Journal of Computer Science

Research Article

Benchmarking Lightweight Deep Learning Models for Field Based Sorghum Leaf Disease Detection

Prince Daughin Ngqabutho Ncube

Department of Networking and IT Support, Walter Sisulu University, East London, South Africa

Article history
Received: 09-01-2026
Revised: 01-05-2026
Accepted: 29-06-2026

Email: pncube@wsu.ac.za

Abstract: Sorghum is a climate-resilient cereal crop central to food and livelihood security across Southern Africa, yet its productivity is frequently reduced by foliar diseases that smallholder farmers struggle to diagnose reliably in the field. While deep learning has demonstrated strong potential for automated plant disease identification, sorghum remains underrepresented in the literature, and no prior study has provided a comprehensive, deployment-oriented benchmark of modern architectures across its major leaf diseases. This study evaluates three widely used convolutional neural networks, MobileNetV3-Large, EfficientNet-B0, and ResNet-50, using the publicly available Mendeley Sorghum Disease Image Dataset, consisting of 7,167 field-derived RGB images across six disease classes, under a unified and standardized training pipeline. MobileNetV3-Large and ResNet-50 achieved the highest classification accuracy (99.54%), followed by EfficientNet-B0 (98.79%), while MobileNetV3-Large provided the most favorable balance between accuracy, model size, and inference latency. To assess robustness and generalization, additional experiments were conducted, including multi-seed stability, data augmentation sensitivity, robustness to image perturbations, and background ablation. Results show that performance is stable across training runs but sensitive to image degradation, particularly blur, and to the removal of contextual information, which led to a marked decline in accuracy. This indicates that models rely not only on localized disease features but also on broader contextual cues, raising important considerations for robustness and generalization in real-world deployment. This study establishes a reproducible, efficiency-aware benchmarking framework for sorghum disease classification and provides new insight into robustness, context dependence, and deployment constraints in field-based plant disease detection, supporting the development of lightweight diagnostic tools for resource-constrained environments.

Keywords: Deep Learning, Digital Agriculture, Sorghum Disease Detection, Lightweight Convolutional Neural Networks, Transfer Learning

Introduction

Sorghum (Sorghum bicolor) is a drought-tolerant cereal crop central to food and livelihood security across Southern Africa (Dunjana et al., 2022). Its resilience to high temperatures, low rainfall, and poor soil fertility makes it a key component of climate-adaptation strategies in arid and semi-arid regions (Tolosa et al., 2023; Mwamahonje et al., 2024). However, productivity is frequently constrained by foliar diseases such as anthracnose, smut, and rust, which reduce grain quality and yield and are often visually difficult to distinguish in the field (Ahn et al., 2023). Disease pressure is projected to intensify under climate change as pathogen distributions and epidemiological patterns shift (Bebber, 2015; Savary et al., 2019).

Timely and accurate disease diagnosis is therefore critical, yet smallholder farmers, who produce much of the region's sorghum, often lack access to plant health specialists, diagnostic services, and rapid extension support (Sithole and Olorunfemi, 2024). Although many farmers draw on Indigenous Knowledge (IK) systems inherited across generations, including experiential observation of plant symptoms, seasonal patterns, and locally adapted management practices, these approaches are increasingly constrained by the accelerating impacts of climate change. Emerging sorghum diseases, shifts in pathogen behavior, and atypical symptom expression reduce the reliability of traditional diagnostic cues, leading to delayed or incorrect diagnoses and, ultimately, suboptimal management decisions.

Deep learning has emerged as a powerful tool for automated plant disease detection, consistently outperforming traditional machine-learning methods across multiple crops and imaging conditions (Too et al., 2019; Chen et al., 2020). Transfer learning and data augmentation enhance performance on small or imbalanced datasets (Shorten and Khoshgoftaar, 2019), while lightweight architectures, such as MobileNet and EfficientNet, enable real-time inference on mobile and low-power devices (Ramcharan et al., 2017; Kumar et al., 2025a). However, the effectiveness of specific augmentation strategies and the robustness of lightweight models under field-level image variability, including lighting changes and blur, remain insufficiently examined in sorghum disease classification.

Despite this progress, these challenges are particularly pronounced in sorghum, which remains underrepresented in AI-for-agriculture research. Existing work focuses mainly on phenotyping tasks, including panicle detection, stomatal trait analysis, and leaf counting, or examines individual diseases using small, custom datasets with limited geographic coverage (Senthil et al., 2024; Gonzalez et al., 2024; Varur et al., 2024; Desai et al., 2025). Recent reviews emphasize that systematic sorghum-disease benchmarks are notably absent, especially for models designed for mobile or field deployment (Kutyauripo et al., 2025). Notably, there is a lack of peer-reviewed studies comparing modern transfer learning architectures on a publicly available dataset encompassing all six major sorghum foliar diseases.

This study addresses this gap by establishing a standardized and reproducible benchmark for sorghum leaf disease classification using modern lightweight and conventional transfer learning architectures. Specifically, the study:

i. Compares multiple models under an identical and rigorously controlled training pipeline
ii. Jointly evaluates predictive performance and deployment-relevant efficiency metrics, including model size, inference latency, and throughput
iii. Assesses model reliability through additional analyses, including multi-seed stability, augmentation sensitivity, robustness under image perturbations, and background ablation. These contributions collectively establish a comprehensive and deployment-oriented evaluation framework for developing field-ready diagnostic systems in smallholder farming contexts

To guide this investigation, the study addresses the following research question.

How do MobileNetV3-Large, EfficientNet-B0, and ResNet-50 compare in predictive performance and computational efficiency for sorghum leaf disease classification?

Models are evaluated under a consistent experimental setup, with performance assessed using accuracy, precision, recall, F1-score, model size, and CPU-based inference metrics, including latency and throughput, to evaluate their suitability for deployment in resource-constrained agricultural environments.

Related Work

Recent advances in plant disease detection research underscore the growing importance of timely and accurate diagnosis in safeguarding agricultural productivity and food security. Traditional approaches, including Indigenous Knowledge (IK), visual assessment by farmers, and expert field scouting, remain valuable but face increasing challenges such as labor intensity, symptom subjectivity, limited specialist availability, and reduced reliability under shifting climate conditions. These constraints have motivated the integration of digital technologies not to replace traditional knowledge systems, but to complement and strengthen them by providing scalable, objective, and rapid diagnostic support. Against this background, recent studies have explored a range of computational approaches for plant disease recognition, with deep learning emerging as the most prominent and effective direction.

