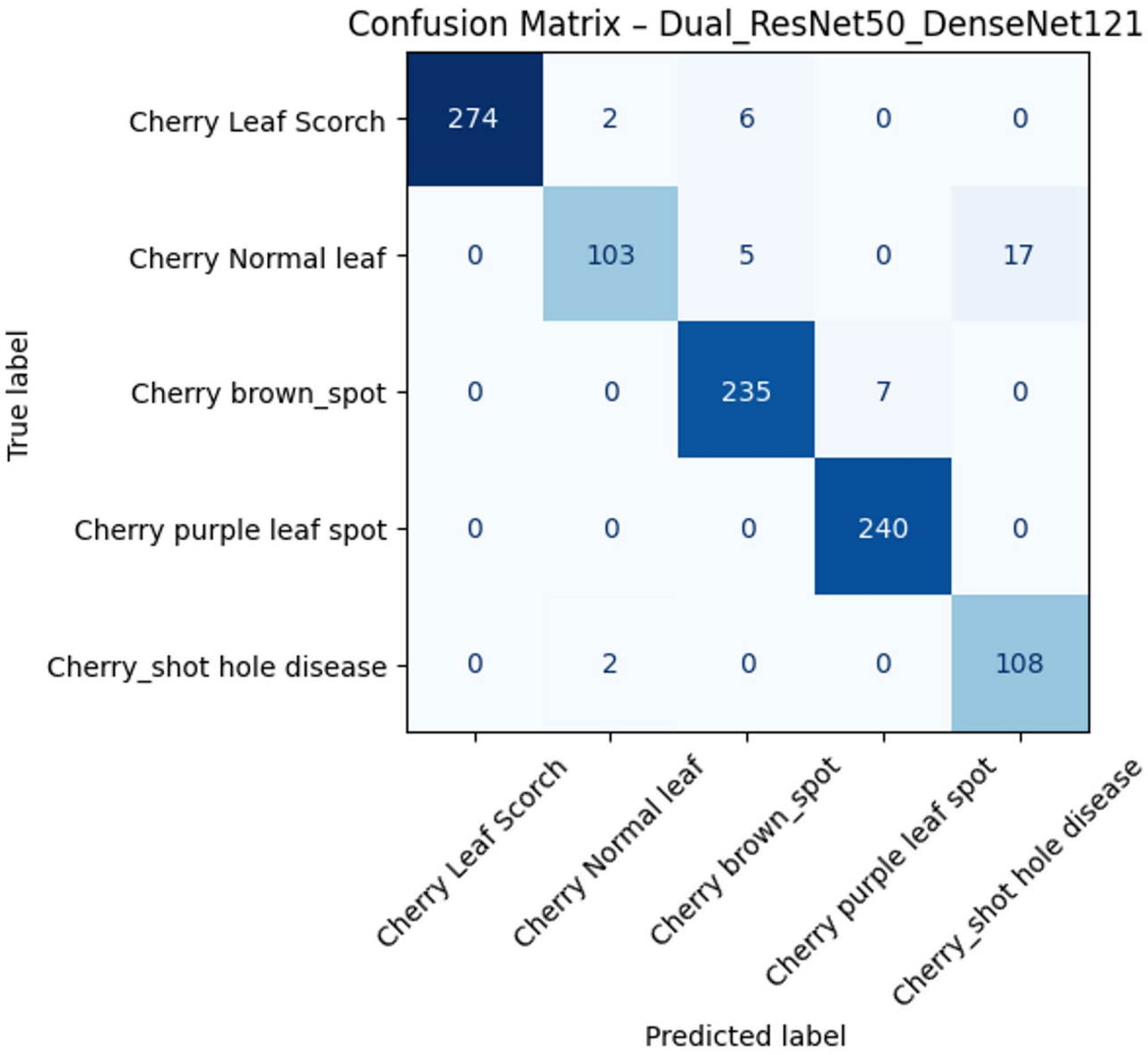

Fig. 11. Confusion Matrix-Dual_ResNet50_DenseNet121.

and CNN–Transformer hybrid models offer promising directions by capturing long-range spatial dependencies that conventional CNNs may overlook. Additionally, architectures such as Swin Transformer, ConvNeXt, and attention-weighted feature fusion networks should be explored to address the limitations observed in naïve multi-backbone designs. Beyond architectural diversity, the proposed framework should be systematically applied to different plant species and agricultural contexts, including leaf diseases in apple, grape, citrus, and tomato crops, as well as field-acquired imagery from drones or mobile devices. Repeating the analysis across diverse datasets and imaging conditions will strengthen the generalizability of the conclusions and support the development of scalable, interpretable, and deployment-ready artificial intelligence systems for precision agriculture.

