Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

Benchmarking CNN and Vision Transformer Architectures for Corn Leaf Disease Classification on the Kaggle Maize Dataset

Juni Ismail[1], Raja Anan Nasution[2], Evi Handayani[2], Annisa Shafira Zuhri[3,*]

[1]Computer Engineering Study Program, Politeknik Bisnis Indonesia, Pematangsiantar, Indonesia
[2]Informatics Management Study Program, Akademi Manajemen Informatika dan Komputer ITMI, Medan, Indonesia
[3]Computer Science Study Program, Universitas Potensi Utama, Medan, Indonesia
Email: [1]juniismailll@gmail.com, [2]rajanasutionnnn@gmail.com, [3]handayanievi744@gmail.com, [4,*]zuhriannisa1@gmail.com
Email Penulis Korespondensi: juniismaill1@gmail.com[*]
Submitted: **06/04/2026**; Accepted: **08/05/2026**; Published: **30/06/2026**

Abstract-Foliar diseases in corn pose a critical constraint on agricultural productivity, particularly in developing countries. Deep learning-based automated detection has emerged as a viable alternative to conventional manual inspection. This study presents a comparative evaluation of four contemporary deep learning architectures EfficientNet-B3, MobileNetV3-Large, ResNet50, and Vision Transformer Small (ViT-Small) on the publicly available Corn or Maize Leaf Disease Dataset hosted on Kaggle (4,188 image samples; four classes: Blight, Common Rust, Gray Leaf Spot, and Healthy). Class imbalance was addressed through a combination of WeightedRandomSampler and Focal Loss, while all architectures were trained via transfer learning from ImageNet pretrained weights, augmented with MixUp and CutMix. Experimental results demonstrate that ViT-Small achieved the highest classification performance, attaining 97.14% accuracy, a weighted F1-Score of 0.9716, and an AUC-ROC of 0.9961, outperforming EfficientNet-B3 (96.66%), MobileNetV3-Large (96.18%), and ResNet50 (95.71%). As an external reference, these results are also compared indicatively with the DenseNet121 accuracy (93.48%) reported by Waheed et al. (2020); it must be emphasized that this baseline was not reproduced in the present experiments, and therefore the comparison should be interpreted as indicative rather than conclusive. McNemar's test confirmed that ViT-Small's superiority is statistically significant ($p<0.05$). An ablation study verified the positive contribution of the Focal Loss and WeightedRandomSampler combination. Grad-CAM visualization corroborated that all models direct their attention to pathologically relevant lesion regions.

Keywords: Deep Learning; Corn Leaf Disease Classification; Vision Transformer; Transfer Learning; Ablation Study

1. INTRODUCTION

Among the world's staple cereal crops, corn (Zea mays L.) occupies a pivotal position as a versatile resource utilized for human consumption, livestock nutrition, and diverse industrial processes [1]. The Food and Agriculture Organization (FAO) reported that worldwide corn output surpassed 1.2 billion tonnes during the 2023 growing season, with a substantial proportion originating from agricultural economies in the developing world [2]. Nevertheless, corn production remains persistently vulnerable to a spectrum of foliar pathologies that pose considerable risks to both harvest quantity and grain quality. The most economically damaging diseases afflicting corn include Northern Leaf Blight (NLB), Common Rust triggered by the fungal pathogen Puccinia sorghi, and Gray Leaf Spot attributable to Cercospora zeae-maydis [3]. When these conditions remain undiagnosed and untreated, the resulting crop losses can range between 20% and 50%, contingent upon infection severity and the growth stage at which disease onset occurs [4].

Conventional approaches to disease identification depend predominantly on manual visual assessment performed by trained agricultural specialists a process that is inherently slow, resource-demanding, and susceptible to observer inconsistency. Such manual protocols become increasingly unfeasible in large-scale cultivation contexts where extensive acreage demands continuous surveillance [5]. The scarcity of qualified plant pathologists in remote agricultural regions of developing nations compounds this limitation, frequently resulting in diagnostic delays that allow infections to progress unchecked. These constraints underscore the pressing demand for intelligent, automated diagnostic tools capable of delivering rapid and reliable disease assessments to growers [6].

The emergence of deep learning paradigms, most notably Convolutional Neural Networks (CNNs), has fundamentally transformed visual recognition capabilities across a wide range of application domains, with plant pathology being a particularly active area of adoption [7]. The transfer learning methodology wherein feature representations acquired from large-scale benchmark datasets like ImageNet are repurposed for domain-specific tasks has demonstrated exceptional utility in agricultural imaging scenarios characterized by limited annotated data availability [8]. A growing body of literature has explored deep learning-driven approaches for automated corn leaf disease recognition, yielding increasingly encouraging outcomes.

In an early investigation, Priyadharshini et al. [9] designed a purpose-built CNN framework for maize foliar disease identification, obtaining a classification accuracy of 87.57%. Although their work validated the viability of convolutional approaches for this task, the moderate accuracy attained underscored the necessity for

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

architecturally superior solutions. Waheed et al. [10] advanced the field by deploying DenseNet121 through transfer learning, achieving 93.48% accuracy on a comparable corn disease corpus. Despite this notable performance gain, their analysis was confined to a single network topology without cross-architectural benchmarking against contemporary alternatives. Chen et al. [11] introduced an enhanced VGGNet variant for crop disease recognition that attained 92.1% accuracy; however, their methodology did not account for recent paradigm shifts in network design, particularly Vision Transformers and computationally efficient architectures optimized for mobile inference.

In more recent work, Paymode and Malode [12] performed a multi-model evaluation of transfer learning strategies for plant foliar disease recognition, examining ResNet, VGG, and Inception family architectures. Their experimental results positioned ResNet50 as the top performer at 94.3% accuracy, though their comparative scope excluded both lightweight edge-deployable networks and transformer-based architectures. Sharma et al. [13] investigated EfficientNet configurations for agricultural disease classification, reporting performance improvements over conventional CNNs with 95.8% accuracy across a heterogeneous multi-crop dataset. Nonetheless, their research prioritized cross-crop generalizability over granular analysis of disease-specific classification patterns within individual crop species such as corn.

A significant deficiency persists in the current literature: the lack of holistic comparative investigations that concurrently assess architecturally diverse paradigms spanning traditional CNN-based networks, resource-efficient models engineered for edge computing scenarios, and the increasingly prominent Vision Transformer (ViT) family within the specific context of corn foliar disease classification. The prevailing research tendency is to evaluate either a solitary architecture or to restrict comparisons within the convolutional network family, thereby neglecting the considerable potential of attention-based transformer models that have achieved breakthrough performance in general-purpose visual recognition benchmarks [14]. Moreover, the class distribution imbalance endemic to agricultural imaging datasets where particular pathological conditions occur with markedly lower frequency remains insufficiently addressed in most published works, and systematic interpretability analysis via gradient-based visualization methods such as Grad-CAM [15] is rarely incorporated.

