Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2026, 2026:38

https://doi.org/10.28919/cmbn/9797

ISSN: 2052-2541

FUSION OF EFFICIENTNET-B0 AND MOBILENETV2 FOR TOMATO DISEASE CLASSIFICATION

SEBASTIANUS ADI SANTOSO MOLA[1,*], BERTHA SELVIANA DJAHI[1], ANDREA STEVENS KARNYOTO[2,3], CLARISSA ELFIRA AMOS PAH[1], YASINTA LETEK KLEDEN[1], ASNAT NOFRI KENLOPO[1], BENS PARDAMEAN[2,3]

[1]Faculty of Science and Technology, University of Nusa Cendana, Kupang, 85148, Indonesia

[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, BINUS Graduate Program - Master of Computer Science Bina Nusantara University, Jakarta 11480, Indonesia

Abstract: The study proposes a fusion architecture for tomato disease classification that combines two complementary backbones—EfficientNet-B0 and MobileNetV2—via feature concatenation and a hierarchical fusion head. Each branch applies global pooling, dropout, and fully connected layers before feature aggregation; the concatenated representation is then processed by the fusion head to produce the final prediction. Evaluation on an independent test set (n = 1000) shows that the fusion model reduces loss and improves overall accuracy relative to individual backbones, effectively correcting many systematic errors made by MobileNetV2 while achieving performance comparable to EfficientNet-B0. Paired statistical testing using McNemar's test indicates the reduction in misclassifications is practically significant, suggesting the improvements are not attributable to random variation. The concatenation-based aggregation preserves full channel-wise information from both backbones, enabling the fusion head to learn cross-channel interactions and selectively reweight complementary signals to mitigate the weaknesses of each backbone. These findings support the use of feature-level fusion to enhance the robustness and accuracy of plant disease

*Corresponding author

E-mail address: adimola@staf.undana.ac.id

Received January 24, 2026

classification systems.

Keywords: tomato leaves diseases; EfficientNet-B0; MobileNetV2; fusion model; feature concatenation.

2020 AMS Subject Classification: 68T05, 68T07, 68T45.

1. INTRODUCTION

Tomato (*Solanum lycopersicum*) is consumed worldwide due to its rich nutritional content, including vitamins B and E and β-carotene [1]. Euromonitor International reports that tomatoes are processed into a variety of products, including paste, soups, pizza, and ketchup [2]. Given their substantial contributions to food security and health, tomatoes are a vital agricultural commodity. The Food and Agriculture Organization of the United Nations (2025), with major processing by Our World in Data, reports that in 2025, global tomato production was dominated by China with 70.12 million tons, followed by India with 20.43 million tonnes [3].

Plant disease outbreaks are one of the factors affecting tomato production, alongside agronomic factors such as farm management, soil type and fertility, and climate change. According to Ref. [4], tomato diseases primarily attack the leaves. Diseases such as leaf spot, leaf blight, and viral infections can reduce photosynthetic capacity, stunt growth, and even cause plant mortality. Moreover, tomato diseases can spread rapidly and cause substantial yield losses. Acute, widespread outbreaks can reduce tomato production by 20–30%. These production declines have serious economic consequences [5], including reduced farmers' incomes, threats to food security, and potentially higher consumer prices.

Early and accurate diagnosis of tomato diseases is crucial for mitigating their negative impacts and ensuring sustainable production. The main challenges for early detection are farmers' limited knowledge and restricted access to information due to resource constraints, as well as the high diversity of disease types. These challenges often result in delayed intervention, leading to further disease spread and irreversible crop damage before effective measures can be implemented. Diagnosis is typically performed by observing the affected plant parts and then subjectively assessing the disease. This approach is often slow, inaccurate, and inefficient, making it challenging to implement on a large scale in agricultural management.

The advancement of computational technologies has introduced a new paradigm for addressing these challenges. In general, there are two approaches to diagnosing tomato diseases: machine learning (ML) and deep learning (DL). Several ML methods have been reported to deliver strong classification results, such as Support Vector Machine (SVM) [5], [6], Random Forest (RF)[7],

and K-Nearest Neighbor (KNN) [8]. Although ML methods can provide good classification performance, they require preprocessing of leaf images and manual feature extraction. Deep learning overcomes the limitations of traditional, domain-expert-dependent feature engineering by automatically learning features from raw image data.

DL methods offer greater flexibility regarding input images because the feature-extraction step can be bypassed. Convolutional Neural Network (CNN)-based models dominate this approach. The following studies have applied pretrained CNN models such as VGG [9], [10], [11], [12], ResNet[13], [14], [15], [16], AlexNet [17], [18], [19], DenseNet [20], [21], [22], [23], Inception [24], [25], [26], Xception [27], [28], [29], EfficientNet-B0 [30], [31], [32] and MobileNet [33], [34]. These studies indicate that convolutional mechanisms have been widely used to detect tomato diseases.

This study focuses specifically on lightweight architectures: EfficientNet-B0 and MobileNetV2, because they offer favorable trade-offs between accuracy, parameter count, and inference cost, making them suitable for real-world deployment on edge devices and in resource-constrained farming environments. We propose a novel fusion model that extracts features in parallel from both EfficientNet-B0 and MobileNetV2 backbones and then concatenates the resulting feature streams. The rationale for model fusion and concatenation is threefold:

- Complementary representations: different backbones emphasize different inductive biases. Parallel extraction enables the model to capture complementary, multi-scale, and texture-oriented features that a single backbone might miss.
- Richer feature space via concatenation: concatenation preserves the complete set of features from both backbones, ensuring a comprehensive representation. By concatenating feature maps, the fusion head receives a higher-dimensional, information-rich representation that downstream layers can selectively exploit to form stronger, discriminative patterns.
- Practical and flexible aggregation: concatenation is simple and parameter-efficient as an operation (no learnable blending weights required initially), and attention mechanisms to learn optimal combinations without losing raw features at the merge point.

By addressing these aspects, our fusion model aims to achieve superior diagnostic accuracy and robustness, particularly in challenging real-world agricultural settings.

Contributions of this study include: (1) a fusion architecture that combines backbone models in parallel for feature extraction, (2) feature aggregation from two heterogeneous backbones using

concatenation, and (3) a comprehensive evaluation comparing backbone performance and fused model performance. We hypothesize that this fusion approach will enrich image representations and improve classification performance while remaining practical for deployment because the chosen backbones are lightweight.

2. RELATED WORKS

In research on tomato disease classification, both machine learning (ML) and deep learning (DL) approaches have been widely investigated. ML methods offer greater transparency into which hand-crafted features drive classification decisions, but they require an explicit feature-extraction stage before model training. For example, Ref. [5] extracted a comprehensive set of statistical and texture descriptors, including contrast, energy, correlation, mean, homogeneity, entropy, variance, standard deviation, root mean square, skewness, and kurtosis, and found that SVM achieved the highest per-class accuracy among the compared ML classifiers. Similarly, Ref. [6] reported superior SVM performance when features derived from the Gray-Level Co-Occurrence Matrix (GLCM) and Scale-Invariant Feature Transform (SIFT) were used. In contrast, Ref. [7] showed that Random Forest (RF) produced competitive results using KAZE key point descriptors, while Ref. [8] observed that K-Nearest Neighbors (KNN) reached an accuracy of 97%, albeit with the caveat that KNN is an instance-based (lazy) learner and therefore does not yield a compact, generalized model representation. Together, these studies indicate that although ML pipelines can deliver strong accuracy, their reliance on manual feature design and the variation in best-performing algorithms motivate the exploration of end-to-end DL methods and hybrid strategies for more robust, deployable solutions, especially for large-scale, automated agricultural applications.