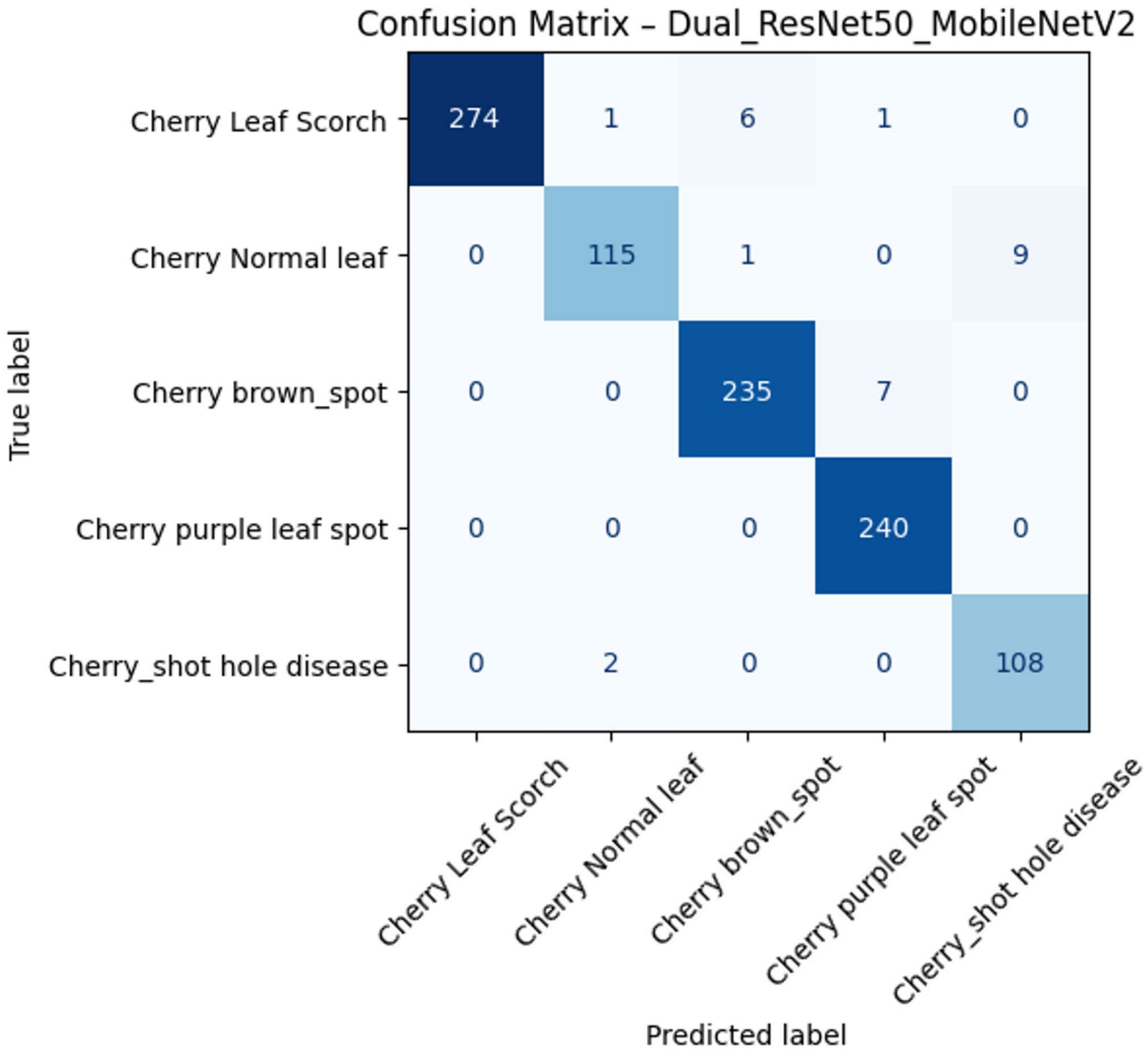

Fig. 12. Confusion Matrix-Dual_ResNet50_MobileNetV2.

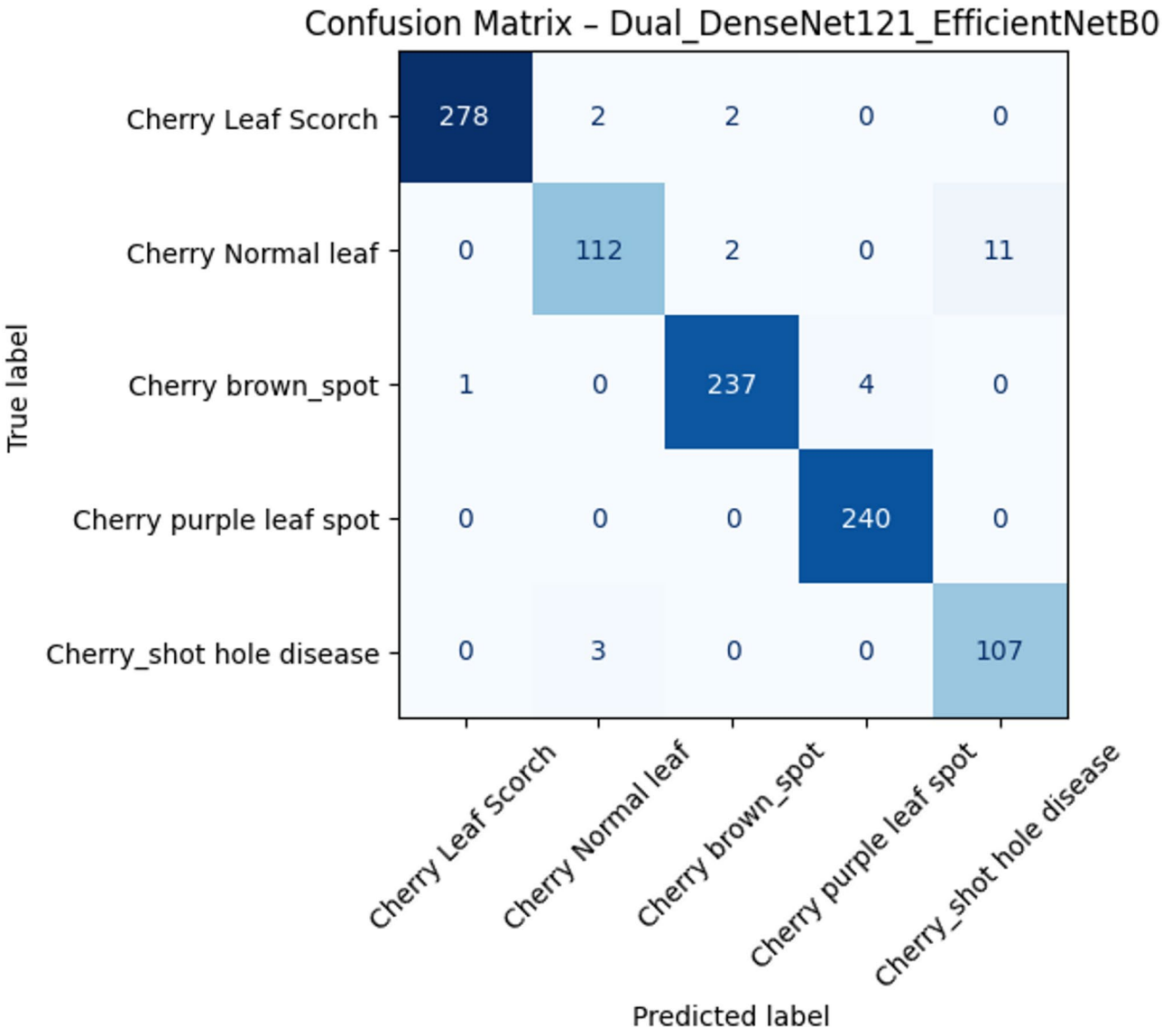

Fig. 13. Confusion Matrix-Dual_ DenseNet121_ EfficientNetB0.