Deep Learning and Transfer Learning for Plant Disease Detection

Over the past decade, deep learning has transformed image-based plant disease detection. Early work by Mohanty et al. (2016); Ferentinos (2018) showed that Convolutional Neural Networks (CNNs) can accurately classify leaf diseases across multiple crop species and imaging conditions. Comparative studies and surveys in (Too et al., 2019; Chen et al., 2020; Khalid and Romle, 2023) consistently report that deep learning methods outperform traditional machine learning approaches when sufficient labeled images and appropriate augmentation strategies are employed.

Transfer learning has been particularly important in agricultural domains where labeled datasets are comparatively small. Fine-tuning pre-trained architectures, such as VGG, ResNet, and Inception, on plant disease datasets yields substantial gains over training from scratch (Too et al., 2019; Shah et al., 2023). More recent studies have focused on architectures designed for efficiency, including

MobileNet and EfficientNet, which enable real-time or near real-time inference on edge or mobile devices (Ramcharan et al., 2017; Kumar et al., 2025a). These works emphasize the need to balance accuracy with computational cost to support deployment in low-resource settings.

Sorghum Specific Deep Learning Applications

In contrast to major cereals such as maize, rice, and wheat, sorghum has received comparatively limited attention in computer vision and deep learning research. Comparative studies also note that deep-learning applications for sorghum remain fewer and less mature than those developed for major global staples (e.g., rice, wheat, maize) (Li et al., 2022; Waldamichael et al., 2022). Existing sorghum-focused studies predominantly address phenotyping and yield estimation, including sorghum head and panicle detection (Ghosal et al., 2019; Lin and Guo, 2020; Santiago et al., 2024), stomatal trait phenotyping (Bheemanahalli et al., 2021), and automated leaf counting (Miao et al., 2021). Additional work includes weed segmentation in sorghum fields using UAV imagery (Genze et al., 2022) and pest-related tasks such as sorghum aphid detection (Grijalva et al., 2023).

A smaller but growing literature addresses sorghum diseases directly. Gonzalez et al. (2024) used deep neural networks to quantify charcoal rot leaf symptoms in field-grown plants. Senthil et al. (2024) applied CNN models for leaf blight detection, while a study in Varur et al. (2024) focused on tar disease severity estimation. Desai et al. (2025) investigated the severity detection of fall armyworm infestation on sorghum leaves using deep learning. Although these studies demonstrate the feasibility of deep learning for sorghum pathology, they tend to rely on custom datasets with limited geographic coverage and often focus on a single disease or a narrow group of symptoms.

Gaps in Sorghum Disease Benchmarks

Existing sorghum disease studies do not cover the full set of six major foliar diseases, Anthracnose and Red Rot, Cereal Grain Mold, Covered Kernel Smut, Head Smut, Loose Smut, and Rust, represented in the Mendeley Sorghum Disease Image Dataset (Sandip and Kailas, 2024). Most work, such as Senthil et al. (2024), considers only a small number of disease classes; Gonzalez et al. (2024) relies on a custom field-image dataset collected in a specific region, without providing a reusable open benchmark; and Ramcharan et al. (2017) evaluates a single deep CNN architecture without any systematic analysis of computational efficiency. Recent analyses of AI applications in sorghum and millet farming emphasize that sorghum remains significantly underrepresented in digital agriculture research and that disease diagnostics represent a key area for further development (Kutyauripo et al., 2025).

Comparative benchmarks on related crops, such as Shah et al. (2023); Too et al. (2019), have demonstrated the value of evaluating multiple architectures within a unified pipeline to understand the trade-offs between accuracy, model size, and inference speed. Computational efficiency, including model size and inference latency, is increasingly recognized as a critical factor for practical deployment on mobile phones and low-power edge devices (Ramcharan et al., 2017; Kumar et al., 2025a). This emphasis aligns with broader advances in AI for resource-constrained and IoT environments, where lightweight and hybrid deep learning models are increasingly used to support real-time decision-making under constrained computational and energy budgets (Kumar et al., 2025b). However, systematic comparisons of this nature are largely absent from sorghum disease detection research.

Sorghum is a climate-resilient small grain of growing importance in Southern Africa, yet it remains relatively unexplored in digital agriculture research despite significant disease pressures intensified by changing climate patterns. This study contributes to addressing these gaps by benchmarking three widely used transfer learning architectures on a public sorghum disease dataset under controlled and consistent experimental conditions, and by jointly analyzing predictive and computational performance.

Methods

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), which structures data-driven research into six iterative stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Chapman et al., 2000). The framework is adopted to ensure transparency, reproducibility, and methodological rigor in benchmarking deep learning models for sorghum disease classification.

Figure 1 presents the experimental workflow as an operational sequence. The blocks shown in the figure correspond directly to specific CRISP-DM stages: Dataset loading and inspection align with Data Understanding; image preprocessing and data augmentation represent Data Preparation; model training and fine-tuning constitute the Modeling stage; and evaluation and benchmarking reflect the Evaluation stage.

The Business Understanding stage is addressed in the Business Understanding subsection, where the problem formulation and study motivation are defined. The Deployment stage is reflected in the Deployment Considerations subsection, which examines computational efficiency and real-world feasibility. In this way, all stages of the CRISP-DM framework are incorporated, while Figure 1 emphasizes the operational pipeline implemented in this study.

Business Understanding

Under the Business Understanding stage of the CRISP-DM framework, the task was defined as a supervised image classification problem aimed at distinguishing six major sorghum foliar diseases from field-acquired RGB images. The primary objective was to benchmark representative transfer learning architectures with respect to both predictive performance and deployment-relevant efficiency, reflecting the constraints of mobile and low-resource agricultural settings.

To support fair and reproducible comparison, the study prioritized the use of a publicly available dataset, a unified training and evaluation pipeline, and consistent performance metrics across all models. Particular attention was given to computational efficiency, including model size and inference speed, as these factors directly influence feasibility for real-world diagnostic tools. These objectives guided the selection of models, preprocessing strategies, and evaluation criteria implemented across subsequent CRISP-DM stages.