The application of pre-trained deep learning models, combined with transfer learning, has been widely investigated for tomato disease classification. Several studies highlight the strengths of particular architectures. VGG-19 was employed in Ref. [9], achieving 92.5% accuracy. Ref. [10] also used VGG-19 but incorporated HSV-space image segmentation as preprocessing; freezing the convolutional layers during training yielded an accuracy of 99.72%. A VGG-16 variant with transfer learning yielded 95.71% accuracy in Ref. [11].

Pretrained ResNet models have likewise been applied: Ref. [13] reported ResNet-50 delivering >99% accuracy across three dataset-split scenarios (70:30, 80:20, and 90:10). ResNet-101 was evaluated in Ref. [14] with hyperparameter tuning (learning rate, batch size, and

number of epochs) and produced high accuracy. Ref. [15] compared multiple ResNet variants (ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, Res2NeXt-50, and Res2Net-50d) and found Res2NeXt-50 to be the best performer in terms of classification accuracy.

The performance of alternative backbone architectures—such as AlexNet, DenseNet, Inception, Xception, EfficientNet, and MobileNet—has also been extensively investigated for tomato disease classification. AlexNet, augmented with hand-crafted descriptors such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), demonstrated improvements in Ref. [17], and further gains were reported in Ref. [18] through hyperparameter tuning of the AlexNet backbone. DenseNet variants have also shown strong performance: hyperparameter optimization via Particle Swarm Optimization (PSO) yielded the best results in Ref. 18, and DenseNet201 achieved 98.70% accuracy in Ref. [21].

Inception-family and related architectures were also successful. Ref. [24] proposed a multi-kernel Inception extension, Ref. [25] fine-tuned InceptionV3 with favorable outcomes, and Ref. [26] integrated InceptionV4 into an AlexNet hybrid. In several studies, Inception variants were employed both as pre-trained feature extractors, where the convolutional base produces deep representations that are pooled and fed to lightweight classifiers, and as end-to-end fine-tuned backbones, with the choice depending on dataset size and overfitting risk. Pretrained Xception models achieved 82.89%[27], 85.84% [28], and 90% [29] in different studies. Xception has been used either as a fixed feature extractor (extracting high-level convolutional features for downstream classifiers) or as a fine-tuned backbone, often providing improved representational capacity for texture-rich leaf imagery.

A comprehensive survey of EfficientNet variants (B0–B7) in Ref. [30] identified EfficientNet-B5 as the best performer; other works report high accuracy with EfficientNet-B0 and B1[31], [32]. EfficientNet employs a compound scaling strategy that jointly balances network depth, width, and input resolution, combined with MBConv blocks and squeeze-and-excitation modules to achieve high representational power per parameter. These design choices make EfficientNet particularly effective as a pre-trained feature extractor and as a fine-tuned backbone for leaf disease classification, offering strong multi-scale texture modeling while remaining computationally efficient. For practical trade-offs, B0/B1 are recommended for resource-constrained or real-time deployments, while B5 frequently provides the best accuracy-to-cost balance on larger datasets. For further background and implementation guidance on compound scaling (EfficientNet) [35].

Finally, MobileNet demonstrated an excellent efficiency–accuracy trade-off, achieving

approximately 98% accuracy with very few parameters [33], [34]. In particular, MobileNetV2 introduces inverted residuals and linear bottlenecks in combination with depthwise-separable convolutions, enabling preserved information flow, reduced parameter count, and low computational cost, properties that make it especially suitable as a lightweight backbone or pre-trained feature extractor for edge and mobile deployments. In fusion frameworks, MobileNetV2 can provide complementary, low-latency feature streams that pair well with larger backbones for multi-scale ensemble or attention-based fusion, enabling high overall accuracy with reduced inference time and memory footprint.

3. PRELIMINARIES

3.1 Research Methodology

3.1.1 EfficientNet-B0 Model

EfficientNet-B0 is built on a compact, performance-oriented design that combines mobile inverted bottleneck convolution (MBConv) blocks with squeeze-and-excitation (SE) modules to achieve high representational power with a low parameter and FLOPs budget; this construction preserves both channel and spatial sensitivity while remaining computationally efficient for downstream fine-tuning. The model's defining methodological contribution is compound scaling. This principled rule jointly scales network depth, width, and input resolution using a set of optimized coefficients rather than adjusting any single dimension in isolation. Compound scaling enables a family of models (with B0 as the baseline) to achieve consistently improved accuracy per additional computational cost compared with naïve scaling strategies [35].

The MBConv block is the principal computational unit in EfficientNet: it consists of a 1×1 pointwise convolution for channel expansion, a $k \times k$ depthwise convolution for spatial processing at low FLOPs, an integrated SE module for channel reweighting, and a final 1×1 projection to restore the output channel dimensionality. When the input and output spatial sizes match, and the stride is 1, a residual skip connection is employed. MBConv's inverted-bottleneck pattern increases representational capacity during the expansion phase, while depthwise convolution maintains computational efficiency. The projection and skip preserve low-level information flow. The strategic combination of these elements within the MBConv block enables EfficientNet-B0 to achieve a remarkable balance between high representational power and computational efficiency, making it suitable for tasks that require both accuracy and deployability. The SE module performs channel-wise recalibration by first aggregating each channel via global

average pooling, then passing the resulting vector through a small bottleneck consisting of two fully connected layers with a non-linearity and a sigmoid output to generate per-channel attention weights. These weights scale the input feature maps so that informative channels are emphasized and less relevant channels are suppressed. SE thus captures global channel dependencies and enhances discriminative power with a modest increase in parameter count. MBConv blocks are organized into multiple sequential stages, and some blocks within a stage perform spatial downsampling (stride = 2) to reduce the spatial resolution progressively. This staged design creates a hierarchical feature representation: early stages capture local, high-resolution details; intermediate stages aggregate mid-level patterns; and later stages encode high-level semantic features with large effective receptive fields. The controlled allocation of block repeats and channel widths across stages balances representational richness and computational cost.

3.1.2 MobileNetV2

MobileNetV2 is based on the inverted residual architecture with a linear bottleneck. Each residual module first expands the channel dimension to enable rich nonlinear feature extraction via depthwise separable convolutions, then projects back to a low-dimensional linear bottleneck to avoid information loss through non-linearities. This arrangement retains the benefits of residual connectivity while minimizing computational cost and preserving the integrity of essential low-dimensional manifolds. The architecture was designed explicitly for resource-constrained environments, offering a favorable trade-off between latency and representational power.

Architecturally, MobileNetV2 combines depthwise separable convolutions, an adjustable width multiplier, and compact bottleneck blocks to produce lightweight, fast encoders that maintain stable gradient flow and robust low-level feature preservation. The linear bottleneck mitigates feature collapse that can occur when activations are low-dimensional, while the inverted residual skip connections facilitate optimization and promote feature reuse across layers. These properties yield features that are compact, locally descriptive, and stable under quantization or other deployment optimizations.