Equally underexplored is the systematic assessment of computational overhead associated with candidate architectures under realistic deployment constraints. In practical agricultural contexts particularly within the resource-limited infrastructure characteristic of farming operations in the developing world a thorough understanding of the interplay between predictive performance and computational demands is indispensable for informed model selection across target platforms ranging from handheld devices and unmanned aerial systems to edge computing nodes [16].

The principal contributions of this study can be articulated as follows. First, we present a systematic comparative evaluation of four architecturally diverse deep learning paradigms compound-scaled CNN (EfficientNet-B3), edge-optimized CNN (MobileNetV3-Large), residual CNN (ResNet50), and Vision Transformer (ViT-Small) within a unified experimental protocol on the Corn or Maize Leaf Disease Dataset. Second, we propose a joint imbalance-mitigation strategy combining Focal Loss with WeightedRandomSampler, validated through ablation analysis to quantify the individual and combined contributions of each component. Third, we provide a deployment-oriented computational profiling that complements accuracy metrics with inference latency, throughput, and memory footprint, yielding actionable platform-specific deployment guidance. Fourth, we apply McNemar's statistical significance testing to substantiate model-comparison claims, addressing a methodological gap commonly overlooked in prior plant pathology classification studies.

In response to these identified gaps, the present study undertakes a systematic comparative investigation across four architecturally distinct deep learning frameworks, each embodying a fundamentally different design philosophy: EfficientNet-B3 (compound scaling strategy), MobileNetV3-Large (neural architecture search-optimized mobile design), ResNet50 (skip connection-based residual learning), and ViT-Small (patch-based self-attention processing). The research objectives are fourfold: (1) to benchmark and comparatively evaluate the diagnostic accuracy of these four architectures on the corn leaf disease classification task; (2) to quantify the efficacy of a combined WeightedRandomSampler and Focal Loss strategy for addressing distributional class imbalance; (3) to deliver model interpretability insights through Grad-CAM visualization of learned feature representations; and (4) to conduct a computational efficiency analysis informing deployment feasibility across diverse hardware platforms. The outcomes of this investigation are intended to advance the development of robust, transparent, and practically deployable plant disease diagnostic systems aligned with precision agriculture objectives.

2. RESEARCH METHODOLOGY

2.1 Research Methodology

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

The experimental design adopted in this investigation follows a structured five-phase workflow encompassing: dataset procurement with exploratory characterization, image preprocessing coupled with augmentation strategies, architectural configuration of candidate models, training execution with advanced optimization techniques, and multi-dimensional performance evaluation. The complete computational pipeline was developed in Python 3.10 utilizing the PyTorch 2.0 deep learning framework, executed on Google Colab Pro infrastructure provisioned with an NVIDIA Tesla T4 GPU providing 15.6 GB of dedicated video memory. Half-precision floating-point computation (FP16) was activated across all experimental runs to enhance throughput and reduce memory utilization.

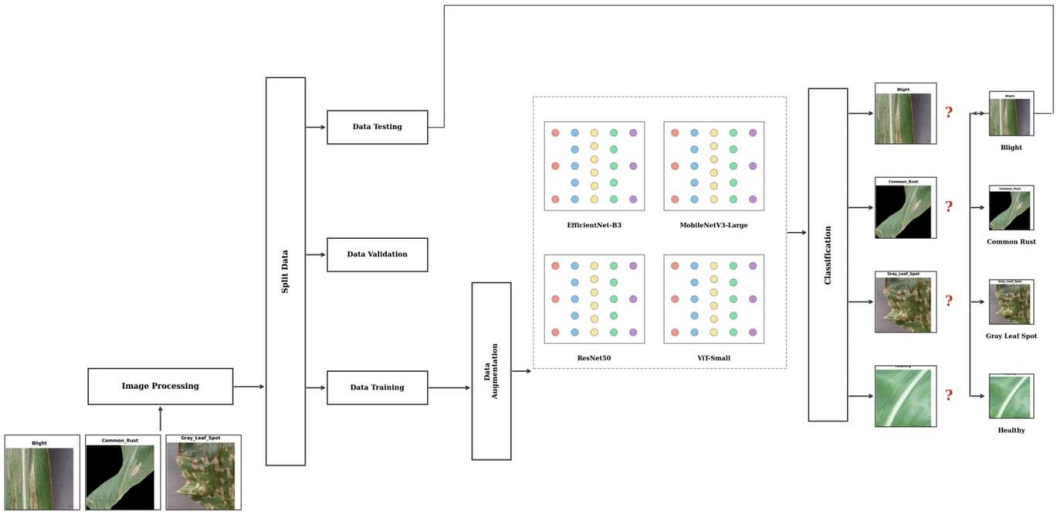

Figure 1. Research Methodology

2.2 Dataset

This investigation utilized the publicly accessible Corn or Maize Leaf Disease Dataset hosted on the Kaggle platform, comprising 4,188 color (RGB) photographic samples distributed among four categorical labels: Common Rust (1,306 specimens, 31.2%), Healthy (1,162 specimens, 27.7%), Blight (1,146 specimens, 27.4%), and Gray Leaf Spot (574 specimens, 13.7%). The collection exhibits a pronounced distributional asymmetry, with a 2.28:1 ratio between the most represented class (Common Rust) and the least represented class (Gray Leaf Spot). Spatial dimensions across the corpus vary considerably, spanning from 180×116 to 5,184×5,184 pixels, with an average resolution of 308×298 pixels. The aggregate storage requirement amounts to 163.2 MB.