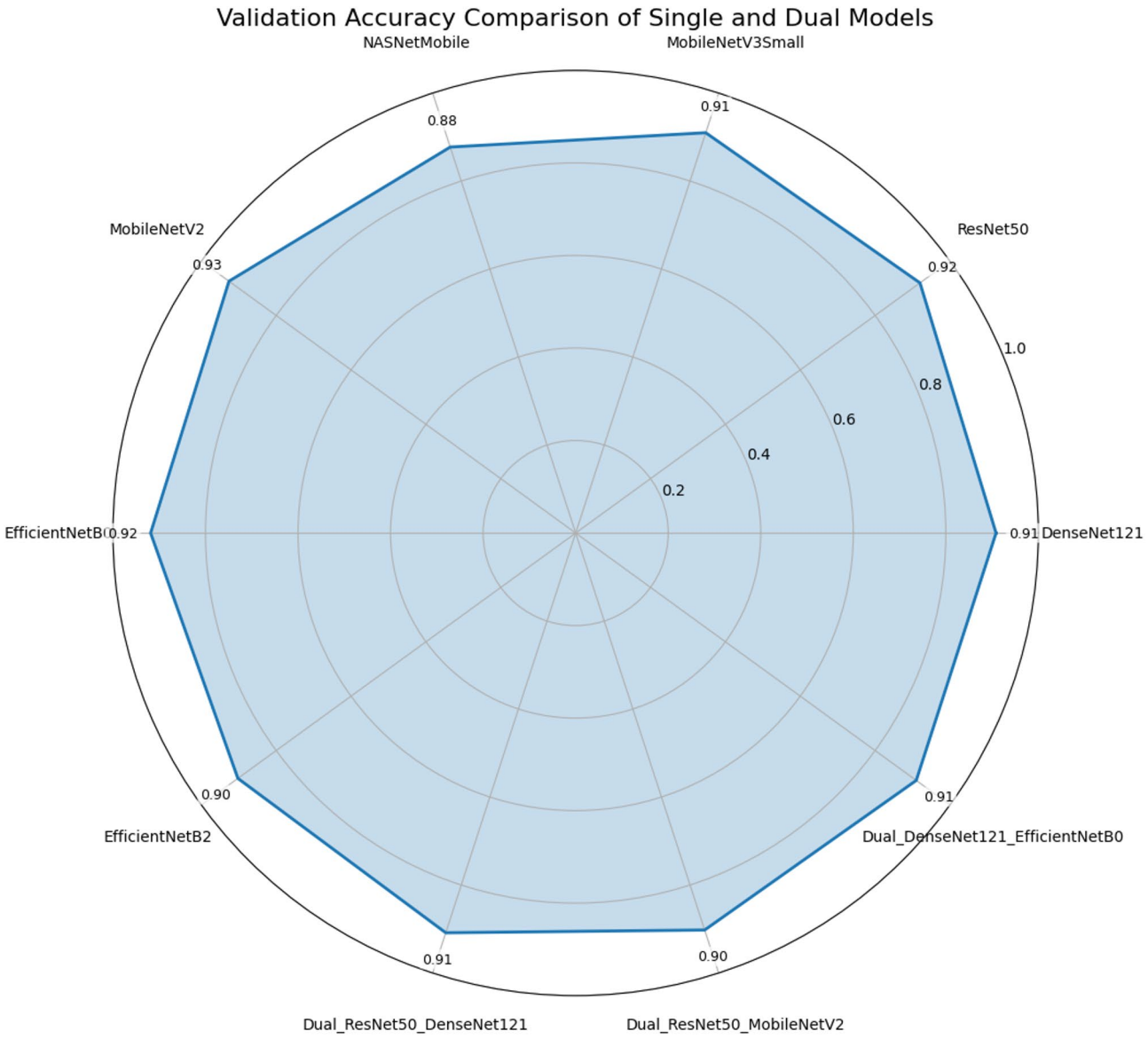

Fig. 14. Radar Chart.

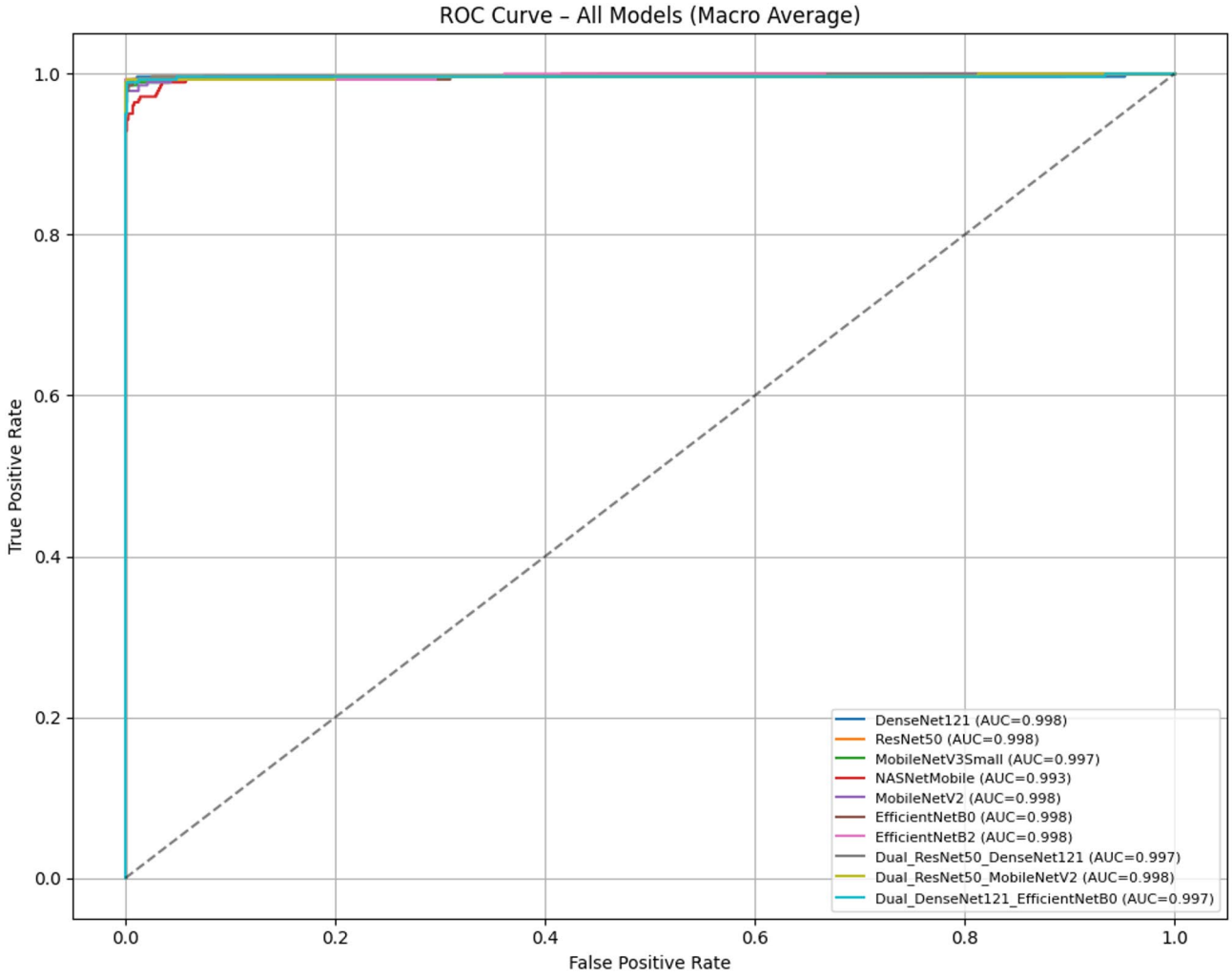

Fig. 15. ROC Curve.