3.1.3 Proposed Fusion Model

We propose a parallel fusion architecture (Figure 1) that processes the same input image with two complementary convolutional backbones: EfficientNet B0 and MobileNetV2, whose high-level feature tensors are aligned via lightweight projection layers, then concatenated element-wise to form a unified embedding for downstream classification or dense prediction heads. The parallel design preserves the independent inductive biases of each backbone, allowing the network to learn

modality-specific feature transforms, while the average aggregation enforces an information-balanced representation. This fusion scheme targets a practical trade-off: it retains rich semantic channel interactions from EfficientNet-B0 while preserving the spatially sensitive, low-cost representations characteristic of MobileNetV2, producing a model that is both discriminative and deployable.

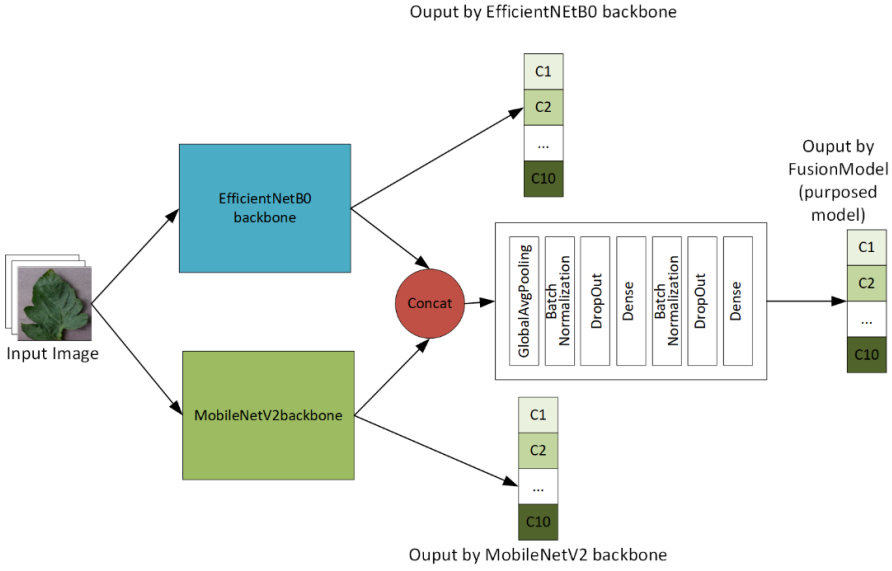


Figure 1. Proposed Fusion Model

When used as one arm of a fusion backbone, EfficientNet-B0 provides high-capacity, channel-rich representations that complement lighter, more linear encoders. Its learned inductive biases, concentrated receptive field aggregation, channel attention via SE, and systematic scale handling make its feature outputs well-suited for downstream fusion strategies that require informative, multi-scale embeddings. Conversely, MobileNetV2 serves as a complementary encoder, supplying concise, linearly preserved features that stabilize and regularize the fused representation. Its inductive bias toward retaining low-dimensional manifold structure and providing efficient spatial detail blends well with the richer, channel-attended embeddings from EfficientNet-B0. In practice, MobileNetV2's low computational footprint allows designers to allocate more capacity to the complementary backbone or to the fusion mechanism itself without exceeding resource constraints, thereby improving the overall scalability and deployability of the fused model.

Global average pooling (GAP) is adopted as the spatial aggregation mechanism because it substantially reduces the number of trainable parameters relative to fully connected classifiers, while establishing a direct correspondence between spatial feature maps and class logits. GAP

enhances generalization and interpretability while mitigating overfitting. Dropout is incorporated as a stochastic regularizer to curb co-adaptation in the terminal classifier and to temper the increased capacity introduced by the dual backbones. Empirically, dropout promotes balanced contributions from each pathway and improves resilience to label noise and distributional shifts. Together, GAP and dropout are critical for managing the increased complexity of a dual-backbone fusion model, ensuring that the combined features yield a generalized, robust classifier rather than overfitting to the training data.

We adopt feature concatenation as the fusion strategy to integrate the representations from the backbones, maintaining the complementary information captured by each backbone without scaling bias. Concatenation merges the feature vectors along the channel dimension, allowing the fused representation to retain all distinctive characteristics captured independently by each backbone. This technique avoids any implicit assumption of linearity or equal contribution, providing flexibility for subsequent layers to learn complex, nonlinear interactions across the combined features. The rationale behind concatenation lies in preserving maximal information diversity, enabling richer feature expressiveness compared to element-wise operations. Technically, concatenation avoids magnitude distortion or bias towards any backbone, as it does not involve direct arithmetic combination that might alter activation scales. This design choice supports stable training dynamics and helps in effectively leveraging the complementary strengths of both EfficientNet-B0 and MobileNetV2.

3.2 Evaluation Method

In the performance evaluation, the principal metrics employed are accuracy, precision, recall (sensitivity), and the F1-score, all of which are derived from the confusion-matrix elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Mathematically, accuracy is defined as

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

For a given class k, precision and recall are defined as

$$Precision_k = \frac{TP_k}{TP_k+FP_k} \tag{2}$$

$$Recall_k = \frac{TP_k}{TP_k+FN_k} \tag{3}$$

respectively, and the class F1-score is the harmonic mean of precision and recall

$$F1-\text{Score}_k = \frac{2 \times Precision_k \times Recall_k}{Precision_k+Recall_k} \tag{4}$$

4. MAIN RESULTS

4.1 Dataset

The study uses the Tomato Leaf Disease Detection dataset [36], which comprises 10,000 tomato leaf images annotated across 10 classes: Mosaic virus, Target spot, Bacterial spot, Tomato Yellow leaf curl virus, Late blight, Leaf mold, Early blight, Two-spotted spider mite, Septoria leaf spot, and Healthy. A balanced dataset ensures reliable performance metrics and a true assessment of the classifier's generalization across all disease types, preventing skew from class imbalance. Each class contains 1,000 images, yielding a balanced class distribution. The dataset was partitioned into training (80%, 8,000 images), validation (10%, 1,000 images), and test (10%, 1,000 images) subsets. Representative examples for each class are shown in Figure 2. The training set was used to fit model parameters, the validation set to monitor learning and tune hyperparameters to mitigate overfitting, and the test set to evaluate final model performance.

4.2 Training

All models were trained using the Adam optimizer with a learning rate of 0.0001, a batch size of 32, and a dropout rate of 0.3 applied to the fully connected heads. Both backbones were initialized with ImageNet-pretrained weights and fine-tuned for up to 20 epochs; the fusion strategy was the concatenation of the two head outputs (Table 1).

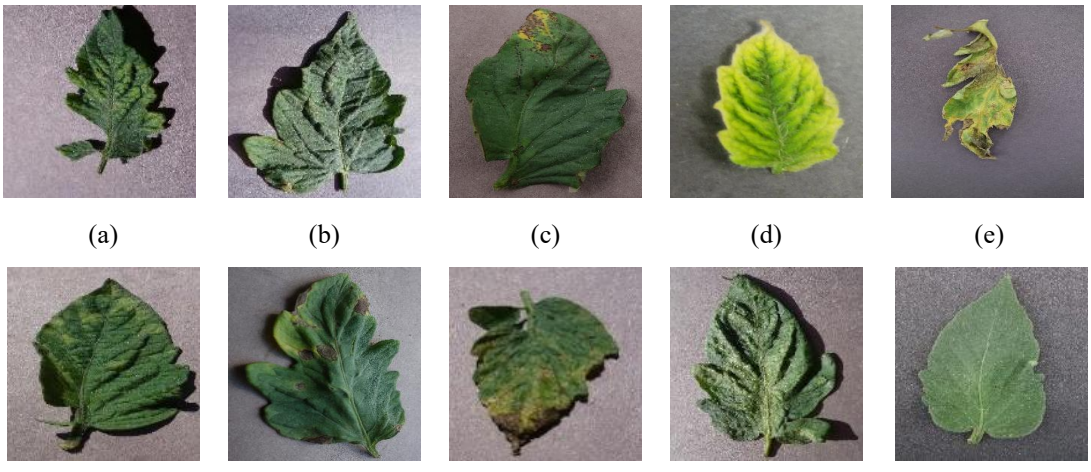


Figure 2. Example images of each class: (a) Mosaic virus, (b) Target spot, (c) Bacterial spot, (d) Tomato Yellow leaf curl virus, (e) Late blight, (f) Leaf mold, (g) Early blight, (h) Septoria leaf spot, (i) Two-spotted spider mite, (j) Tomato Healthy