Data Understanding

The experiments use the publicly available Mendeley Sorghum Disease Image Dataset (Sandip and Kailas, 2024), which contains 7167 RGB images across six foliar disease categories: Anthracnose and Red Rot, Cereal Grain Mold, Covered Kernel Smut, Head Smut, Loose Smut, and Rust. The images were collected under natural field and field-like conditions and exhibit variation in illumination, background clutter, leaf orientation, and symptom severity. This variability provides a realistic basis for evaluating the robustness of deep learning models for agricultural use. The dataset is considerably imbalanced, particularly between Rust and Covered Kernel Smut, which has the potential to bias model learning. These characteristics informed the decisions taken in the Data Preparation and Modeling stages.

Data Preparation

All images were manually inspected to remove corrupted files and to ensure a consistent directory structure. No images were excluded based on disease severity, symptom visibility, or image clarity. The dataset was then partitioned into training (70%), validation (15%), and test (15%) subsets, following common practice in image-based classification studies (Too et al., 2019). Stratified sampling was applied to maintain proportional class representation, consistent with recommendations for handling imbalanced datasets (Krawczyk, 2016).

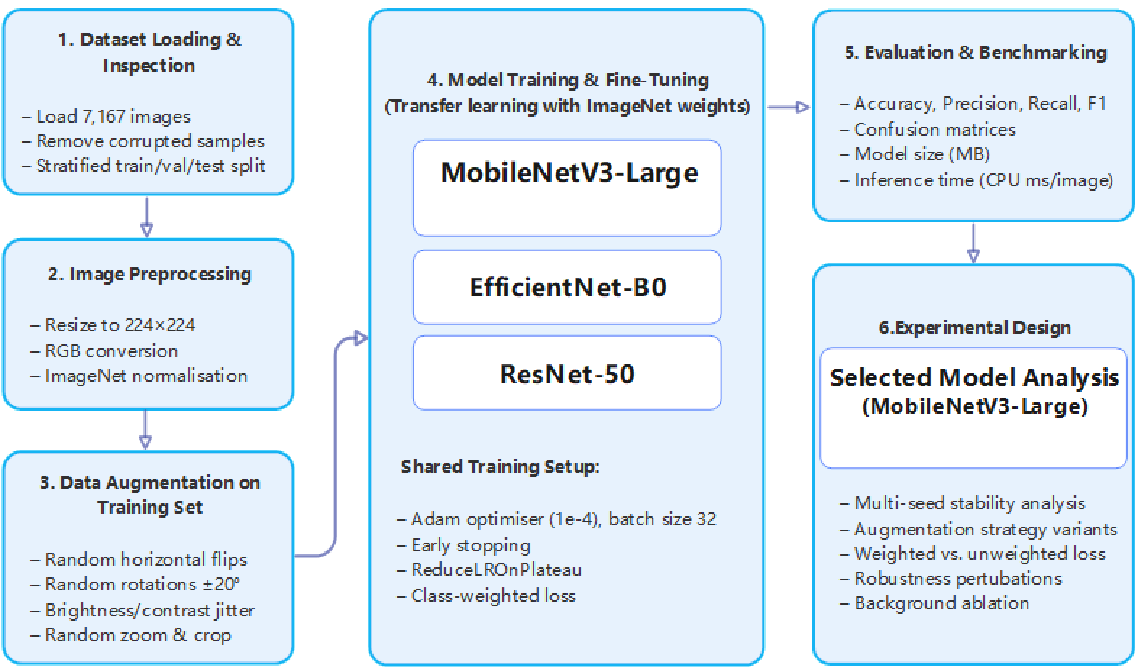

Fig. 1: Experimental pipeline illustrating the CRISP-DM workflow adopted in this study. The process begins with dataset loading and inspection, followed by image preprocessing and data augmentation applied to the training set. Three transfer learning architectures, MobileNetV3-Large, EfficientNet-B0, and ResNet-50, are then trained under a shared configuration using ImageNet-pretrained weights. Baseline performance is evaluated and benchmarked using classification metrics and computational efficiency indicators. Based on this comparison, the selected model is subjected to further analysis through controlled experiments, including stability assessment, training sensitivity analysis, robustness evaluation, and background ablation

Each image was resized to 224 by 224 pixels, converted to RGB format, and normalized using the ImageNet mean and standard deviation to maintain compatibility with ImageNet-pretrained models. To enhance generalization and reflect the variability encountered in real field conditions, the training set was augmented using transformations that simulate natural variation in leaf orientation, illumination, and framing. These transformations included horizontal flips to mimic wind-induced movement, rotational adjustments to account for non-aligned field captures, controlled brightness and contrast changes to replicate fluctuating sunlight, and random zooming with cropping to approximate differences in camera distance. Such augmentations follow established recommendations for deep learning on agricultural imagery and are intended to improve robustness to field-level variability (Shorten and Khoshgoftaar, 2019). Validation and test images were processed only through resizing and normalization to ensure consistent and unbiased evaluation. Given the pronounced imbalance across disease categories in the dataset, class-weighted cross-entropy loss was incorporated during training to reduce bias toward majority classes while preserving the natural class distribution.

The final distribution of images across the training, validation, and test subsets after stratified sampling is presented in Table 1. Table 1 illustrates the class imbalance present in the dataset, which further motivated the use of class-weighted loss.

Modeling

Three transfer learning architectures were selected to represent distinct design strategies within contemporary deep learning. MobileNetV3-Large (Howard et al., 2019) is designed for efficiency on mobile and embedded hardware through depthwise separable convolutions and lightweight attention mechanisms. EfficientNet-B0 (Tan and Le, 2019) employs compound scaling to balance network depth, width, and resolution and is known for its favorable accuracy-to-efficiency trade-off. ResNet-50 (He et al., 2016) provides a strong conventional baseline with residual connections that address vanishing gradients in deeper models. All models were initialized with ImageNet-pretrained weights and adapted to the six-class sorghum disease classification task by replacing the final classification layer with a fully connected layer that produces six outputs. The models were fine-tuned end-to-end, allowing the feature extractors to adapt to the characteristics of the sorghum imagery.