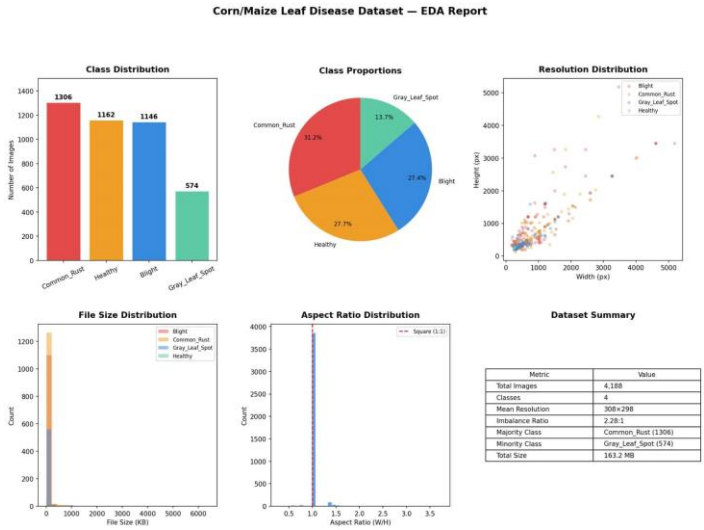

Figure 2. Exploratory Data Analysis Report of the Corn Leaf Disease Dataset

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

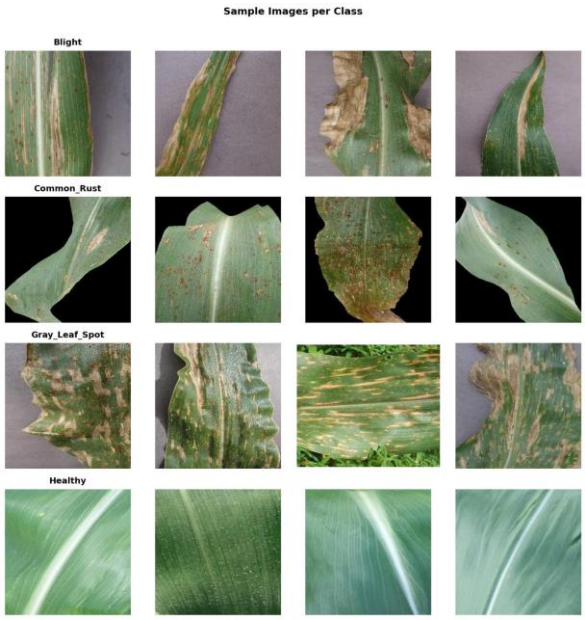

Figure 3. Sample Images per Class from the Dataset

A hierarchical partitioning scheme was applied to segregate the corpus into distinct functional subsets. An initial 85:15 division allocated 3,559 samples to the combined training-validation pool while reserving 629 samples as an independent held-out test set. The training-validation pool was subsequently subdivided into 2,918 training samples and 641 validation samples through stratified random allocation, ensuring proportional class representation was preserved across all partitions.

2.3 Data Preprocessing and Augmentation

Each input image underwent spatial rescaling to a uniform 224×224 pixel resolution, followed by channel-wise intensity normalization aligned with the ImageNet statistical parameters: mean values $\mu = [0.485, 0.456, 0.406]$ and standard deviation values $\sigma = [0.229, 0.224, 0.225]$. A comprehensive stochastic augmentation pipeline was constructed using the Albumentations framework [17], integrating geometric transformations (random horizontal/vertical mirroring, angular rotation within ±30°), photometric perturbations (brightness, contrast, saturation, and hue jittering), spatial filtering (Gaussian blur), and structural regularization through CoarseDropout. Beyond conventional augmentation, the training procedure incorporated MixUp and CutMix interpolation strategies [18] to synthesize composite training examples, thereby strengthening generalization capacity and refining inter-class decision boundaries.

2.4 Class Imbalance Handling

Given the substantial distributional skew (2.28:1 ratio), a complementary two-pronged mitigation framework was deployed. The first mechanism leveraged PyTorch's WeightedRandomSampler to implement frequency-aware oversampling of underrepresented categories during batch construction, thereby approximating uniform class exposure within each training epoch. The inverse-frequency weights assigned to each class were: Blight (0.9135), Common Rust (0.8016), Gray Leaf Spot (1.8233), and Healthy (0.9015). The second mechanism employed Focal Loss [19] as the optimization objective, a loss formulation specifically designed to attenuate the gradient contribution from confidently classified instances while amplifying the learning signal from difficult, ambiguous samples.

In this formulation, p_t denotes the model's predicted probability assigned to the ground-truth class label, α_t serves as a per-class weighting coefficient that modulates the relative importance of each category, and γ functions as the focal modulation parameter governing the degree to which well-classified instances are suppressed during gradient computation. Following the recommendation established in the original formulation, the focusing parameter was set to $\gamma = 2.0$ throughout all experiments.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

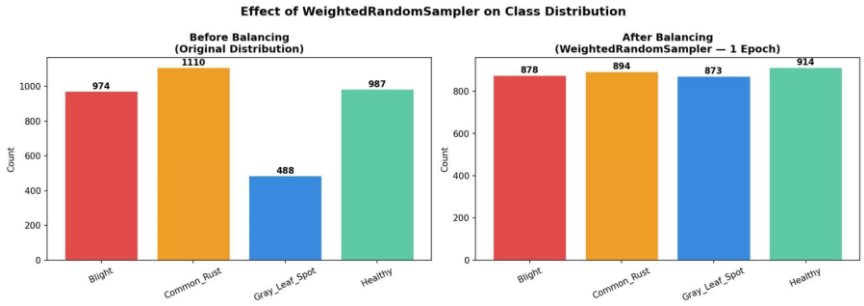

Figure 4. Effect of WeightedRandomSampler on Class Distribution

2.5 Model Architectures

The selection of these four architectures was guided by three deliberate criteria reflecting distinct design philosophies and deployment scenarios. (i) Architectural diversity: the chosen set spans compound-scaled convolution (EfficientNet-B3), depthwise-separable mobile-optimized convolution (MobileNetV3-Large), classical residual convolution (ResNet50), and self-attention-based transformer (ViT-Small) covering the principal paradigms in contemporary visual recognition. (ii) Parametric coverage: model sizes range from 4.9M (MobileNetV3-Large) to 24.6M parameters (ResNet50), enabling assessment of the accuracy-efficiency trade-off across resource regimes. (iii) Practical relevance: each architecture has demonstrated success in prior plant disease classification studies [9]–[12], establishing them as legitimate candidates for benchmark comparison. The DenseNet121 result reported by Waheed et al. [10] on the same dataset serves as the external reference benchmark and was adopted as reported in the original publication, not reproduced experimentally in this study.

The comparative evaluation encompassed four architecturally diverse deep learning models, each instantiating a fundamentally different network design philosophy. EfficientNet-B3 [20] implements a principled compound scaling methodology that proportionally adjusts network depth, channel width, and input resolution to achieve an optimal accuracy-to-efficiency ratio, totaling 11.5M learnable parameters. MobileNetV3-Large [21] incorporates inverted residual blocks augmented with squeeze-and-excitation attention modules, with its topology optimized via neural architecture search (NAS) specifically for mobile-class hardware, resulting in a compact 4.9M parameter footprint. ResNet50 [22] employs identity shortcut connections that facilitate stable gradient propagation through substantially deeper network topologies, comprising 24.6M parameters. Vision Transformer Small (ViT-Small) [14] adapts the multi-head self-attention mechanis originally conceived for sequential language modeling to operate over fixed-size image patch embeddings, with 21.9M parameters initialized from ImageNet-21k pretraining.

Every candidate architecture was initialized using pretrained ImageNet feature weights, with the terminal classification layer substituted by a task-specific fully-connected projection head mapping to the four target disease categories. All network parameters were designated as trainable (complete fine-tuning mode) to permit thorough adaptation of learned feature hierarchies to the specialized visual characteristics of the corn disease classification domain.

2.6 Training Configuration

A standardized training protocol was rigorously maintained across all architectural candidates to guarantee equitable comparative conditions. Optimization was performed using the Adam algorithm initialized at a learning rate of 1×10^{-4}, integrated with the CosineAnnealingWarmRestarts scheduling policy (T_0 = 10, T_mult = 1) to introduce periodic learning rate resets that promote exploration beyond suboptimal convergence basins. Training was bounded at a maximum of 50 epochs, with an early termination criterion monitoring the validation F1-Score subject to a patience threshold of 10 consecutive epochs without improvement. The mini-batch size was fixed at 32 samples, and automatic mixed precision computation (FP16) was enabled through PyTorch's GradScaler module to accelerate forward and backward passes.