Table 1. Hyperparameter

Hyperparameter	Value
Learning rate	0.0001
Epoch	20
Optimizer	Adam
Dropout rate	0.3
Batch size	32
Pretrained weight	ImageNet
Fusion strategy	Concatenation

The training curves (Figure 3) for the proposed fusion model exhibit a rapid initial reduction in training loss, followed by a smooth progression toward convergence, with validation metrics closely tracking the training trajectory throughout most epochs. This behavior indicates effective optimization and stable learning dynamics, suggesting that the fusion architecture successfully integrated complementary representations from the constituent backbones without inducing significant generalization gaps. Minor fluctuations observed in the validation trace are limited in magnitude. They are consistent with expected sampling variability or occasional noisy examples in the validation set rather than clear signals of systematic overfitting.

These results suggest that the current hyperparameter choices and fusion strategy yield a model that generalizes well to held-out data, although targeted measures could further enhance robustness. Specifically, employing early stopping based on smoothed validation loss, augmenting borderline cases, and validating on an external dataset would strengthen claims of generalizability.

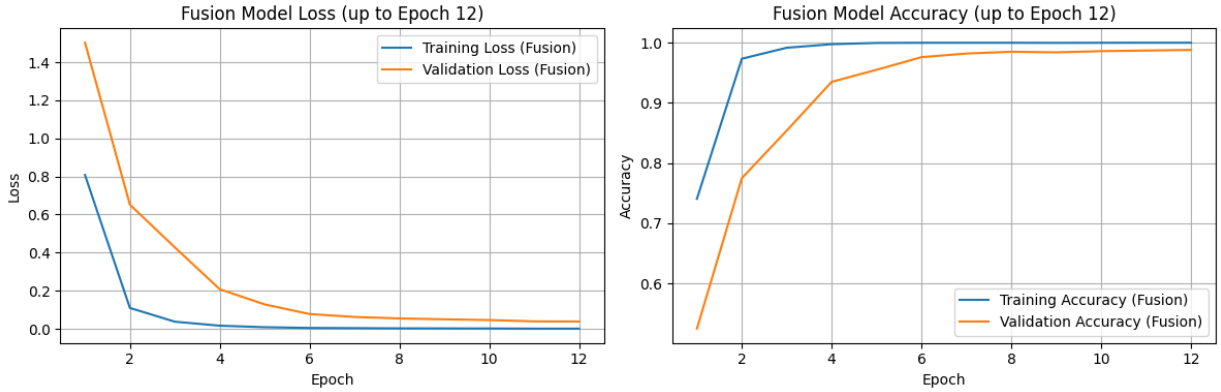


Figure 3. Training and validation curves for the proposed fusion model, showing rapid convergence of training loss/accuracy and stable validation performance, indicative of effective feature integration and good generalization

4.3 Model Performances

4.3.1 EfficientNet-B0 backbone performance

The classification results (Table 2) indicate that the proposed model achieved robust performance across all tomato disease categories, with an overall accuracy of 97.9%. Class-wise precision ranged from 94.90% (Early blight) to 99.01% (Mosaic virus and Healthy). In comparison, recall ranged from 92.0% (Late blight) to 100% across several classes, including Bacterial spot, Yellow leaf curl virus, Mosaic virus, and Healthy. Corresponding F1-scores were similarly high, reaching 99.5% for Mosaic virus and Healthy, and comparatively lower for Early blight (93.94%) and Late blight (95.34%). Macro-average and weighted-average metrics were approximately 97.9%, reflecting consistent performance across classes and balanced predictive behavior relative to the label distribution. Collectively, these findings demonstrate the model's strong discriminative capability for tomato disease detection. Although targeted improvements in sensitivity for Early blight and Late blight may further enhance diagnostic reliability in cases with similar visual manifestations, the model's current performance is notable. Overall, EfficientNet-B0 demonstrates strong accuracy in tomato disease classification, establishing a strong baseline for the fusion model.

Table 2. EfficientNet-B0 backbone classification report

Class	Precision	Recall	F1-Score
Bacterial spot	0.9709	1	0.9852
Early blight	0.949	0.93	0.9394
Late blight	0.9892	0.92	0.9534
Leaf mold	0.9899	0.98	0.9849
Septoria leaf spot	0.99	0.99	0.99
Two-spotted spider mite	0.9802	0.99	0.9851
Target spot	0.9608	0.98	0.9703
Yellow leaf curl virus	0.9804	1	0.9901
Mosaic virus	0.9901	1	0.995
Healthy	0.9901	1	0.995
accuracy	0.979	0.979	0.979
macro avg	0.9791	0.979	0.9788
weighted avg	0.9791	0.979	0.9788

The normalized confusion matrix (Figure 4) for the EfficientNet-B0 backbone demonstrates predominantly strong diagonal values, indicating excellent per-class discrimination and overall model robustness on the test set; several classes (e.g., Bacterial spot, Yellow leaf curl virus, Mosaic

virus, and Healthy) exhibit near-perfect recall, while most others exceed 0.98. Notably, albeit limited, confusions occur between visually similar conditions: Early blight shows a diagonal value of approximately 0.93 with small proportions (~2% and ~4%) misclassified as Bacterial spot and Target spot, respectively, and Late blight has a diagonal near 0.92 with modest misclassification toward Early blight and minor spillover to Leaf mold and Yellow leaf curl virus. These patterns suggest the model reliably separates distinct disease phenotypes but encounters challenges with borderline or visually overlapping presentations. Taken together, the results demonstrate the strong discriminative capability of EfficientNet-B0 for tomato disease identification on the current dataset.