All models were trained with identical hyperparameters to ensure fair comparison. Training used the Adam optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-4}) with a batch size of 32, following the optimization method introduced by (Kingma and Ba, 2017). Models were trained for up to 25 epochs, with early stopping activated after seven epochs without a decrease in validation loss. A ReduceLROnPlateau scheduler was applied to reduce the learning rate when validation performance plateaued for three consecutive epochs.

Because the dataset exhibits considerable class imbalance, for example, Rust contains substantially more samples than Covered Kernel Smut, balanced class weights from scikit-learn were incorporated into the cross-entropy loss to penalize misclassification of minority classes. This approach aims to reduce the bias towards majority classes without altering the underlying dataset distribution.

Following the baseline benchmarking of architectures, the model demonstrating the most favorable balance between predictive performance and computational efficiency was selected for further analysis. To ensure that performance observations were not solely dependent on a single training configuration or dataset representation, additional experiments were conducted to evaluate model behavior under multiple controlled conditions. These experiments examine the stability of training outcomes, sensitivity to data augmentation and class weighting, robustness to input perturbations, and the influence of contextual background information on classification performance.

All experiments were conducted in Google Colab, a cloud-based development environment that provides managed GPU resources (Bisong, 2019). Model training was executed on NVIDIA GPUs, typically Tesla T4 accelerators designed for deep-learning inference and training workloads (NVIDIA Corporation, 2018). The software stack consisted of Python 3.x (Rossum and Drake, 2009), PyTorch 2.x for model implementation and optimization (Paszke et al., 2019), and the CUDA libraries supplied through the Colab runtime for GPU acceleration (NVIDIA Corporation, 2025).

Table 1: Distribution of sorghum disease classes across the training, validation, and test sets after stratified sampling

Disease Class	Train	Val	Test	Total
Anthracnose and Red Rot	709	152	152	1,013
Cereal Grain Molds	854	183	183	1,220
Covered Kernel Smut	205	43	44	279
Head Smut	349	75	75	499
Loose Smut	1,244	267	266	1,777
Rust	1,665	357	357	2,379
Total	5,026	1,077	1,077	7,167

DOI: 10.3844/jcssp.2026.2992.3006

Experimental Design

To provide a comprehensive evaluation of model performance, a structured experimental design was adopted to complement the baseline benchmarking and to assess model behavior under multiple controlled conditions. This design examines the consistency, sensitivity, robustness, and input dependence of the selected architecture while maintaining the same underlying dataset and training framework.

Stability Analysis

To assess the consistency of model performance, multiple training runs were conducted using different random seeds. This approach enables the evaluation of variability in performance metrics and ensures that observed results are not dependent on a specific initialization or stochastic training process. Performance statistics across runs were used to quantify the stability of the model under identical training conditions.

Training Sensitivity Analysis

The sensitivity of the model to training configurations was investigated through two complementary experiments. First, the impact of data augmentation strategies was examined to determine their contribution to generalization performance. Variations in augmentation pipelines were evaluated to determine how effectively these transformations capture field-level variability in sorghum imagery.

Second, the effect of class-weighted loss was analyzed by comparing models trained with and without class weighting. This experiment focused on differences in performance across majority and minority classes, with particular attention to underrepresented disease categories.

Robustness Evaluation

To evaluate model performance under conditions representative of real-world image acquisition, controlled perturbations were applied to input images. These perturbations included variations in brightness, contrast, and image quality degradation, such as image blurring. The objective of this analysis was to assess the model's resilience to common distortions encountered in field-based scenarios and to determine the extent to which these variations influence classification performance.

Background Ablation

To investigate the influence of contextual background information on model predictions, a segmentation-based background ablation approach was employed. Foreground leaf regions were isolated using the GrabCut algorithm (Rother et al., 2004), which iteratively separates foreground and background based on color and spatial continuity.

The segmentation process was performed offline as a preprocessing step to generate foreground-only images. The selected model was subsequently retrained using these segmented inputs under the same training configuration as the baseline. This enables a direct comparison between models trained on original images and those trained on background-removed data, thereby isolating the contribution of foreground disease features to classification performance.

Evaluation

All three architectures were assessed using a unified experimental protocol to ensure that any differences in performance reflected intrinsic model characteristics rather than variations in preprocessing, augmentation, or training configurations. Evaluation was carried out on the held-out test set and included overall accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. Confusion matrices were generated to examine performance patterns and to identify persistent misclassification trends.

In addition to overall performance metrics, class-level behavior was analyzed to provide deeper insight into model decision behavior, particularly in the presence of class imbalance and visually similar disease categories. To support the experimental design, comparative evaluation was conducted across different training conditions and input variations, enabling assessment of model stability, sensitivity, and robustness under controlled scenarios.

Computational efficiency was assessed using multiple indicators to support the evaluation of real-world feasibility. Model size in megabytes was measured to reflect storage requirements, and inference latency per image was measured on a CPU within the Google Colab environment using a virtualized x86_64 processor (2 cores) with approximately 12.7 GB of RAM. Throughput, expressed as images processed per second, was computed as the inverse of inference latency to provide an additional measure of processing capacity. Together, these measurements were used to estimate the relative computational demands of each architecture. Although absolute values may vary across hardware platforms, this setup provides a consistent comparative basis for evaluating deployment suitability.

Deployment Considerations

Deployment considerations address the practical viability of the evaluated architectures within low-resource agricultural contexts. Devices used by smallholder farmers often have limited memory, processing capacity, and network connectivity. These constraints require models that offer a balance between predictive accuracy and computational efficiency. To support this assessment, model size and inference time were used as empirical indicators of deployment feasibility across architectures. Lightweight architectures, such as MobileNetV3-Large and EfficientNet-B0, are better suited for mobile or offline applications, while ResNet-50 offers a stronger accuracy baseline that may be more suitable for deployment in higher-performance environments. Although full deployment is beyond the scope of this study, this stage of the CRISP-DM framework evaluates the readiness of each architecture for future integration into practical diagnostic tools.