Statistic	Value
Number of paired observations (N)	10
Wilcoxon statistic (W)	0
p-value (two-tailed)	0.00195
Z-score (approx.)	−2.80
Effect size (r = |Z| / √N)	0.89
Effect magnitude	Large

Table 2. Wilcoxon Signed-Rank Test Results and Effect Size (Accuracy vs. Recall).

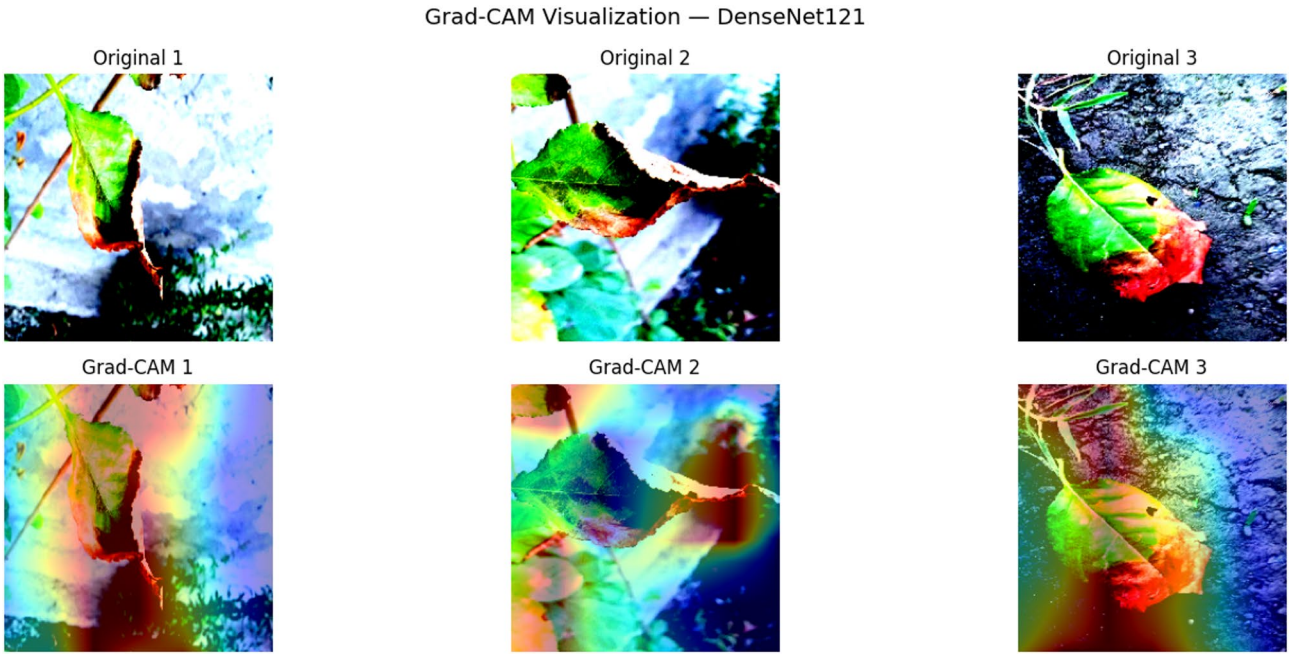

Fig. 16. Grad-Cam Visualization of DenseNet121.

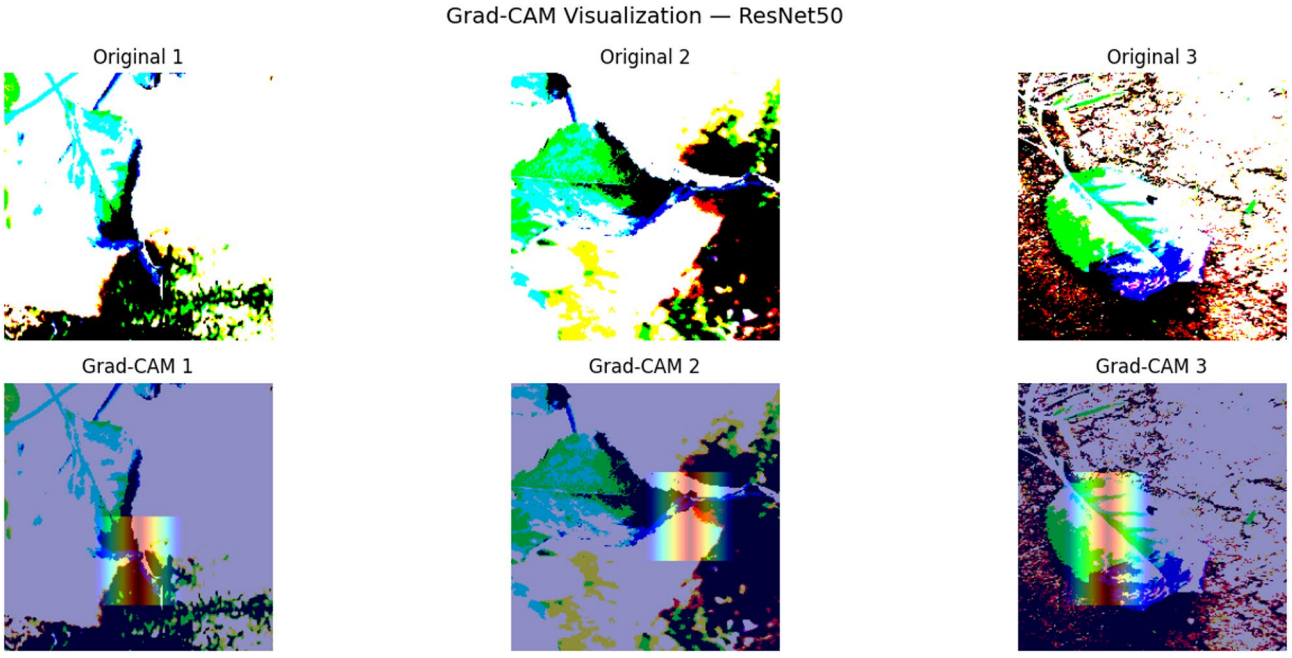

Fig. 17. Grad-Cam Visualization of ResNet50.