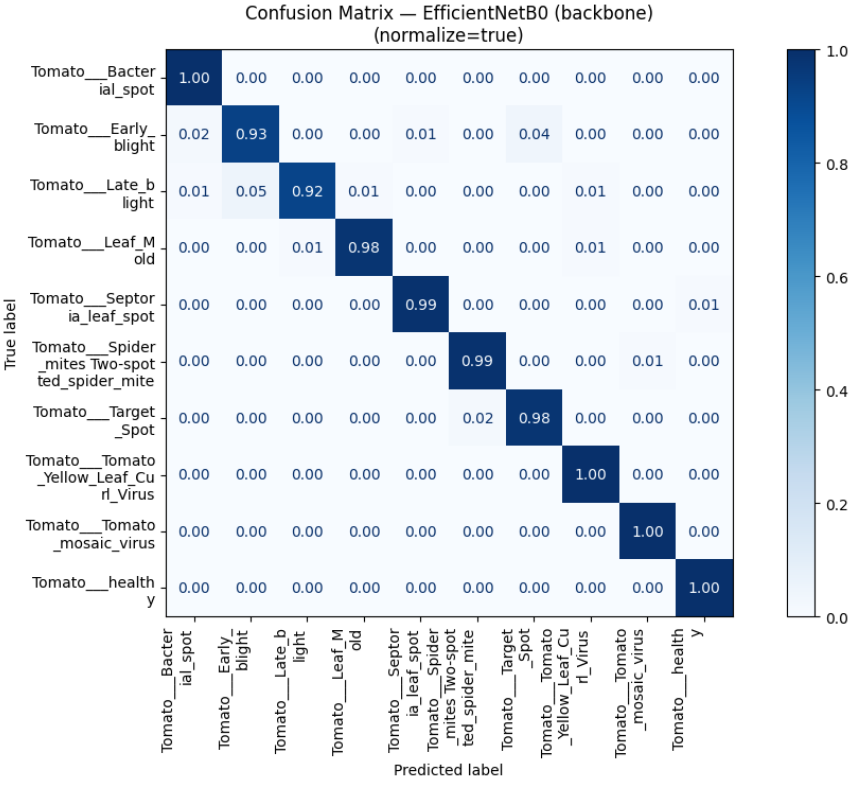


Figure 4. Confusion matrix of EfficientNet-B0 backbone

4.3.2 MobileNetV2 backbone performance

The MobileNetV2 backbone model achieved an overall accuracy of 95.3%, with consistent macro- and weighted-average scores (approximately 95.3% for precision, recall, and F1), indicating balanced and discriminative performance across disease categories. Several classes attained near-perfect metrics. For instance, Mosaic virus and healthy samples achieved 100% recall and an F1 score of 99.01%, while Yellow leaf curl virus achieved 98.99% precision and an F1 score of 98.49%, reflecting the model's capacity to identify conditions with distinctive visual signatures.

Nevertheless, certain classes require further attention: Tomato Early blight showed a relatively low recall (85.0%) despite high precision (93.41%; F1 = 89.01%), whereas Tomato Target spot presented lower precision (87.85%) but high recall (94.0%; F1 = 90.82%), suggesting a tendency toward false negatives for Early blight and ambiguity/false positives for Target spot; Tomato Septoria leaf spot displayed high recall (96.0%) but moderate precision (90.57%; F1 = 93.20%). These error patterns likely stem from visual similarity among lesion types, class imbalance, and intra-class variability (Table 3).

Table 3. MobileNetV2 backbone classification report

Class	Precision	Recall	F1-Score
Bacterial spot	0.9697	0.96	0.9648
Early blight	0.9341	0.85	0.8901
Late blight	0.9688	0.93	0.949
Leaf mold	0.9596	0.95	0.9548
Septoria leaf spot	0.9057	0.96	0.932
Two-spotted spider mite	0.9697	0.96	0.9648
Target spot	0.8785	0.94	0.9082
Yellow leaf curl virus	0.9899	0.98	0.9849
Mosaic virus	0.9804	1	0.9901
Healthy	0.9804	1	0.9901
accuracy	0.953	0.953	0.953
macro avg	0.9537	0.953	0.9529
weighted avg	0.9537	0.953	0.9529

The provided normalized confusion matrix for the MobileNetV2 backbone demonstrates robust per-class classification performance, with most diagonal entries approaching 1.00, indicating near-perfect accuracy for classes such as Bacterial spot, Yellow leaf curl virus, Mosaic virus, and Healthy. Minor confusions are observed (Figure 5): Early blight shows a diagonal value of approximately 0.93 with roughly 2% of instances misclassified as Bacterial spot and 4% as Target spot; Late blight presents a diagonal near 0.92 with about 5% of samples misattributed to Early blight; Leaf mold and Septoria leaf spot retain very high diagonal values (~0.98–0.99) with negligible leakage (e.g., Septoria leaf spot → Healthy ≈1%); and Two-spotted spider mites exhibits a diagonal around 0.99 with approximately 1% confusion toward Mosaic virus.

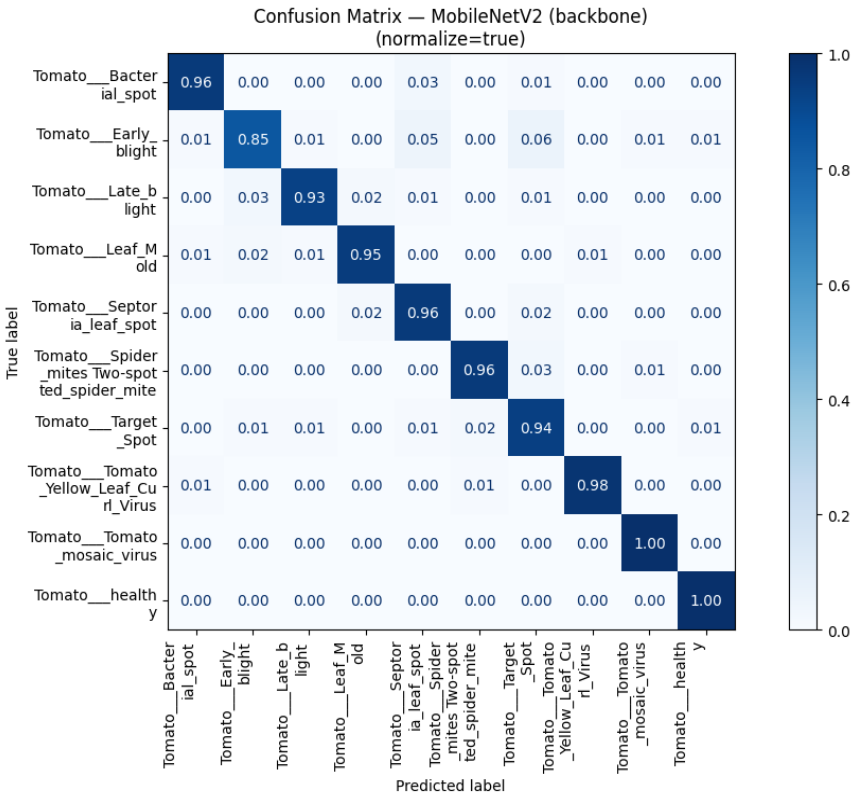


Figure 5. Confusion matrix of MobileNetV2 backbone

4.3.3 Fusion model performance

The fusion model achieved an overall accuracy of 98.2%, with macro- and weighted-average precision, recall, and F1-score converging at approximately 98.22%, indicating consistently high performance across classes (Table 4). Per-class results further underscore this robustness: Bacterial spot, Mosaic virus, and Healthy achieved perfect recall (1.00) with F1-scores ≥ 0.985, while Leaf mold and Yellow leaf curl virus exhibited precision and F1 values close to 0.99, reflecting strong discriminability. A slightly lower recall for Early blight (0.94) and modestly reduced precision for Target spot (0.9515) indicate isolated false negatives and false positives, respectively; yet, both classes retain high F1-scores (≥ 0.959), suggesting a limited practical impact on overall performance. Collectively, these metrics demonstrate the effectiveness of the fusion strategy in delivering balance and high-fidelity classification.