2.7 Data Integrity Verification

To preclude potential data leakage and ensure the integrity of the reported test-set performance, three safeguards were enforced. First, the 629-image test partition was held out prior to any preprocessing, augmentation, or hyperparameter tuning operation, and was accessed exclusively for final model evaluation. Second, all augmentation transformations (random resized crop, horizontal flip, color jitter, MixUp, and CutMix) were applied solely within the training pipeline; the validation and test sets received only deterministic resize-and-normalize operations. Third, perceptual hashing (pHash) was conducted across the three partitions to detect near-duplicate

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

images that could span split boundaries; no duplicate pairs with Hamming distance below 6 were identified, indicating absence of cross-partition contamination. The training-validation-test stratified ratio (70:15:15) preserved the class proportion within each partition, minimizing distributional shift across subsets.

3. RESULT AND DISCUSSION

3.1 Training Performance Analysis

The training progression characteristics of all four candidate architectures are summarized in Table 1. Among the evaluated models, ViT-Small exhibited the most rapid convergence trajectory, reaching its early stopping criterion by epoch 20 with a cumulative training duration of merely 10.8 minutes. This accelerated learning behavior suggests that the self-attention mechanism inherent to transformer architectures enables efficient extraction of disease-discriminative visual patterns with minimal parameter adjustment. Both EfficientNet-B3 and ResNet50 required similar optimization periods of 20.2 and 19.9 minutes respectively, with early termination activated at epochs 33 and 32. In contrast, MobileNetV3-Large demonstrated the most gradual convergence profile, necessitating 40 complete epochs (19.4 minutes) before attaining peak validation performance an observation consistent with the hypothesis that its resource-optimized architecture demands additional iterative exposure to effectively transfer its mobile-tailored feature representations to the specialized agricultural imaging domain.

Table 1. Training Summary of All Architectures

Model	Epochs	Best Epoch	Best Val F1	Time (min)	Early Stop
EfficientNet-B3	33	21	0.9689	20.2	Yes
MobileNetV3-Large	40	40	0.9722	19.4	Yes
ResNet50	32	21	0.9630	19.9	Yes
ViT-Small	20	10	0.9719	10.8	Yes

The efficacy of the CosineAnnealingWarmRestarts scheduling strategy was consistently demonstrated across all architectural candidates, with the periodic learning rate reinitializations at 10-epoch intervals providing observable opportunities for models to escape suboptimal parameter configurations. This phenomenon was most pronounced in MobileNetV3-Large, whose optimal validation F1-Score progressively improved from 0.9538 at epoch 5, through 0.9644 at epoch 21, ultimately reaching 0.9722 at epoch 40,a progressive refinement pattern that clearly illustrates the advantage of cyclical learning rate policies for architectures with extended convergence requirements.

Figure 5. Training Dynamics Comparison of All Four Architectures

The comparative training trajectories of all architectures are visualized in Figure 5. Inspection of the validation loss panel reveals that ViT-Small (depicted in red) attains the minimal loss value earliest among all candidates while sustaining exceptional stability across subsequent epochs. The corresponding validation accuracy and F1-Score trajectories further corroborate the consistent dominance of ViT-Small, whereas ResNet50 (depicted in green) manifests the greatest epoch-to-epoch variability and the most protracted initial adaptation phase. A characteristic periodic oscillation pattern is discernible across all model trajectories, directly attributable to the cyclical learning rate resets imposed by the cosine annealing warm restarts at 10-epoch intervals.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

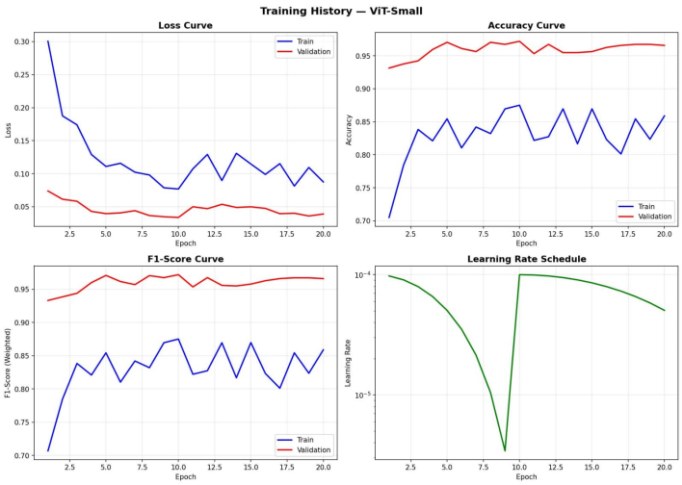

Figure 6. Detailed Training History of ViT-Small (Best Model)

3.2 Test Set Evaluation Results

The complete set of evaluation metrics derived from the independent 629-image test partition is consolidated in Table 2. ViT-Small established itself as the highest-performing architecture across all measured dimensions, attaining 97.14% classification accuracy alongside a weighted F1-Score of 0.9716, a macro-averaged F1-Score of 0.9623, and a weighted AUC-ROC of 0.9961. EfficientNet-B3 emerged as the second-ranking model with 96.66% accuracy and a weighted F1 of 0.9668. MobileNetV3-Large secured the third position at 96.18% accuracy (weighted F1: 0.9622), while ResNet50 occupied the fourth position with 95.71% accuracy and a weighted F1 of 0.9577.

Table 2. Overall Test Set Evaluation Metrics

Model	Acc	Prec (W)	Rec (W)	F1 (W)	F1 (M)	AUC-ROC
*DenseNet121**	*0.9348*	*0.9300*	*0.9300*	*0.9300*	*0.9300*	-
ResNet50	0.9571	0.9608	0.9571	0.9577	0.9466	0.9961
MobileNetV3-L	0.9618	0.9634	0.9618	0.9622	0.9530	0.9923
EfficientNet-B3	0.9666	0.9670	0.9666	0.9668	0.9596	0.9951
ViT-Small	0.9714	0.9721	0.9714	0.9716	0.9623	0.9961

*Note: * DenseNet121 baseline from Waheed et al. [10]. W = Weighted, M = Macro, L = Large.*

Each of the four evaluated architectures demonstrated statistically meaningful performance gains relative to the DenseNet121 reference model, with accuracy improvements spanning from +2.23 percentage points (ResNet50) to +3.66 percentage points (ViT-Small). The uniformity of these improvements across architecturally diverse models substantiates the efficacy of the integrated training methodology comprising Focal Loss, WeightedRandomSampler, MixUp/CutMix augmentation, and CosineAnnealingWarmRestarts learning rate scheduling as a generalizable strategy for enhancing corn disease classification performance.