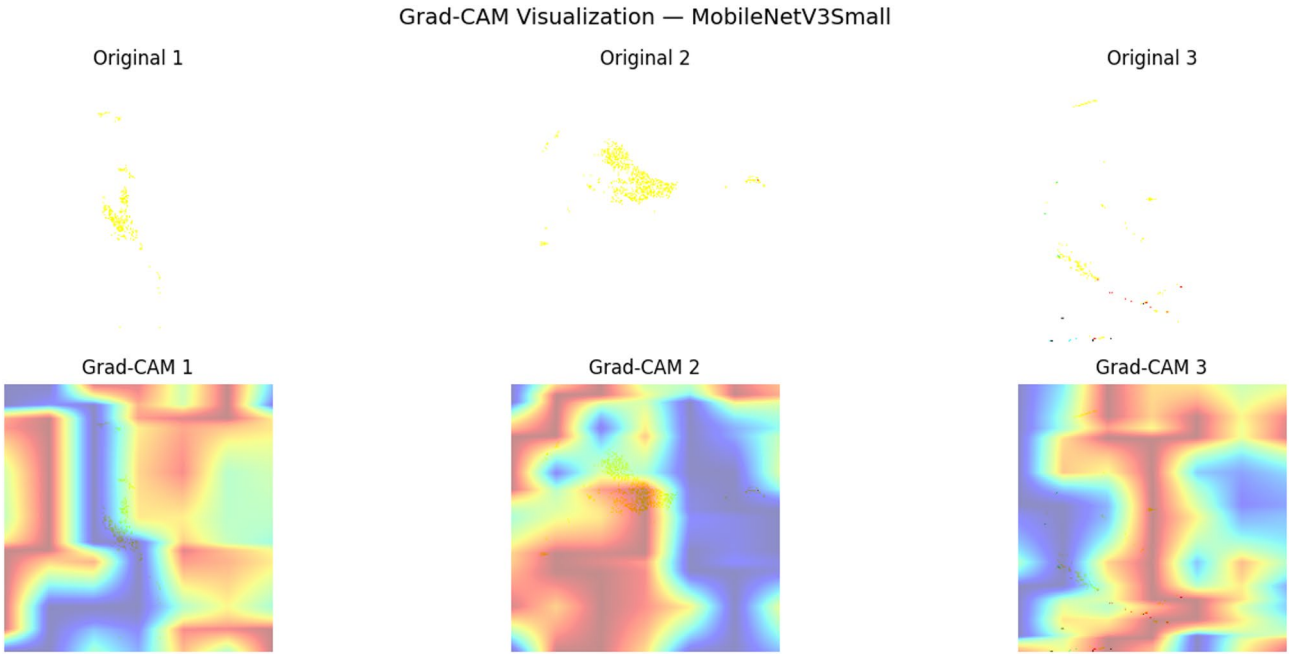

Fig. 18. Grad-Cam Visualization of MobileNetV3Small.

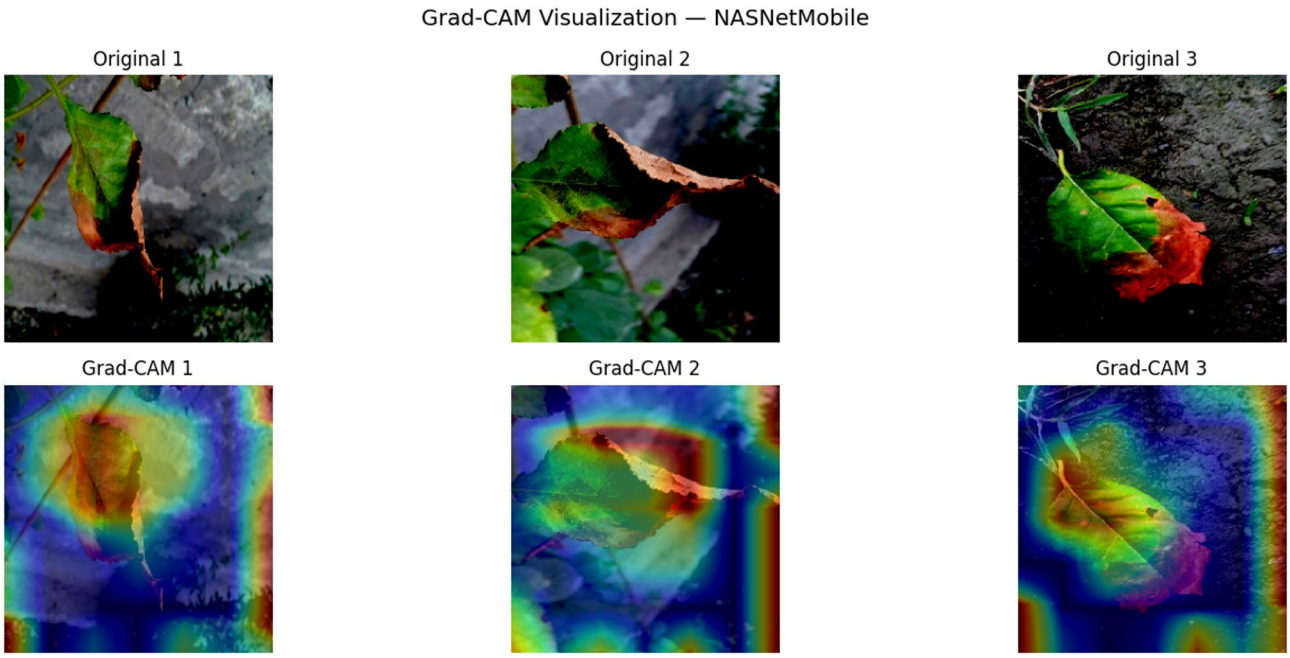

Fig. 19. Grad-Cam Visualization of NASNetMobile.