Table 4. Fusion model classification report

Class	Precision	Recall	F1-Score
Bacterial spot	0.9804	1	0.9901
Early blight	0.9792	0.94	0.9592
Late blight	0.9898	0.97	0.9798
Leaf mold	0.99	0.99	0.99
Septoria leaf spot	1	0.97	0.9848
Two-spotted spider mite	0.98	0.98	0.98
Target spot	0.9515	0.98	0.9655
Yellow leaf curl virus	0.99	0.99	0.99
Mosaic virus	0.9709	1	0.9852
Healthy	0.9901	1	0.995
accuracy	0.982	0.982	0.982
macro avg	0.9822	0.982	0.982
weighted avg	0.9822	0.982	0.982

The normalized confusion matrix for the fusion model (Figure 6) demonstrates near-perfect class discrimination, with dominant diagonal entries equal to 1.00 for Bacterial spot, Mosaic virus, Yellow leaf curl virus and Healthy, and majority of remaining classes exhibiting true-positive rates between 0.92 and 0.99; Early blight shows a reduced true-positive rate (~0.93) with small proportions of instances misclassified as Bacterial spot and Target spot, while Late blight (≈0.92) presents limited confusion primarily toward Early blight and marginally toward Yellow leaf curl virus. Leaf mold, Septoria leaf spot, Two-spotted spider mites, and Target spot each retain high per-class fidelity (≈0.98–0.99), with only trace off-diagonal leakage to visually related classes (e.g., Leaf mold → Tomato mosaic virus; Septoria leaf spot → healthy; Spider mites → Tomato mosaic virus; Target spot ← Septoria leaf spot). The localized, and anatomically plausible pattern of errors suggests that residual misclassifications are driven by intrinsic lesion-level ambiguity and intra-class variability rather than by systematic model bias.

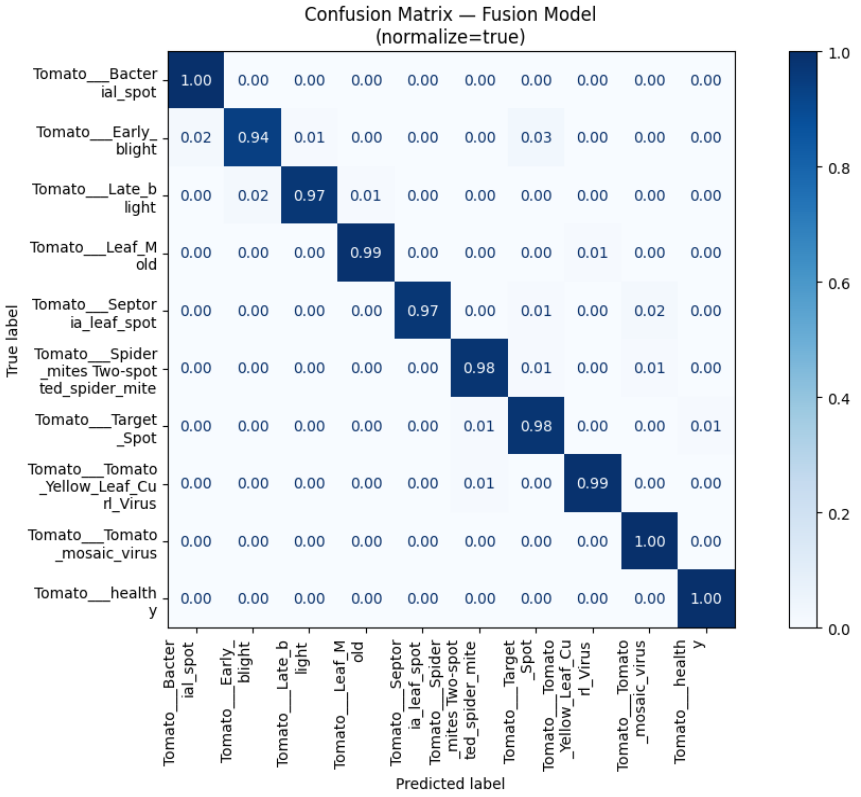


Figure 6. Confusion matrix of the fusion model

Table 5 shows that the fusion model achieved the highest performance, with an accuracy of 98.2% (982 correct, 18 incorrect), outperforming EfficientNet-B0 (97.9% (979 correct, 21 incorrect)) and MobileNetV2 (95.3% (953 correct, 47 incorrect)).

Table 5. Model accuracy comparison.

Model	Accuracy	Data Test		
		Correctly classified	Misclassified	Total
EfficientNet-B0	97.9	979	21	1000
MobileNetV2	95.3	953	47	1000
Fusion model	98.2	982	18	1000

The absolute difference between the fusion model and EfficientNet-B0 is slight (0.3 percentage points; a reduction of 3 misclassifications, ≈14.3% relative reduction), whereas the improvement relative to MobileNetV2 is more pronounced (2.9 percentage points; a reduction of 29 misclassifications, ≈61.7% relative reduction). These results indicate that the fusion approach provides a consistent improvement in overall accuracy and error reduction compared with single-

backbone models, most notably against MobileNetV2. At the same time, the advantage over EfficientNet-B0 is modest (Table 6).

Table 6. Misclassified by EfficientNet-B0 and MobileNetV2 but correctly predicted by the fusion model.

Model	Misclassified (by model)	Correctly classified by the Fusion model
EfficientNet-B0	21	11
MobileNetV2	47	33

Based on Tables 7 and 8, the McNemar's paired test (n = 1000) showed no statistically significant difference between the fusion model and EfficientNet-B0 (b = 11, c = 8; x^2_{cont}= 0.21, p_{exact} = 0.648), indicating comparable classification performance. In contrast, the fusion model significantly outperformed MobileNetV2 (b = 33, c = 4; x^2_{cont} = 21.19, p_{exact} = 4.2×10^{-6}), demonstrating that the fusion approach corrected substantially more MobileNetV2 errors (two-sided exact test, α = 0.05). These statistical findings formally confirm that, while the fusion model achieves performance on par with the high-performing EfficientNet-B0, its primary strength lies in systematically overcoming the limitations of the more lightweight MobileNetV2, making it a robust and efficient solution for practical deployment.

Table 7. 2×2 contingency table of paired classification outcomes for Fusion vs EfficientNet-B0 on the test set (n = 1000).

	The fusion model misclassified	The fusion model correctly classified	Total
EfficientNet-B0 misclassified	10	11	21
EfficientNet-B0 correctly classified	8	971	979
Total	18	982	1000

Table 8. 2×2 contingency table of paired classification outcomes for Fusion vs MobileNetV2 on the test set (n = 1000).

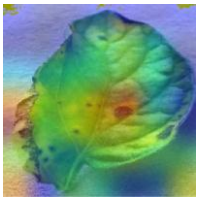
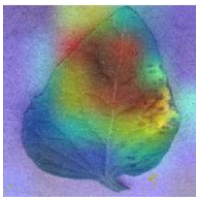
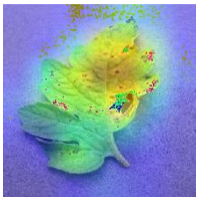
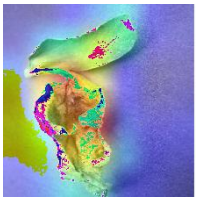
	The fusion model misclassified	The fusion model correctly classified	Total
MobileNetV2 misclassified	14	33	47
MobileNetV2 correctly classified	4	949	953
Total	18	982	1000

4.3.3 Visualization

Grad-CAM [37] visualizations (Table 9) indicate that the model frequently localizes to lesion-specific features for classes characterized by distinct, high-contrast symptoms. For instance, Target spot and Early blight show strong, concentrated activations in concentric necrotic rings; Septoria leaf spot presents multiple spatially distributed saliency peaks aligned with small spot lesions; and Late blight yields broad, contiguous activations that follow extensive necrotic or water-soaked regions. Bacterial Spot and Leaf mold likewise produce heatmaps that correlate with discrete speckles or patchy-textured areas on the lamina. These lesion-driven localizations demonstrate the model's ability to exploit class-specific visual signatures when lesions are sufficiently salient and spatially well-defined.