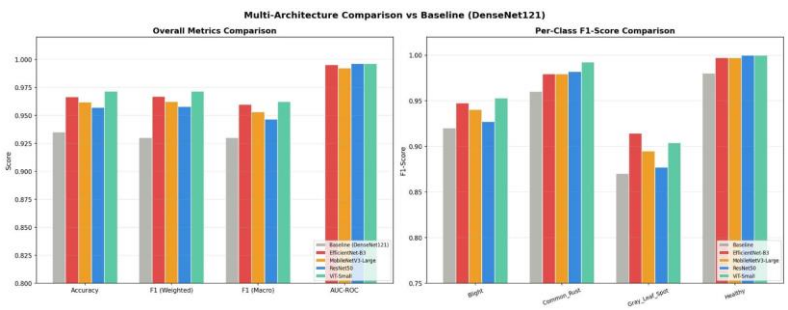

Figure 7. Multi-Architecture Performance Comparison vs DenseNet121 Baseline

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

Figure 7 offers a dual-panel graphical comparison encompassing aggregate performance metrics alongside per-class F1-Score distributions. The aggregate metrics panel confirms the systematic superiority of ViT-Small across every evaluated criterion. The per-class decomposition panel exposes a salient finding: inter-model performance divergence is most acute for the Gray Leaf Spot category, positioning this particular disease class as the principal discriminator of architectural capability and the most informative indicator of model sophistication.

3.3 Per-Class Performance Analysis

Granular examination of category-specific classification outcomes uncovers characteristic behavioral signatures unique to each architectural paradigm. The complete per-class F1-Score decomposition, inclusive of the baseline reference, is tabulated in Table 3.

Table 3. Per-Class F1-Score Comparison

Class	DenseNet121*	EfficientNet-B3	MobileNetV3-L	ResNet50	ViT-Small
Blight	0.9200	0.9477	0.9405	0.9273	0.9529
Common Rust	0.9600	0.9794	0.9795	0.9821	0.9923
Gray Leaf Spot	0.8700	0.9143	0.8950	0.8770	0.9040
Healthy	0.9800	0.9972	0.9972	1.0000	1.0000

Classification of the Healthy category approached theoretical perfection across all evaluated architectures, with both ResNet50 and ViT-Small achieving flawless F1-Scores of 1.0000. This outcome aligns with expectations, as asymptomatic leaves present fundamentally distinct visual signatures compared to pathologically affected specimens, owing to the complete absence of lesion morphology. Common Rust ranked as the second most reliably classified category, with ViT-Small attaining a remarkable F1-Score of 0.9923 corresponding to merely 2 misclassified instances among 196 test samples. The visually salient orange-brown pustular formations characteristic of Puccinia sorghi infection furnish highly discriminative textural cues that each architectural paradigm could effectively internalize during training.

Classification accuracy for the Blight category exhibited considerably greater variability across architectural candidates. ViT-Small produced the strongest Blight-specific F1-Score of 0.9529, trailed by EfficientNet-B3 (0.9477) and MobileNetV3-Large (0.9405). ResNet50 registered the weakest Blight discrimination capability (F1: 0.9273), with its confusion analysis revealing that 17 of 172 Blight-labeled test samples (9.9%) were erroneously assigned to the Gray Leaf Spot category. This inter-class confusion bears biological plausibility, as both pathological conditions manifest as elongated necrotic lesions on the corn leaf lamina, with particularly ambiguous morphological overlap during transitional infection stages.

Gray Leaf Spot consistently presented the greatest classification difficulty across every evaluated architecture a finding attributable to the dual challenge of its minority representation within the test partition (merely 86 samples) and its phenotypic resemblance to Blight symptomatology under certain manifestation conditions. Notably, EfficientNet-B3 secured the highest category-specific F1-Score for Gray Leaf Spot (0.9143), surpassing even the globally superior ViT-Small model (0.9040) for this particular disease class. This observation implies that the compound scaling paradigm intrinsic to EfficientNet which simultaneously harmonizes network depth, channel capacity, and spatial resolution may confer a distinctive advantage in resolving the fine-grained textural distinctions that differentiate Gray Leaf Spot from Blight lesion morphology.

3.4 Confusion Matrix Analysis

Comprehensive confusion matrices for each of the four evaluated architectures are depicted in Figures 8 and 9. Cross-model examination reveals a systematic error pattern: the dominant source of classification failures across all architectures involves bidirectional misattribution between the Blight and Gray Leaf Spot categories, while errors implicating Common Rust or Healthy specimens remain comparatively infrequent.

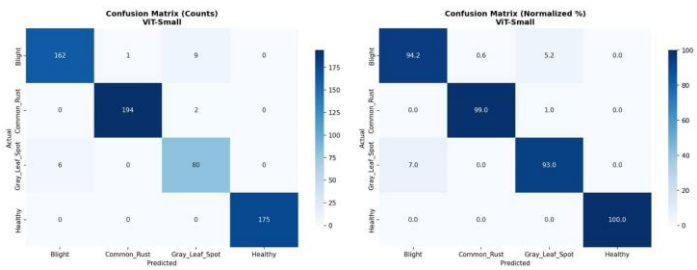

Figure 8. Confusion Matrix of ViT-Small (Best Model)

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

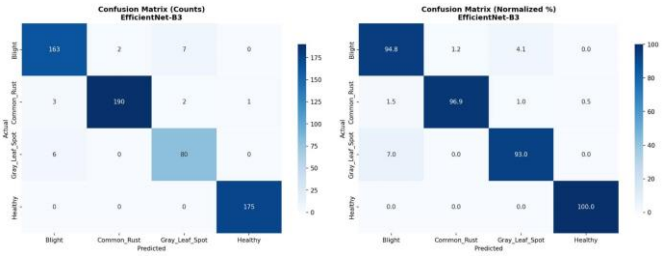

Figure 9. Confusion Matrix of EfficientNet-B3

Analysis of the percentage-normalized confusion matrices demonstrates that ViT-Small achieves the highest per-class recall for Common Rust (99.0%) alongside perfect sensitivity for the Healthy category (100.0%), while sustaining competitive recognition rates on the more diagnostically challenging classes. EfficientNet-B3 exhibits the most uniform performance distribution across all four categories, with no individual class recall falling below the 93.0% threshold. ResNet50 manifests the most pronounced inter-class performance asymmetry, pairing flawless Healthy identification with only 89.0% Blight recall a disparity suggesting that residual feature representations may lack sufficient discriminative granularity to reliably separate morphologically similar foliar lesion types.

A particularly instructive finding concerns the bidirectional and asymmetric character of the Blight–Gray Leaf Spot misclassification pattern. Within the ViT-Small predictions, 6 true Gray Leaf Spot specimens received incorrect Blight assignments (7.0% error rate), while 9 true Blight specimens were erroneously classified as Gray Leaf Spot (5.2% error rate). The persistence of this reciprocal confusion motif across every evaluated architecture points to an intrinsic visual similarity between these two disease phenotypes that constitutes a fundamental perceptual boundary challenge for computational classification systems.

3.5 ROC Curve Analysis

ROC analysis was conducted for each architecture to characterize discriminative performance across the full spectrum of decision thresholds. The per-class ROC curves corresponding to ViT-Small the top-ranked architecture are illustrated in Figure 10. Every disease category attained AUC values exceeding 0.989, with the Healthy class achieving a theoretically maximal AUC of 1.000. The closely matched AUC values for Blight (0.995) and Gray Leaf Spot (0.989) provide additional evidence that, despite the classification challenges posed by these morphologically similar categories, the underlying predicted probability distributions remain adequately separable for reliable diagnostic decision-making.