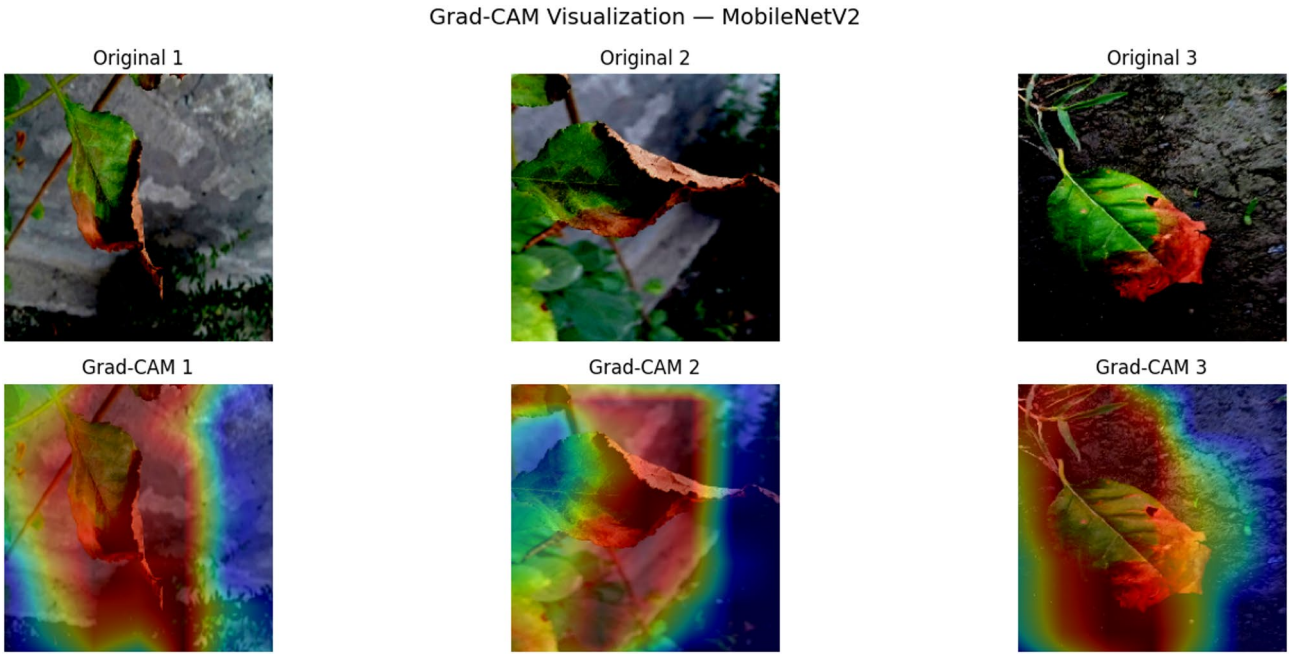

Fig. 20. Grad-Cam Visualization of MobileNetV2.

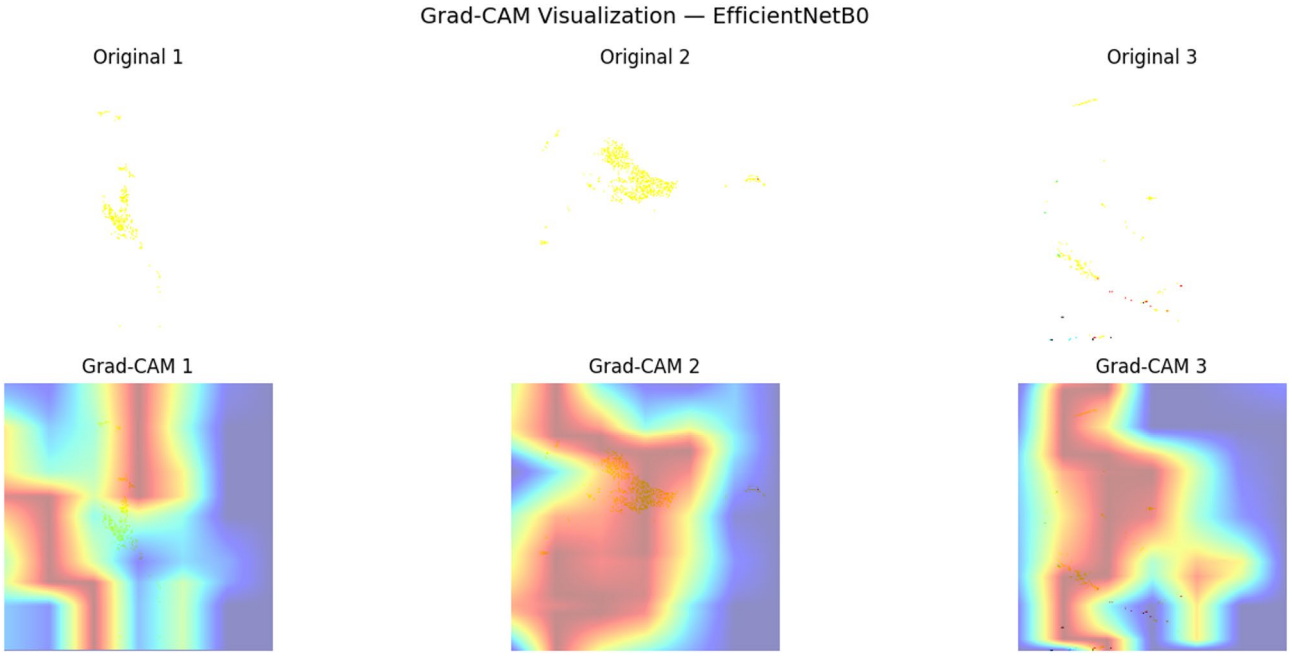

Fig. 21. Grad-Cam Visualization of EfficientNetB0.

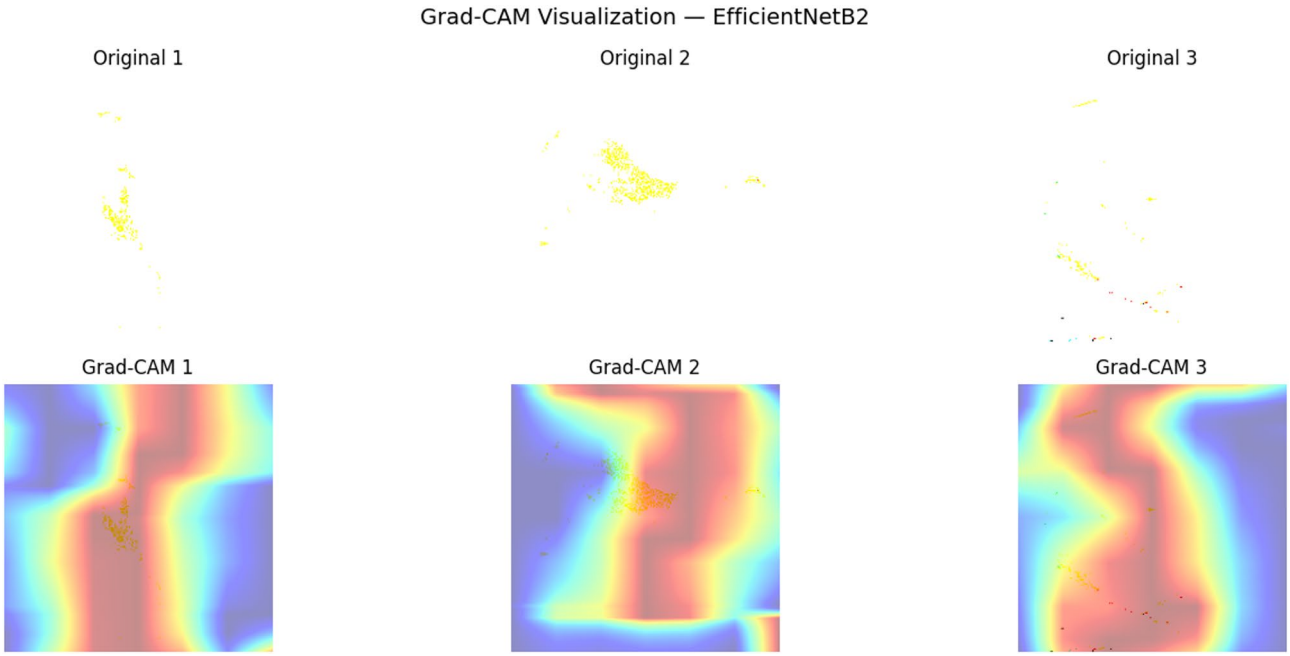

Fig. 22. Grad-Cam Visualization of EfficientNetB2.