Concurrently, Grad-CAM also reveals systematic reliance on structural or positional cues in cases with diffuse, low-contrast, or morphology-dominant symptoms. For viral and diffuse conditions (e.g., Mosaic virus and Yellow leaf curl virus), activations often extend to mosaic-like chlorotic regions or shift toward leaf margins and petioles; healthy leaves elicit more diffuse laminar attention rather than focal hotspots; and subtle, low-contrast damage, such as Two-spotted spider mites feeding, yields comparatively dispersed saliency. These patterns imply a dual behavior of the classifier: robust lesion-focused discrimination where visual pathology is prominent, and compensatory use of leaf shape, venation, or framing as proxies where lesions are less localized or less contrastive. Together, these findings highlight both the model's capacity to capture diagnostically relevant visual cues and its tendency to leverage morphological/contextual correlates, with direct implications for per-class interpretability and case-level reliability of automated diagnoses.

Table 9. Grad-CAM visualization

Class	Original Images	Grad-CAM Images	Class	Original Images	Grad-CAM Images
Tomato Mosaic virus			Leaf mold		
Target spot			Early blight		
Bacterial spot			Septoria leaf spot		
Yellow leaf curl virus			Late blight		
Healthy			Two-spotted spider mite		

CONCLUSION

The fusion architecture leverages the complementary strengths of its constituent backbones to produce more robust classification. EfficientNet-B0 provides compact yet expressive visual representations that capture higher-resolution features and complex structural patterns. At the same time, MobileNetV2 supplies lightweight, texture- and contrast-sensitive features that can be especially informative for subtle lesions. By concatenating features, the model preserves complete channel-wise information from both backbones, allowing downstream fusion layers to learn optimal cross-channel interactions and reweight or suppress signals as needed, rather than

discarding them through early averaging or reduction. This design enables the fusion head to correct errors introduced by one backbone using complementary signals from the other backbone. Paired statistical testing (McNemar's test) corroborates these observations: the fusion model's overall accuracy was comparable to that of EfficientNet-B0, while it significantly reduced the misclassifications made by MobileNetV2, indicating that the observed gains reflect a genuine correction of systematic errors rather than random sample variation.

For future work, we recommend: dataset curation and augmentation, and external validation and systematic testing on additional, independent datasets (including datasets with different acquisition conditions, geographic origins, or disease prevalence) to assess generalization and robustness to domain shift.

ACKNOWLEDGMENTS

This research was supported by the Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology, Republic of Indonesia, with contract number: 092/C3/DT.05.00/PL/2025. The authors thank the Directorate General for its financial support. The study used the Open Data Tomato leaf disease detection dataset (https://www.kaggle.com/datasets/kaustubhb999/tomatoleaf).

AUTHOR CONTRIBUTIONS

Conceptualization, S.A.S.M. and B.S.D.; methodology, S.A.S.M., A.N.K.; software, C.E.A.P. and B.S.D.; validation, A.S.K. and C.E.A.P.; formal analysis, S.A.S.M. and Y.L.K.; investigation, S.A.S.M. and Y.L.K.; resources, S.A.S.M. and B.S.D.; data curation, B.S.D. and S.A.S.M.; writing—original draft preparation, A.S.K., Y.L.K., B.P. and S.A.S.M.; writing—review and editing, Y.L.K., S.A.S.M., A.S.K., A.N.K., B.S.D., and C.E.A.P.; Visualization, A.S.K. and Y.L.K.; supervision, S.A.S.M. and B.P.; project administration, S.A.S.M.; funding acquisition, S.A.S.M. All authors have read and agreed to the published version of the manuscript.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] L. Giri, M. Hussain, J.C. Angmo, G. Mustafa, B. Singh, et al., Enhancing Tomato (Solanum Lycopersicum) Yield and Nutrition Quality Through Hydroponic Cultivation with Treated Wastewater, Food Chem. 463 (2025),

141079. https://doi.org/10.1016/j.foodchem.2024.141079.

[2] Euromonitor International, Global Market Report Processsed Tomatoes, Euromonitor International, 2023. https://tomatonews.com/wp-content/uploads/2025/09/Global-Market-Report-Processed-Tomatoes.pdf.

[3] R. Hannah, P. Rosado, M. Roser, Data Page: Tomato Production, Data Page: Tomato Production, 2025. https://archive.ourworldindata.org/20250909-093708/grapher/tomato-production.html.

[4] S.W. Zhang, Y.J. Shang, L. Wang, Plant Disease Recognition Based on Plant Leaf Image, 2015, J. Anim. Plant Sci. 25 (2015), 42-45.

[5] S. Sagar, J. Singh, An Experimental Study of Tomato Viral Leaf Diseases Detection Using Machine Learning Classification Techniques, Bull. Electr. Eng. Inform. 12 (2023), 451-461. https://doi.org/10.11591/eei.v12i1.4385.

[6] R. Khan, N. Ud Din, A. Zaman, B. Huang, Automated Tomato Leaf Disease Detection Using Image Processing: An SVM‐Based Approach with GLCM and SIFT Features, J. Eng. 2024 (2024), 9918296. https://doi.org/10.1155/2024/9918296.

[7] T. Loganayaki, M. Poongothai, Tomato Foliage Disease Recognition System Using Random Forest and Convolutional Neural Networks, in: Algorithms for Intelligent Systems, Springer, Singapore, 2023: pp. 333-343. https://doi.org/10.1007/978-981-99-1431-9_26.

[8] S.S. Harakannanavar, J.M. Rudagi, V.I. Puranikmath, A. Siddiqua, R. Pramodhini, Plant Leaf Disease Detection Using Computer Vision and Machine Learning Algorithms, Glob. Transitions Proc. 3 (2022), 305-310. https://doi.org/10.1016/j.gltp.2022.03.016.

[9] C. Sharma, G. Ansari, K. Yadav, S.K. Shah, A. Alkhayyat, Detecting Alternaria Solani in Tomatoes: Identification with VGG-19 Deep Learning for Early Detection, in: A. Tomar, S. Mishra, Y.R. Sood, P. Kumar (Eds.), Proceedings of 4th International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication, Springer Nature Singapore, Singapore, 2024: pp. 1–11. https://doi.org/10.1007/978-981-97-5231-7_1.

[10] T.H. Nguyen, T.N. Nguyen, B.V. Ngo, A VGG-19 Model with Transfer Learning and Image Segmentation for Classification of Tomato Leaf Disease, AgriEngineering 4 (2022), 871-887. https://doi.org/10.3390/agriengineering4040056.

[11] A.S. Paymode, V.B. Malode, Transfer Learning for Multi-Crop Leaf Disease Image Classification Using Convolutional Neural Network VGG, Artif. Intell. Agric. 6 (2022), 23-33. https://doi.org/10.1016/j.aiia.2021.12.002.

[12] S.A.S. Mola, T.D.I.D. Ole, A.S. Karnyoto, N.F.T. Udju, A.L. Hipir, B. Pardamean, Fine-Tuning VGG16 Model for Driver Behavior Classification, Commun. Math. Biol. Neurosci. 2025 (2025), 47. https://doi.org/10.28919/cmbn/9165.