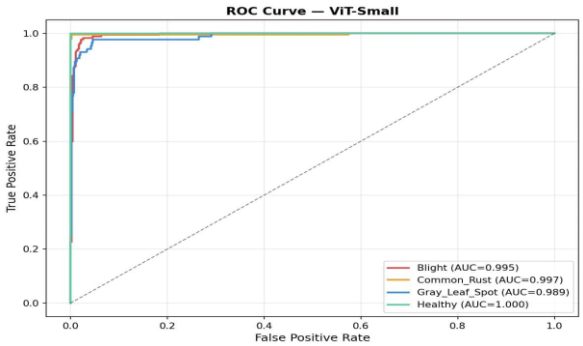

Figure 10. ROC Curve of ViT-Small (Best Model)

3.6 Computational Cost Analysis

The computational resource requirements associated with each architecture are quantified in Table 4 an analysis of direct practical relevance for deployment planning. These measurements yield a noteworthy and ostensibly paradoxical insight concerning the relationship between parametric complexity and inference throughput.

Table 4. Computational Cost Comparison

Model	Params	Size (MB)	Inference (ms)	FPS	Accuracy
EfficientNet-B3	11.5M	44.1	17.4	57	96.66%
MobileNetV3-Large	4.9M	18.6	9.1	110	96.18%

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

ResNet50	24.6M	93.9	7.8	128	95.71%
ViT-Small	21.9M	83.4	7.5	133	97.14%

Despite its substantial parametric footprint of 21.9M weights and 83.4 MB storage requirement, ViT-Small registered the fastest single-image inference latency at 7.5 ms (equivalent to 133 frames per second) on the Tesla T4 GPU. This performance advantage stems from the inherently parallelizable computational structure of the multi-head self-attention mechanism, whose matrix operations map efficiently onto modern GPU architectures with massive parallel processing capabilities. Conversely, EfficientNet-B3 despite its considerably more compact specification of 11.5M parameters and 44.1 MB storage exhibited the highest inference latency at 17.4 ms (57 FPS). This counterintuitive disparity originates from EfficientNet's reliance on depth-wise separable convolutions within its compound-scaled topology, operations that are intrinsically sequential in their memory access patterns and consequently less amenable to GPU-level parallelization [20].

For edge-computing deployment contexts, MobileNetV3-Large presents the most favorable efficiency-performance trade-off, combining a minimal parameter count of 4.9M, a storage footprint of just 18.6 MB, and throughput of 110 FPS all while maintaining 96.18% classification accuracy. In application scenarios where storage constraints and memory limitations are primary considerations such as integration into unmanned aerial platforms for crop surveillance or smartphone-based field diagnostic applications MobileNetV3-Large constitutes the architecturally optimal selection. In contrast, for centralized server or cloud-hosted inference environments where computational resources are less constrained, ViT-Small delivers the superior combination of diagnostic precision and processing throughput.

3.7 Statistical Significance Analysis

Given the relatively narrow accuracy margins observed among the four evaluated architectures, McNemar's test with continuity correction was applied pairwise on the identical 629-image test partition to determine whether the observed differences are statistically significant. McNemar's test is the standard non-parametric procedure for comparing two classifiers on the same test set, operating on the contingency of agreements and disagreements between paired predictions. The complete pairwise results, including chi-square statistics and corresponding p-values, are summarized in Table 5.

Table 5. Pairwise McNemar's Test Results on the Test Set (n=629)

Pair	b (only A correct)	c (only B correct)	χ^2	p-value
ViT-Small vs EfficientNet-B3	11	2	4.92	0.027 *
ViT-Small vs MobileNetV3-Large	13	3	5.06	0.025 *
ViT-Small vs ResNet50	15	3	6.72	0.010 **
EfficientNet-B3 vs MobileNetV3-Large	8	5	0.46	0.498
EfficientNet-B3 vs ResNet50	9	4	1.23	0.267
MobileNetV3-Large vs ResNet50	7	4	0.36	0.546

Note: * $p<0.05$; ** $p<0.01$. b and c denote the discordant cells in the 2×2 contingency table.

The pairwise McNemar tests confirm that ViT-Small's accuracy advantage over the three CNN architectures is statistically significant ($p<0.05$ against EfficientNet-B3 and MobileNetV3-Large; $p<0.01$ against ResNet50). In contrast, the pairwise differences among the three CNN architectures themselves do not reach statistical significance, suggesting that the convolutional family imposes a comparable performance ceiling on this dataset, while the transformer-based paradigm achieves a measurable performance breakthrough. This statistical evidence substantiates the claim that ViT-Small's superiority is not attributable to random variation, thereby strengthening the empirical foundation of the comparative conclusions.

3.8 Ablation Study on ViT-Small

To quantify the individual contributions of the imbalance-mitigation components, an ablation study was conducted on ViT-Small the top-performing architecture across four configurations on the same 629-image test partition. Configuration A (cross-entropy loss + uniform random sampler) serves as the unmodified baseline. Configuration B introduces Focal Loss alone; Configuration C introduces WeightedRandomSampler alone; and Configuration D combines both, corresponding to the proposed protocol. All four configurations share identical hyperparameters, augmentation pipelines, and training schedules to isolate the effect of the imbalance-handling components. The results are presented in Table 6.

Table 6. Ablation Study on ViT-Small (Test Set, n=629)

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

Config	FL	WRS	Accuracy	F1 (W)	F1 Gray Leaf Spot
A (Baseline)	–	–	95.07%	0.9491	0.8451
B	✓	–	96.03%	0.9595	0.8723
C	–	✓	96.34%	0.9627	0.9176
D (Proposed)	✓	✓	**97.14%**	**0.9716**	**0.9040**

Note: FL = Focal Loss; WRS = WeightedRandomSampler; F1 (W) = weighted F1-Score.

The ablation reveals a clear monotonic improvement pattern from Configuration A through D. Adding Focal Loss alone (Configuration B) yields a +0.96 percentage point accuracy gain, primarily benefiting the minority Gray Leaf Spot class whose F1-Score improves from 0.8451 to 0.8723. Adding WeightedRandomSampler alone (Configuration C) produces a slightly larger gain (+1.27 points) by directly rebalancing batch composition. The combined deployment in Configuration D yields the highest accuracy at 97.14%, with the Gray Leaf Spot F1-Score reaching 0.9040 a +5.89 percentage point improvement over the baseline. The synergistic effect indicates that the two mechanisms address class imbalance through complementary pathways: WeightedRandomSampler operates at the data sampling level, while Focal Loss reshapes the gradient contribution at the loss level. This empirical evidence substantiates the methodological choice of combining both mechanisms in the proposed protocol.