Data availability

The image dataset used in this study was obtained from a publicly available secondary source hosted on the Kaggle platform. Specifically, cherry leaf images were accessed from the Cherry Leaf Disease on PlantCity Dataset (2025), which is derived from the PlantCity image collection. The dataset is openly available at: [https://www.kaggle.com/datasets/codewithsk/cherry-leaf-disease-on-plantcity-2025]All images were used in accordance with Kaggle's open data usage policy and were subsequently curated, preprocessed, and partitioned into training, validation, and testing subsets for model development and evaluation. No new data were generated during the current study.

Received: 14 January 2026; Accepted: 20 April 2026
Published online: 21 April 2026

References

1. Ferentinos, K. P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **145**, 311–318 (2018).
2. Bacelar, E. et al. Impacts of climate change and mitigation strategies for some abiotic and biotic constraints influencing fruit growth and quality. *Plants* **13**, 1942 (2024).
3. Sharma, S. & Chauhan, O. *in Fruits and Vegetables Technologies: Postharvest Processing and Packaging* 103–173 (Springer, 2025).
4. Seethapathy, P., Gothandaraman, R., Gurudevan, T. & Malik, I. A. *in Handbook of plum fruit* 133–176 (CRC, 2022).
5. Mohanty, S. P., Hughes, D. P. & Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **7**, 215232 (2016).
6. Zainab, Z. & Mahum, R. Plant disease detection using deep learning techniques. *ICCK J. Image Anal. Process.* **1**, 36–44 (2025).
7. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. big data*. **6**, 1–48 (2019).
8. Pacal, I. et al. A systematic review of deep learning techniques for plant diseases. *Artif. Intell. Rev.* **57**, 304 (2024).
9. Upadhyay, A. et al. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artif. Intell. Rev.* **58**, 92 (2025).
10. Li, C. et al. An advancing GCT-Inception-ResNet-V3 model for arboreal pest identification. *Agronomy* **14**, 864 (2024).
11. Kaya, Y. & Gürsoy, E. A review of deep learning architectures for plant disease detection. *Turkish J. Biology*. **49**, 459–497 (2025).
12. Radhakrishnan, M., Monish, N., Dev, P. S., Kesavan, N. & Thomas, N. S. Implementation of explainable Ai in deep learning methods for multiclass classification of plant diseases in mango lLeaves. *ELCVIA Electron. Lett. Comput. Vis. image Anal.* **24**, 104–117 (2025).
13. Srinivasan, S. et al. DBA-ViNet: an effective deep learning framework for fruit disease detection and classification using explainable AI. *BMC Plant Biol.* **25**, 965 (2025).
14. Goyal, A. & Lakhwani, K. Integrating advanced deep learning techniques for enhanced detection and classification of citrus leaf and fruit diseases. *Sci. Rep.* **15**, 12659 (2025).
15. Van Zyl, C., Ye, X. & Naidoo, R. Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Appl. Energy*. **353**, 122079 (2024).
16. Ercan, U. et al. Identification and Classification of Snack-Type Watermelon (Citrullus lanatus) Genotypes Using Seed Morphology and Machine Learning Techniques. *Foods* **14**, 4069 (2025).
17. Naveen, K. & Parvathi, R. AI powered multi feature fusion framework for retrieving images using color, texture and shape descriptors. *Scientific Reports* (2025).
18. Khan, M. S., Nisa, K., Ahmad, I., Zubair, M. & Alshammari, K. PlantCity: A Comprehensive Image Based on Multi Crop Leaves in Pakistan. *Data Brief*, **112130** https://www.kaggle.com/datasets/codewithsk/cherry-leaf-disease-on-plantcity-2025 (2025). (accessed on 10 January 2026).