Results

Baseline Performance and Benchmarking

This section presents the experimental findings, beginning with training behavior and followed by test-set performance across the three architectures. Figures 2 and 3 illustrate training and validation dynamics, while Figures 4a-4c present the corresponding confusion matrices.

Training Performance

All models trained stably, with no evidence of divergence or optimization collapse. As shown in Figure 2, MobileNetV3-Large and EfficientNet-B0 exhibited rapid convergence in training and validation loss, reaching low and stable values within the first five to seven epochs. ResNet-50 also converged reliably, although with slightly greater fluctuation in validation loss, consistent with its larger depth and parameter count.

Early stopping occurred at different points for each model. ResNet-50 reached its optimal validation loss by epoch 14, while MobileNetV3-Large achieved its best validation performance at epoch 21. EfficientNet-B0 demonstrated highly stable convergence behavior, reaching its optimal validation performance earlier at epoch 8 and maintaining consistently low validation loss thereafter.

The corresponding training and validation accuracy trajectories are presented in Figure 3. All three architectures achieved high validation accuracy early in training, exceeding 95% within the first few epochs and approaching saturation thereafter. The close alignment between training and validation accuracy curves indicates effective optimization with limited evidence of overfitting across all models.

Test Set Performance

Overall Accuracy

The classification performance of the three architectures on the held-out test set is summarized in Table 2. All models achieved high accuracy, with MobileNetV3-Large and ResNet-50 attaining the highest performance at 99.54%, followed by EfficientNet-B0 at 98.79%.

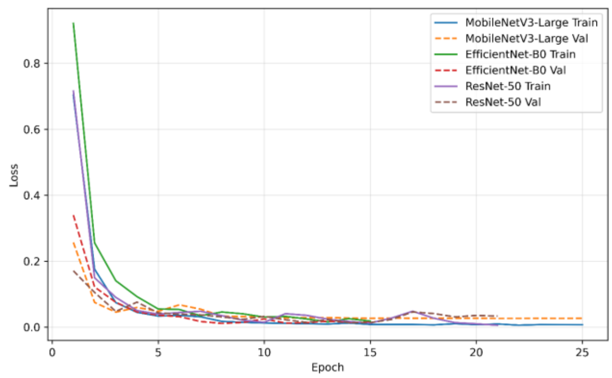

Fig. 2: Training and validation loss curves for MobileNetV3-Large, EfficientNet-B0, and ResNet-50

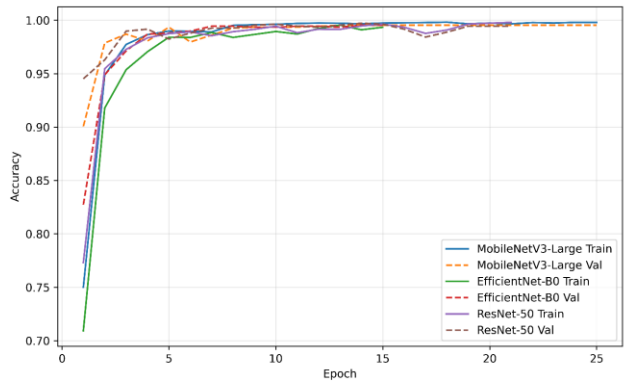

Fig. 3: Training and validation accuracy curves for MobileNetV3-Large, EfficientNet-B0, and ResNet-50

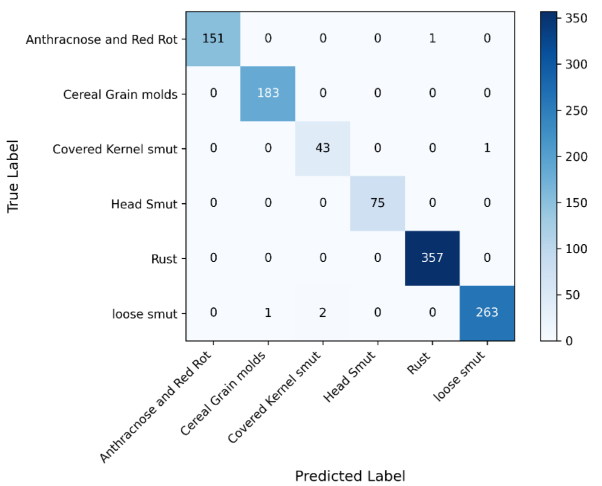

Fig. 4a: Confusion matrix for MobileNetV3-Large

DOI: 10.3844/jcssp.2026.2992.3006

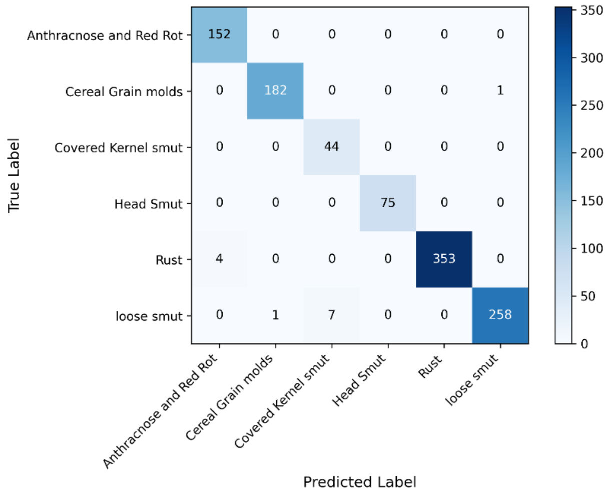

Fig. 4b: Confusion matrix for EfficientNet-B0

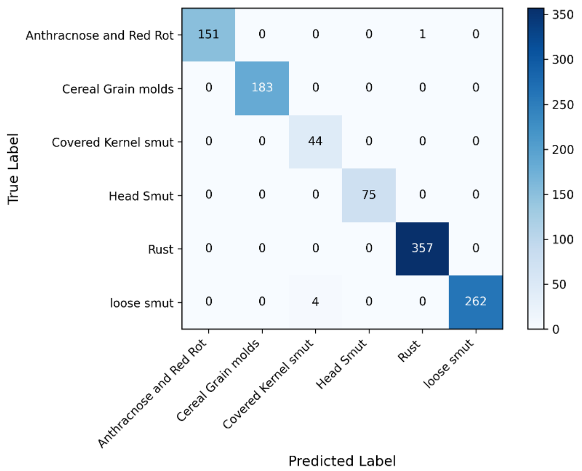

Fig. 4c: Confusion matrix for ResNet-50

In terms of macro-averaged performance, MobileNetV3-Large achieved the highest F1-score (0.9919), followed closely by ResNet-50 (0.9907), while EfficientNet-B0 achieved 0.9808. These results indicate that all three architectures are capable of effectively learning discriminative features for sorghum disease classification, with only marginal differences in predictive performance across models.