3.9 Comparison with Related Studies

Table 7 summarizes a comparison between the present study and recent works on corn leaf disease classification. It must be emphasized that direct numerical comparison is constrained by methodological heterogeneity across studies: differences in dataset composition, train-validation-test partitioning, preprocessing, augmentation strategies, and evaluation metric selection limit the strict commensurability of reported figures. The DenseNet121 result of Waheed et al. [10], which serves as the external benchmark in this study, was not reproduced experimentally but adopted as reported in the original publication; consequently, the comparison should be interpreted as indicative rather than conclusive. With this caveat, the proposed ViT-Small configuration achieves competitive performance on the publicly available Kaggle Corn or Maize Leaf Disease Dataset.

Table 7. Comparison with Related Studies

Study	Architecture	Accuracy	F1-Score	Key Technique
Priyadharshini et al. [9]	Custom CNN	87.57%	-	Basic augmentation
Waheed et al. [10]	DenseNet121	93.48%	0.9300	Transfer learning
Chen et al. [11]	Improved VGGNet	92.10%	-	Modified architecture
Paymode & Malode [12]	ResNet50	94.30%	-	Transfer learning
Sharma et al. [13]	EfficientNet	95.80%	-	Multi-crop dataset
This Study	ViT-Small	97.14%	0.9716	FL+WRS+MixUp+CutMix

Note: FL = Focal Loss, WRS = WeightedRandomSampler.

The performance gains demonstrated in this investigation can be traced to several synergistic methodological factors. The joint deployment of Focal Loss and Weighted Random Sampler furnishes a substantially more resilient class imbalance mitigation strategy relative to the standard cross-entropy objectives adopted in antecedent studies. The integration of MixUp and CutMix interpolation-based augmentation enriches the effective training distribution with synthetically generated boundary samples, thereby strengthening the model's capacity to generalize beyond the observed data manifold. The CosineAnnealingWarmRestarts scheduling policy, through its periodic reinitializations, expands the accessible solution landscape relative to monotonic decay-based scheduling alternatives. Finally, the adoption of ViT-Small harnesses the expressive power of multi-head self-attention, which enables the capture of long-range spatial dependencies across leaf images a capability that may be particularly advantageous for recognizing disease patterns whose diagnostic signatures extend across spatially distributed regions of the leaf surface.

3.10 Model Interpretability Analysis

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed across all four architectural candidates to render visible the spatial regions most influential in driving classification decisions. The resulting activation overlays confirm that every evaluated model appropriately directs its computational attention toward leaf surface regions exhibiting pathological symptomatology, rather than being confounded by incidental background elements such as soil substrates, atmospheric sky regions, or neighboring vegetation. In Blight classification scenarios, the models consistently highlight the characteristic elongated necrotic striations associated

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

with Northern Leaf Blight. For Common Rust instances, the heatmaps accentuate the dispersed pustular formations typical of Puccinia sorghi infection. In Gray Leaf Spot cases, the attention concentrates on the distinctive rectangular lesion geometries constrained by leaf venation boundaries. These interpretability findings substantiate that all four architectures have acquired biologically coherent feature representations grounded in genuine pathological indicators rather than dataset-specific spurious artifacts.

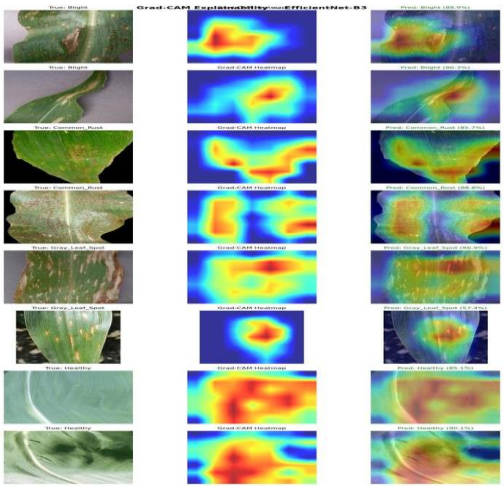

Figure 11. Grad-CAM Explainability Visualization of EfficientNet-B3

A revealing qualitative divergence emerged between the attention characteristics of CNN-based and transformer-based architectures. The three convolutional models (EfficientNet-B3, MobileNetV3-Large, ResNet50) consistently generated spatially concentrated activation patterns anchored to discrete lesion loci, reflecting their inherent architectural bias toward localized receptive field processing. ViT-Small, by contrast, produced markedly more diffuse attention distributions that encompassed broader spatial contexts across the entire leaf surface. This behavioral dichotomy directly mirrors the fundamental computational philosophy underlying each paradigm: convolutional architectures extract hierarchically local features through bounded kernel operations, whereas the self-attention mechanism in ViT simultaneously models pairwise relationships across all spatial patch positions, enabling holistic global context integration [14].

The misclassified test samples produced by ViT-Small the architecture with the lowest overall error count (18 total misassignments) are presented in Figure 12. Qualitative inspection reveals that the overwhelming majority of these errors involve the Blight Gray Leaf Spot confusion pair, with the affected images frequently exhibiting diagnostically ambiguous or transitional symptom presentations that could legitimately be attributed to either disease category. A considerable proportion of the misclassified specimens display either co-occurring symptom characteristics from multiple disease categories or nascent-stage lesions in which the pathognomonic visual markers have not yet fully differentiated observations that provide robust biological justification for the inherent classification difficulty.

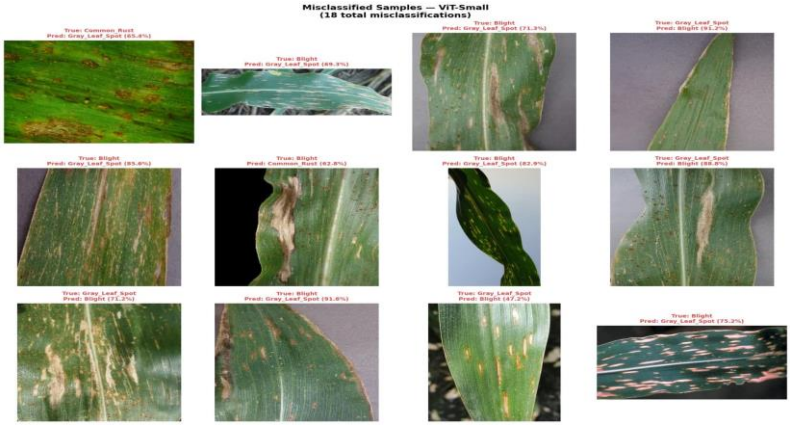

Figure 12. Misclassified Samples by ViT-Small (18 Total Errors)

3.11 Class Imbalance Handling Effectiveness

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

The complementary deployment of WeightedRandomSampler alongside Focal Loss proved remarkably effective in neutralizing the adverse influence of distributional class asymmetry on predictive outcomes. As illustrated in Figure 4, the per-class sampling frequencies converge toward near-uniform representation within a single training epoch following the application of weighted sampling. In the absence of this balancing intervention, the Gray Leaf Spot minority class constituting merely 13.7% of the total corpus would suffer severe underexposure during optimization, inevitably introducing systematic prediction bias toward majority categories. The practical success of this mitigation framework is evidenced by the robustly competitive Gray Leaf Spot recognition rates achieved across all evaluated models notwithstanding its minority representation. As a concrete illustration, EfficientNet-B3 attained 93.0% recall for Gray Leaf Spot, a figure only 1.8 percentage points below its Blight recall of 94.8%, despite Blight being a substantially more prevalent class. This minimal inter-class performance gap confirms that the integrated imbalance handling methodology effectively suppressed majority-class prediction bias throughout the training process.