[13] M. Muslih, A.D. Krismawan, Tomato Leaf Diseases Classification Using Convolutional Neural Networks with Transfer Learning Resnet-50, Kinetik: Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control. 9 (2024), 149-158. https://doi.org/10.22219/kinetik.v9i2.1939.

[14] T.A. Prasetyo, T.L. Gaol, N.F. Sipahutar, T. Siahaan, T.E. Manik, et al., Refining Tomato Disease Recognition: Hyperparameter Tuning on Resnet-101 Architecture for Precise Leaf-Based Classification, Indones. J. Electr. Eng. Comput. Sci. 34 (2024), 1204-1213. https://doi.org/10.11591/ijeecs.v34.i2.pp1204-1213.

[15] I. Kunduracioglu, Utilizing ResNet Architectures for Identification of Tomato Diseases, J. Intell. Decis. Mak. Inf. Sci. 1 (2024), 104-119. https://doi.org/10.59543/jidmis.v1i.11949.

[16] C.L. Candra, Anderies, D. Guan, T.W. Cenggoro, B. Pardamean, Analysing Potential of ResNet for Transfer Learning with Stochastic Depth, in: Y. Tian, T. Ma, M.K. Khan (Eds.), Big Data and Security, Springer, Singapore, 2024: pp. 127–137. https://doi.org/10.1007/978-981-97-4387-2_10.

[17] J. Qiu, X. Lu, X. Wang, C. Chen, Y. Chen, et al., Research on Image Recognition of Tomato Leaf Diseases Based on Improved AlexNet Model, Heliyon 10 (2024), e33555. https://doi.org/10.1016/j.heliyon.2024.e33555.

[18] D.H. Senbatu, B.S. Girma, Y.M. Ayano, Tomato Leaf Disease Detection and Classification Using Custom Modified AlexNet, in: T. Girma Debelee, A. Ibenthal, F. Schwenker (Eds.), Pan-African Conference on Artificial Intelligence, Springer Nature Switzerland, Cham, 2023: pp. 95–113. https://doi.org/10.1007/978-3-031-31327-1_6.

[19] A. Mitchell, E. Edbert, G.N. Elwirehardja, B. Pardamean, Offline Signature Verification Using Transfer Learning and Data Augmentation on Imbalanced Dataset, ICIC Express Lett. 17 (2023), 359-366. http://doi.org/10.24507/icicel.17.03.359.

[20] C.A.D. Lestari, S. Anam, U. Sa'adah, Tomato Leaf Disease Classification with Optimized Hyperparameter: A DenseNet-PSO Approach, in: K. Dharmawan, N.A. Sanjaya Er (Eds.), Proceedings of the First International Conference on Applied Mathematics, Statistics, and Computing (ICAMSAC 2023), Atlantis Press, 2024: pp. 228–239. https://doi.org/10.2991/978-94-6463-413-6_23.

[21] M. Bakr, S. Abdel-Gaber, M. Nasr, M. Hazman, DenseNet Based Model for Plant Diseases Diagnosis, Eur. J. Electr. Eng. Comput. Sci. 6 (2022), 1-9. https://doi.org/10.24018/ejece.2022.6.5.458.

[22] T. Lu, B. Han, L. Chen, F. Yu, C. Xue, A Generic Intelligent Tomato Classification System for Practical Applications Using Densenet-201 with Transfer Learning, Sci. Rep. 11 (2021), 15824. https://doi.org/10.1038/s41598-021-95218-w.

[23] B. Pardamean, H.H. Muljo, T.W. Cenggoro, B.J. Chandra, R. Rahutomo, Using Transfer Learning for Smart Building Management System, J. Big Data 6 (2019), 110. https://doi.org/10.1186/s40537-019-0272-6.

[24] H. Sun, C. Fan, X. Gai, M.A. Al-Absi, S. Wang, et al., Multi-Kernel Inception Aggregation Diffusion Network for Tomato Disease Detection, BMC Plant Biol. 24 (2024), 1069. https://doi.org/10.1186/s12870-024-05797-9.

[25] Z. Li, C. Li, L. Deng, Y. Fan, X. Xiao, et al., Improved AlexNet with Inception-V4 for Plant Disease Diagnosis, Comput. Intell. Neurosci. 2022 (2022), 5862600. https://doi.org/10.1155/2022/5862600.

[26] A. Saeed, A.A. Abdel-Aziz, A. Mossad, M.A. Abdelhamid, A.Y. Alkhaled, et al., Smart Detection of Tomato Leaf Diseases Using Transfer Learning-Based Convolutional Neural Networks, Agriculture 13 (2023), 139. https://doi.org/10.3390/agriculture13010139.

[27] L. Cleetus, A.R. Sukumar, N. Hemalatha, Computational Prediction of Disease Detection and Insect Identification Using Xception Model, bioRxiv: 2021.08.10.455608, (2021). https://doi.org/10.1101/2021.08.10.455608.

[28] H.T. Vo, N.N. Thien, K.C. Mui, Tomato Disease Recognition: Advancing Accuracy Through Xception and Bilinear Pooling Fusion, Int. J. Adv. Comput. Sci. Appl. 14 (2023), 1045-1051. https://doi.org/10.14569/ijacsa.2023.01408113.

[29] N. Arifin, Maratuttahirah, J. Rusman, M.F. Rasyid, Leaf Disease Detection in Tomato Plants Using Xception Model in Convolutional Neural Network Method, J. Tek. Inform. (Jutif) 5 (2024), 571-577. https://doi.org/10.52436/1.jutif.2024.5.2.1926.

[30] Ü. Atila, M. Uçar, K. Akyol, E. Uçar, Plant Leaf Disease Classification Using EfficientNet Deep Learning Model, Ecol. Inform. 61 (2021), 101182. https://doi.org/10.1016/j.ecoinf.2020.101182.

[31] K.K.C. Gonzales, I.A.M. Dioses, EfficientNet Convolutional Neural Network Approach in Classifying Multiple Tomato Diseases, in: 2024 IEEE 12th Conference on Systems, Process and Control (ICSPC), IEEE, 2024, pp. 257-262. https://doi.org/10.1109/icspc63060.2024.10862516.

[32] S.A. Tanim, A.R. Aurnob, Z.H. Anik, M.I. Hossain, Precise Detection of Tomato Leaf Diseases Using Deep Learning Approach with EfficientNet, in: 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023, pp. 1-6. https://doi.org/10.1109/iccit60459.2023.10441130.

[33] H.T. Nguyen, H.H. Luong, L.B. Huynh, B.Q.H. Le, N.H. Doan, et al., An Improved MobileNet for Disease Detection on Tomato Leaves, Adv. Technol. Innov. 8 (2023), 192-209. https://doi.org/10.46604/aiti.2023.11568.

[34] T. Abdullahi, G. George, A. Shehu, A Comprehensive Evaluation of Mobilenet Architecture for Tomato Diseases, Open J. Phys. Sci. 5 (2024), 18-31. https://doi.org/10.52417/ojps.v5i1.585.

[35] M. Tan, Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, arXiv:1905.11946, 2019. https://doi.org/10.48550/arXiv.1905.11946.

[36] B. Kaustubh, Tomato Leaf Disease Detection, 2025. https://www.kaggle.com/datasets/kaustubhb999/tomatoleaf.

[37] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et al., Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization, Int. J. Comput. Vis. 128 (2019), 336-359. https://doi.org/10.1007/s11263-019-01228-7.