Class Wise Performance

Performance across individual disease classes remained consistently high for all three architectures, with class-wise precision, recall, and F1-scores generally exceeding 0.98. Cereal Grain Molds, Head Smut, and Rust were among the most easily discriminated classes, with all models achieving near-perfect classification performance.

Greater variability was observed among visually similar smut-related diseases, particularly Loose Smut and Covered Kernel Smut. While performance remained strong for these classes, slight reductions in precision and recall were observed across all architectures. MobileNetV3-Large and ResNet-50 exhibited more balanced precision-recall trade-offs across these categories, whereas EfficientNet-B0 showed marginally lower overall performance.

These patterns suggest that visually similar disease categories present a greater classification challenge, while more distinct classes are consistently identified with high accuracy.

Confusion Matrices

The confusion matrices for MobileNetV3-Large, EfficientNet-B0, and ResNet-50 are presented in Figures 4a-4c, illustrating class-level prediction performance. Across all three architectures, errors were limited, confirming the strong overall performance observed in the test-set metrics.

MobileNetV3-Large and ResNet-50 exhibited highly similar error profiles, with only a small number of misclassifications concentrated among closely related disease categories. In contrast, classes such as Rust and Head Smut were classified with near-perfect accuracy across all models.

EfficientNet-B0 showed a slightly higher number of errors, with misclassifications distributed across similar class boundaries. In particular, confusion among smut-related categories was more pronounced, and a small number of Rust samples were misclassified.

Overall, the consistency in both the number and distribution of errors across architectures indicates that residual misclassifications are primarily driven by inherent visual similarity between related disease categories, rather than model-specific limitations.

Computational Efficiency

Computational efficiency results are presented in Table 2. The three architectures differed substantially in both storage requirements and inference performance, providing insight into their suitability for deployment in resource-constrained environments.

MobileNetV3-Large and EfficientNet-B0 required only 16.15 MB and 15.48 MB of storage, respectively, whereas ResNet-50 required approximately 89.93 MB.

In terms of inference latency, MobileNetV3-Large achieved the fastest processing time at 27.10 ms per image, followed by EfficientNet-B0 at 56.84 ms, while ResNet-50 was substantially slower at 198.88 ms per image. Throughput measurements follow the same trend, with MobileNetV3-Large achieving the highest processing rate, followed by EfficientNet-B0 and ResNet-50.

Taken together, these results highlight clear trade-offs between predictive performance and computational efficiency. While all three architectures achieved comparable classification performance, MobileNetV3-Large provides the most favorable balance between accuracy, model size, and inference efficiency, making it well-suited for deployment in resource-constrained agricultural settings.

Although all three architectures achieved comparable classification performance, differences became more apparent when computational efficiency was considered (Table 3). MobileNetV3-Large offers the most favorable balance between predictive performance, model size, and inference efficiency, making it particularly suitable for deployment in resource-constrained environments.

Stability Analysis (Multi-seed Experiment)

To assess training stability, MobileNetV3-Large was evaluated across multiple runs with different random seeds. As shown in Table 4, performance remained highly consistent across runs, with only minor variation observed across evaluation metrics. Accuracy exhibited a standard deviation of 0.29 percentage points, while the macro-averaged F1-score showed a deviation of 0.0051.

These results indicate that model performance is stable and reproducible across repeated runs, with limited sensitivity to stochastic initialization. This level of consistency suggests that the model's performance is not dependent on favorable training conditions, reinforcing its reliability for practical deployment.

Training Sensitivity Analysis

Data Augmentation

The impact of data augmentation strategies on model performance is presented in Table 5. Performance varied across augmentation configurations, with simpler transformations yielding the strongest results. Horizontal flipping achieved the highest performance, with an accuracy of 99.91% and a macro F1-score of 0.9992.

Table 2: Baseline test-set performance of the three architectures

Model	Accuracy (%)	Macro Precision	Macro Recall	Macro F1
MobileNetV3-Large	99.54	0.9906	0.9932	0.9919
ResNet-50	99.54	0.9856	0.9964	0.9907
EfficientNet-B0	98.79	0.9713	0.9922	0.9808

Table 3: Computational efficiency comparison of the three architectures

Model	Size (MB)	Inference Latency (ms/image)	Throughput (images/sec)
MobileNetV3-Large	16.15	27.10	36.90
EfficientNet-B0	15.48	56.84	17.60
ResNet-50	89.93	198.88	5.03

Table 4: Multi-seed stability results for MobileNetV3-Large showing mean and standard deviation across repeated runs with different random seeds

Metric	Mean ± Std
Accuracy (%)	99.63±0.29
Precision	0.9919±0.0080
Recall	0.9998±0.0025
F1-score	0.9945±0.0051

Table 5: Impact of data augmentation strategies on MobileNetV3-Large performance

Augmentation Strategy	Accuracy (%)	Macro F1
Horizontal flip	99.91	0.9992
Flip + Rotation	99.63	0.9969
No augmentation	99.72	0.9963
Full augmentation	99.26	0.9869

Moderate augmentation, such as the combination of flipping and rotation, resulted in slightly lower but still competitive performance. The model trained without augmentation also achieved comparable results (macro F1-score of 0.9963), indicating that augmentation provides only marginal improvements under these conditions. Additional transformations incorporating brightness and contrast jitter produced similar outcomes, suggesting limited benefit from increased augmentation complexity.

In contrast, the full augmentation pipeline led to a noticeable decrease in performance, with a macro F1-score of 0.9869. This suggests that excessive transformations may distort disease-relevant features and reduce classification accuracy. These findings indicate that augmentation strategies should be carefully aligned with domain-specific variability rather than applied indiscriminately.