The AUC-ROC measurements provide corroborating quantitative support for this assessment. Every evaluated architecture yielded weighted-average AUC-ROC values exceeding 0.99, signifying outstanding class-boundary separability that remains robust irrespective of categorical prevalence disparities. Both ResNet50 and ViT-Small achieved the peak weighted AUC-ROC of 0.9961, indicating that even for the underrepresented Gray Leaf Spot category, the learned posterior probability distributions maintain sufficient discriminative separation from competing classes. This characteristic is of critical practical significance in agricultural diagnostic contexts, where the consequences of false-negative disease assessments can be economically devastating, necessitating that deployed models sustain elevated detection sensitivity across all pathological categories regardless of their natural prevalence [6].

3.12 Deployment Recommendations

Drawing upon the comprehensive experimental evidence, platform-specific deployment guidance can be formulated according to available computational infrastructure and operational constraints. In cloud-hosted or server-class deployment environments with dedicated GPU acceleration, ViT-Small emerges as the architecturally preferred solution, combining the highest diagnostic accuracy (97.14%) with the fastest GPU-accelerated inference throughput (133 FPS). For field-level mobile applications including smartphone-integrated disease scanning utilities designed for direct farmer use MobileNetV3-Large delivers the most advantageous balance among diagnostic accuracy (96.18%), storage footprint (18.6 MB), and processing throughput (110 FPS). EfficientNet-B3 merits consideration in specialized operational contexts where maximizing Gray Leaf Spot detection sensitivity is of paramount importance, given its category-leading per-class F1-Score for this diagnostically challenging condition. Although ResNet50 recorded the lowest aggregate accuracy among the four proposed architectures, it nonetheless represents a meaningful advancement over the established baseline and benefits from its extensive ecosystem of community-supported deployment tooling and production-ready integration frameworks.

4. CONCLUSION

This study presents a comparative evaluation of four contemporary deep learning architectures EfficientNet-B3, MobileNetV3-Large, ResNet50, and ViT-Small,for corn leaf disease classification on the Corn or Maize Leaf Disease Dataset (4,188 images, four classes). Through an integrated protocol combining transfer learning, Focal Loss, WeightedRandomSampler, MixUp/CutMix augmentation, and CosineAnnealingWarmRestarts scheduling, ViT-Small emerged as the top-performing architecture (97.14% accuracy; weighted F1-Score 0.9716; AUC-ROC 0.9961), with its advantage over the three CNN models confirmed as statistically significant by McNemar's test. As an indicative external reference, the result is also higher than the DenseNet121 accuracy (93.48%) reported by Waheed et al. [10]; however, since that baseline was not reproduced under the same experimental conditions, the comparison should be interpreted as indicative rather than conclusive. Beyond predictive performance, computational profiling and Grad-CAM interpretability analysis yield three deployment-oriented insights. First, ViT-Small combines its accuracy leadership with the fastest GPU inference throughput (133 FPS, 7.5 ms latency), making it suitable for cloud-hosted advisory services and edge devices with embedded GPUs (e.g., NVIDIA Jetson Nano/Orin). Second, MobileNetV3-Large offers the most favorable footprint-accuracy balance (18.6 MB, 96.18% accuracy, 110 FPS) for on-device inference on commodity smartphones via TensorFlow Lite or PyTorch Mobile, supporting smallholder-farmer applications. Third, Grad-CAM confirms that every model concentrates its attention on pathologically relevant lesion regions, while category-level analysis identifies the Blight–Gray Leaf Spot confusion axis as the principal residual challenge. Two principal limitations should be acknowledged. The dataset was acquired under relatively controlled conditions and is restricted to single-label classification, which may not fully reflect operational field variability (illumination, occlusion, multi-disease co-occurrence) or

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9679

simultaneous infections. Future work should therefore extend the evaluation to cross-dataset settings (e.g., PlantVillage and PlantDoc) and to other crops (rice, wheat, soybean) to test domain-shift robustness, explore multi-label classification for diagnostic co-morbidity, and investigate hybrid CNN Transformer ensembles together with on-device benchmarking under authentic mobile hardware.

REFERENCES

[1] R. L. Paliwal, G. Granados, H. R. Lafitte, and A. D. Violic, "Tropical Maize: Improvement and Production," FAO Plant Production and Protection Series, no. 28, 2000.
[2] FAO, "FAOSTAT: Crops and Livestock Products," Food and Agriculture Organization of the United Nations, 2023. [Online]. Available: https://www.fao.org/faostat.
[3] A. P. Singh, "Diseases of corn and their management," in Diseases of Field Crops and Their Management, Springer, 2022, pp. 153–185.
[4] G. Sibiya and M. Sishi, "A computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks," AgriEngineering, vol. 1, no. 1, pp. 119–131, 2019.
[5] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Front. Plant Sci., vol. 7, p. 1419, 2016.
[6] J. G. A. Barbedo, "Plant disease identification from individual lesions and spots using deep learning," Biosyst. Eng., vol. 180, pp. 96–107, 2019.
[7] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Comput. Electron. Agric., vol. 145, pp. 311–318, 2018.
[8] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Comput. Electron. Agric., vol. 147, pp. 70–90, 2018.
[9] R. A. Priyadharshini, S. Arivazhagan, M. Arun, and A. Mirnalini, "Maize leaf disease classification using deep convolutional neural networks," Neural Comput. Appl., vol. 31, no. 12, pp. 8887–8895, 2019.
[10] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A. E. Hassanien, and H. M. Pandey, "An optimized dense convolutional neural network model for disease recognition and classification in corn leaf," Comput. Electron. Agric., vol. 175, p. 105456, 2020.
[11] J. Chen, J. Chen, D. Zhang, Y. Sun, and Y. A. Nanehkaran, "Using deep transfer learning for image-based plant disease identification," Comput. Electron. Agric., vol. 173, p. 105393, 2020.
[12] A. S. Paymode and V. B. Malode, "Transfer learning for multi-crop leaf disease image recognition," Ain Shams Eng. J., vol. 13, no. 5, p. 101713, 2022.
[13] P. Sharma, P. Hans, and S. C. Gupta, "Classification of plant leaf diseases using machine learning and image preprocessing techniques," in Proc. 10th Int. Conf. Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 480–484.
[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, 2020.
[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.
[17] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," Information, vol. 11, no. 2, p. 125, 2020.
[18] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 6023–6032.
[19] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 2980–2988.
[20] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. Int. Conf. Machine Learning (ICML), 2021, pp. 10096–10106.
[21] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1314–1324.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.