19. Khomkham, B. & Pankaseam, Y. Lightweight Convolutional Neural Network Model Based on Modified MobileNetV3 for Plant Disease Classification. *SN Comput. Sci.* **6**, 1028 (2025).
20. Zhao, Z., Bakar, E. B. A. & Razak, N. B. A. & Akhtar, M. N. Corrosion image classification method based on EfficientNetV2. *Heliyon* 10 (2024).
21. Nandhini, S. & Ashokkumar, K. An automatic plant leaf disease identification using DenseNet-121 architecture with a mutation-based henry gas solubility optimization algorithm. *Neural Comput. Appl.* **34**, 5513–5534 (2022).
22. Sumon, R. I. et al. A Deep Learning-Based Approach for Precise Emotion Recognition in Domestic Animals Using EfficientNetB5 Architecture. *Eng* **6**, 9 (2025).
23. Peng, C., Liu, Y., Yuan, X. & Chen, Q. Research of image recognition method based on enhanced inception-ResNet-V2. *Multimedia Tools Appl.* **81**, 34345–34365 (2022).
24. Ahmed, S. F. et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artif. Intell. Rev.* **56**, 13521–13617 (2023).
25. Taye, M. M. Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. *Computation* **11**, 52 (2023).
26. Rajanand, A. & Singh, P. ErfReLU: adaptive activation function for deep neural network. *Pattern Anal. Appl.* **27**, 68 (2024).
27. Jabir, B. & Falih, N. Dropout, a basic and effective regularization method for a deep learning model: A case study. *Indonesian J. Electr. Eng. Comput. Sci.* **24**, 1009–1016 (2021).
28. Kayaalp, K. A deep ensemble learning method for cherry classification. *Eur. Food Res. Technol.* **250**, 1513–1528 (2024).
29. Reyad, M., Sarhan, A. M. & Arafa, M. A modified Adam algorithm for deep neural network optimization. *Neural Comput. Appl.* **35**, 17095–17112 (2023).
30. Anam, M. K., Defit, S., Haviluddin, H., Efrizoni, L. & Firdaus, M. B. Early stopping on CNN-LSTM development to improve classification performance. *J. Appl. Data Sci.* **5**, 1175–1188 (2024).
31. Torghabeh, F. A., Modaresnia, Y. & Khalilzadeh, M. M. Effectiveness of learning rate in dementia severity prediction using VGG16. *Biomedical Engineering: Appl. Basis Commun.* **35**, 2350006 (2023).
32. Dudala, J. M. in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*. 1–7 (IEEE).
33. Sevindik, M. et al. A Comparative Analysis of CNN Architectures, Fusion Strategies, and Explainable AI for Fine-Grained Macrofungi Classification. *Biology* **14**, 1733 (2025).
34. Korkmaz, A. F. et al. Combination Ensemble and Explainable Deep Learning Framework for High-Accuracy Classification of Wild Edible Macrofungi. *Biology* **14**, 1644 (2025).
35. Kumru, E. et al. Explainable convolutional neural network architectures for high-performance taxonomic classification of gasteroid macrofungi. *Sci. Rep.* **15**, 40196 (2025).
36. Kumru, E. et al. Deep Ensemble Learning and Explainable AI for Multi-Class Classification of Earthstar Fungal Species. *Biology* **14**, 1313 (2025).
37. Yang, Y., Khorshidi, H. A. & Aickelin, U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. *Front. Digit. health.* **6**, 1430245 (2024).
38. Jelni, O. S., Radhitya, M. L., Wardhana, G. W., Kusuma, N. W. J. & Desmayani, N. M. M. R. Sentiment Analysis of BRImo Reviews on Google Play Store Using SVM and KNN. *Indonesian J. Data Sci.* **6**, 548–562 (2025).
39. Saheed, Y. K., Balogun, B. F., Odunayo, B. J. & Abdulsalam, M. Microarray gene expression data classification via Wilcoxon sign rank sum and novel Grey Wolf optimized ensemble learning models. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **20**, 3575–3587 (2023).
40. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945).
41. Wilcoxon, F. Probability tables for individual comparisons by ranking methods. *Biometrics* **3**, 119–122 (1947).
42. Lovakov, A. & Agadullina, E. R. Empirically derived guidelines for effect size interpretation in social psychology. *Eur. J. Social Psychol.* **51**, 485–504 (2021).
43. Kansal, K., Chandra, T. B. & Singh, A. ResNet-50 vs. EfficientNet-B0: multi-centric classification of various lung abnormalities using deep learning. *Procedia Comput. Sci.* **235**, 70–80 (2024).
44. Debnath, A. et al. A smartphone-based detection system for tomato leaf disease using efficientNetV2B2 and its explainability with artificial intelligence (AI). *Sensors* **23**, 8685 (2023).
45. Simangunsong, P. B. N., Sihombing, P., Efendi, S. & Fahmi, F. in *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 907–911 (IEEE).
46. Hemavarshini, S. & Arun, R. A. in *2024 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*. 18–22 (IEEE).
47. Salam, A. et al. Mulberry leaf disease detection using CNN-based smart android application. *IEEE Access.* **12**, 83575–83588 (2024).
48. Naskinova, I. Transfer learning with NASNet-Mobile for Pneumonia X-ray classification. *Asian-European J. Math.* **16**, 2250240 (2023).
49. Fauzi, D. R. Comparison of CNN Models Using EfficientNetB0, MobileNetV2, and ResNet50 for Traffic Density with Transfer Learning. *J. Intell. Syst. Technol. Inf.* **1**, 22–30 (2025).
50. Jesupriya, J., Mageswari, P. U. & Alli, A. in *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. 1–9 (IEEE).
51. Tripathi, A., Singh, T., Nair, R. R. & Duraisamy, P. Improving early detection and classification of lung diseases with innovative MobileNetV2 framework. *IEEE Access* (2024).
52. Mutasodirin, M. A. & Falakh, F. M. Efficient weather classification using densenet and efficientnet. *Jurnal Informatika: Jurnal Pengembangan IT*. **9**, 173–179 (2024).
53. An, Q., Chen, W. & Shao, W. A deep convolutional neural network for pneumonia detection in X-ray images with attention ensemble. *Diagnostics* **14**, 390 (2024).
54. Riska, S. Y., Sulistyo, D. A. & Maharani, F. S. S. High-accuracy classification of banana varieties using ResNet-50 and DenseNet-121 architectures. *Indonesian J. Electr. Eng. Comput. Sci.* **39**, 322–335 (2025).
55. Ramzan, M., Bustaman, A., Sarwinda, D. & Ullah, N. in *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 1089–1094 (IEEE).
56. Feng, S., Li, Z., Zhang, B., Chen, T. & Wang, B. DSF2-NAS: Dual-Stage Feature Fusion via Network Architecture Search for Classification of Multimodal Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* (2025).
57. Ennab, M. & Mcheick, H. Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models. *Mach. Learn. Knowl. Extr.* **7**, 12 (2025).
58. Yu, X., Liu, M., Yu, S., Shi, X. & Wang, Q. The Dense Convolutional Network for Rice Disease Recognition Based on Transfer Learning. *Curr. Sci.* **5**, 3595–3606 (2025).
59. Wang, R. et al. MarsTerrNet: A U-Shaped Dual-Backbone Framework with Feature-Guided Loss for Martian Terrain Segmentation. *Remote Sens.* **18**, 35 (2025).
60. Reddy, N. L. & Gopinath, M. Advanced deep learning framework for soil texture classification. *Sci. Rep.* **15**, 34407 (2025).
61. Kadir, M. A., Mosavi, A. & Sonntag, D. in *IEEE 27th International Conference on Intelligent Engineering Systems (INES)*. 000111–000124 (IEEE). 000111–000124 (IEEE). (2023).
62. Kainat, J. et al. Blended Features Classification of Leaf-Based Cucumber Disease Using Image Processing Techniques. *Complexity* 9736179 (2021). (2021).

63. Kondaveeti, H. K. & Simhadri, C. G. Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection. *Sci. Rep.* **15**, 31850 (2025).
64. Georgiou, T., Liu, Y., Chen, W. & Lew, M. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *Int. J. Multimedia Inform. Retr.* **9**, 135–170 (2020).
65. Sahoo, L. et al. Transforming agriculture through artificial intelligence: Advancements in plant disease detection, applications, and challenges. *J. Adv. Biology Biotechnol.* **27**, 381–388 (2024).
66. Manoj, H. et al. AI-driven drone technology and computer vision for early detection of crop disease in large agricultural areas. *Scientific Reports* (2025).

Acknowledgements
None.
Ethics declaration: Not applicable.

Author contributions
Conceptualization, H.T.A., Ö.D., F.E. K.A., M.S., and E.K.; Methodology, M.S.G. and I.A.; Software, Ö.D., K.A., and T.A.; Validation, M.S.G., I.A. K.A., M.S., and F.E.; Formal Analysis, Ö.D., H.T.A. K.A., M.S., and E.K.; Investigation, Ö.D., F.E. and K.A.; Resources, Ö.D., M.S.G., F.E., K.A., I.A and T.A.; Writing – Original Draft Preparation, Ö.D., H.T.A., I.A. and F.E.; Writing – Review & Editing, M.S.G., M.S.,; Visualization, H.T.A., and F.E.; Supervision, M.S.G, I.A. and F.E.

Funding
This research received no external funding.

Declarations

Competing Interests
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.

Additional information
Correspondence and requests for materials should be addressed to M.S.

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

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026