Class Weighting

The effect of class-weighted loss on per-class performance is presented in Table 6. Class weighting had minimal impact on most disease classes, with no observable change for Anthracnose and Red Rot, Rust, and Cereal Grain Molds.

However, slight reductions were observed for the minority classes, particularly Head Smut and Covered Kernel Smut. The largest decrease occurred for Head Smut, where the F1-score declined from 1.0000 to 0.9867, followed by Covered Kernel Smut (from 1.0000 to 0.9888). Loose Smut also exhibited a minor reduction in performance.

These results indicate that class weighting did not improve performance under the current experimental conditions and may slightly reduce performance for minority classes.

Table 6: Effect of class-weighted loss on per-class F1-score for MobileNetV3-Large

Disease Class	F1-score (Unweighted)	F1-score (Weighted)	Δ F1
Head Smut	1.0000	0.9867	-0.0133
Covered Kernel Smut	1.0000	0.9888	-0.0112
Loose Smut	0.9981	0.9962	-0.0019
Anthracnose and Red Rot	1.0000	1.0000	0.0000
Rust	1.0000	1.0000	0.0000
Cereal Grain Molds	0.9973	0.9973	0.0000

This suggests that class imbalance is not a limiting factor for model performance in this dataset, likely due to the strong separability of disease categories.

Robustness Evaluation

The robustness of MobileNetV3-Large under different input perturbations is illustrated in Figure 5. Model performance declined under all perturbation conditions relative to the clean baseline, with varying degrees of sensitivity across distortion types.

Gaussian blur resulted in the largest degradation, with a decrease in macro F1-score of 0.1027, indicating a strong sensitivity to the loss of fine-grained spatial detail. Brightness variations produced moderate performance reductions, with decreases of 0.0593 under high brightness and 0.0486 under low brightness conditions. Similarly, Gaussian noise led to a reduction of 0.0490, reflecting moderate sensitivity to signal distortion.

In contrast, motion blur had a minimal impact on performance, with a decrease of only 0.0065 in macro F1-score, indicating relative robustness to this form of perturbation.

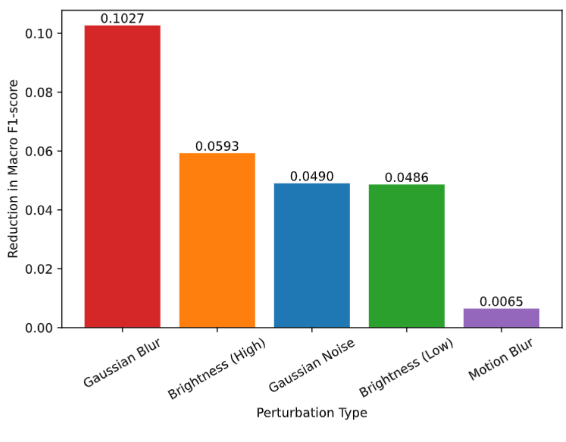

Fig. 5: Robustness of MobileNetV3-Large under controlled image perturbations, expressed as reduction in macro F1-score relative to the clean test set

Table 7: Effect of background ablation on MobileNetV3-Large performance

Training Condition	Accuracy (%)	Macro F1
Original images	99.63	0.9945
Segmented (foreground only)	88.86	0.8644

Overall, the model maintained high classification performance across all perturbation types, although sensitivity to blur and illumination changes highlights the importance of image quality in field-based deployment.

Background Ablation

The effect of background ablation on model performance is presented in Table 7. When trained on segmented images containing only foreground leaf regions, MobileNetV3-Large exhibited a substantial decline in performance compared to the baseline model trained on original images. Accuracy decreased from 99.63% to 88.86%, while the macro F1-score dropped from 0.9945 to 0.8644.

Class-level results indicate that this decline was not uniform across disease categories. Most classes retained relatively strong performance after segmentation, including Cereal Grain Molds (0.9665), Rust (0.9377), Anthracnose and Red Rot (0.9315), Loose Smut (0.9087), and Covered Kernel Smut (0.8810). In contrast, Head Smut exhibited a substantial decrease in F1-score to 0.5613, driven by low precision (0.3989) despite high recall (0.9467), indicating a significant increase in false positive predictions for this class.

These findings demonstrate that while foreground disease features remain informative, contextual information plays a critical role in class separability, particularly for disease categories characterized by broader structural or panicle-level visual patterns.

Discussion

Model Performance and Deployment Considerations

The experimental results demonstrate that all three architectures, MobileNetV3-Large, EfficientNet-B0, and ResNet-50, achieve high classification accuracy for sorghum leaf disease detection under a unified training pipeline. Despite minor differences in performance, the models exhibit broadly comparable predictive capability, with similar class-wise behavior and misclassification patterns. This indicates that each architecture is capable of effectively learning discriminative features for the disease categories represented in the dataset.

Although predictive performance is similar across models, differences become more apparent when computational efficiency is considered. MobileNetV3-Large and ResNet-50 attained the highest accuracy, while EfficientNet-B0 performed slightly lower but remained competitive. However, these differences are small relative to the substantial variation in model size and inference speed. While EfficientNet architectures are designed to achieve strong performance through compound scaling (Tan and Le, 2019), the results in this study suggest that such advantages are less pronounced when disease categories are highly separable.

From a deployment perspective, MobileNetV3-Large provides the most favorable balance between accuracy, model size, and inference latency. Its compact architecture and fast processing time make it particularly suitable for resource-constrained environments, such as smartphone-based or offline diagnostic applications. The suitability of lightweight architectures for mobile agricultural diagnostics has been demonstrated in prior work (Ramcharan et al., 2017; Kumar et al., 2025b). In contrast, ResNet-50, while accurate, incurs significantly higher computational cost, limiting its practicality for edge deployment. EfficientNet-B0 offers a compromise between these extremes but does not match the efficiency of MobileNetV3-Large.

These findings highlight that model selection for agricultural disease classification should not be based solely on accuracy. Instead, deployment considerations, including computational efficiency and hardware constraints, play a critical role in determining the most suitable architecture.

Data and Representation Effects

The results suggest that the dataset provides strong, readily learnable visual signals for disease classification. The six disease categories exhibit relatively distinctive symptom patterns under the captured conditions, enabling high precision and recall with minimal ambiguity for most classes. The near-saturation performance observed across models likely reflects the relatively distinct visual characteristics of the disease categories under these controlled conditions and may not fully capture the variability and ambiguity present in real-world field scenarios.

The training sensitivity analysis supports this interpretation. Data augmentation experiments showed that simple transformations, such as horizontal flipping, yielded the strongest performance, while more complex augmentation strategies provided only marginal benefits and, in some cases, reduced performance. Although augmentation is widely used to improve generalization in image classification tasks (Shorten and Khoshgoftaar, 2019), its limited impact in this study suggests that augmentation strategies must be carefully aligned with domain-specific variability.

Similarly, the class-weighting analysis revealed that rebalancing the loss function did not improve performance and, in some cases, led to slight reductions for minority classes. This indicates that class imbalance is not a primary limiting factor, likely due to the strong separability of disease categories.

Taken together, these findings indicate that model performance is driven primarily by dataset characteristics rather than advanced training strategies. The limited impact of augmentation and class weighting further supports the interpretation that the dataset contains strongly separable visual features, reducing the relative contribution of these techniques under the current experimental conditions.

Robustness and Context Dependence

Although the models achieved strong performance under baseline conditions, the additional experiments reveal important dependencies on both image quality and contextual information. The robustness analysis shows that performance declines under common perturbations, with the greatest sensitivity observed for Gaussian blur, indicating a reliance on fine-grained spatial detail for accurate classification. Moderate degradation under brightness variation and noise further suggests that model predictions are sensitive to variations commonly encountered in field conditions.

More importantly, the background ablation results provide deeper insight into the nature of the learned representations. The substantial decline in performance when trained on foreground-only images indicates that the models do not rely exclusively on localized disease symptoms, but instead incorporate broader contextual cues present in the original images. These cues may include plant structure, spatial distribution of symptoms, and background characteristics associated with specific disease conditions.

The class-specific impact of ablation further supports this interpretation. While most disease categories maintained relatively high performance, the sharp decline observed for Head Smut suggests that some classes depend more on contextual or structural features than on localized leaf texture alone. This indicates that disease discrimination in real-world settings may involve a combination of symptom-level and context-level information, rather than purely lesion-based features.

These findings highlight an important limitation of standard image classification approaches for plant disease detection. Models trained on field images may partially exploit contextual correlations that are not causally related to the disease itself, potentially reducing robustness under changing environmental conditions. At the same time, the results suggest that contextual information is not merely noise, but can provide complementary signals that improve classification performance when appropriately captured.

From a deployment perspective, this dual dependence

DOI: 10.3844/jcssp.2026.2992.3006

has important implications. Reliable performance in real-world applications requires not only high-quality image capture but also sufficient contextual information to support accurate classification. This underscores the need for careful dataset design, robustness evaluation, and potentially the integration of context-aware or explainability-driven approaches to ensure that models generalize beyond the conditions observed during training. This behavior is consistent with emerging evidence that deep learning models may rely on spurious correlations or shortcut features when trained on complex natural images, including background cues that are not causally related to the task, which can limit generalization and robustness (Steinmann et al., 2024; Geirhos et al., 2020).

Comparison With Existing Studies

Compared with existing sorghum disease studies, which often rely on smaller datasets, focus on fewer disease classes, or employ earlier convolutional neural network architectures (Senthil et al., 2024; Gonzalez et al., 2024; Varur et al., 2024; Desai et al., 2025), this work provides a more comprehensive and standardized evaluation of modern deep learning models. By benchmarking three widely used architectures under identical training conditions using a publicly available dataset, the study enables a more controlled and comparable assessment of model performance.

Beyond reporting overall accuracy, this work extends existing literature by incorporating class-wise performance analysis and computational efficiency metrics, which are critical for practical deployment. The inclusion of latency and model size alongside predictive performance provides a more complete evaluation of suitability for real-world applications, particularly in resource-constrained agricultural settings.

In addition, the robustness and background ablation experiments represent important contributions that are less commonly addressed in related work. By explicitly evaluating the impact of image perturbations and contextual information, this study provides deeper insight into model behavior under realistic conditions. It highlights limitations that may not be apparent from standard benchmark evaluations.

Overall, this work contributes to the growing body of AI-driven agricultural research by establishing a reproducible, deployment-oriented benchmark for sorghum disease classification, while helping position sorghum more centrally within the broader AI-for-agriculture literature (Kutyauripo et al., 2025).

Limitations and Directions for Further Validation

Despite the strong performance observed across all models, several limitations should be considered when interpreting the results. First, the dataset represents a specific agronomic and environmental context, which may limit generalization to different regions, cultivars, and imaging conditions, particularly where disease presentation and environmental context differ substantially from those captured in the dataset. Second, the high classification accuracy may partly reflect the relatively distinct visual characteristics of the disease categories, whereas real-world scenarios often involve greater variability in lighting, occlusion, and image quality.

Third, while stability analysis confirmed consistent performance across repeated runs, the study did not include formal statistical significance testing across architectures. Although performance differences between models are small, this study emphasizes practical deployment trade-offs rather than statistical superiority; however, future work could incorporate formal statistical testing to further validate the observed differences.

Fourth, while robustness was evaluated using synthetic perturbations, real-world image degradation is often more complex. Validation on independent datasets or user-generated field images would further strengthen conclusions regarding generalizability and provide a more realistic assessment of model performance across diverse environmental and imaging conditions. Finally, although background ablation highlights the importance of contextual information, the extent to which models rely on disease-specific features versus contextual cues remains unclear. Further analysis using explainability or region-based methods could provide deeper insight into model decision-making.

These limitations highlight the need for broader validation across environments, imaging conditions, and deployment scenarios to ensure reliable real-world performance.

Practical Implications

The combined performance and efficiency results indicate strong potential for deploying lightweight deep learning models in real-world sorghum disease diagnostics. MobileNetV3-Large, in particular, offers a practical balance of accuracy, model size, and inference speed, making it well-suited for smartphone-based or offline applications in resource-constrained farming environments.

The robustness and background ablation findings highlight important deployment considerations. Model performance is sensitive to image quality and the presence of contextual information, suggesting that reliable field use depends on clear image capture and sufficient visual context. This has implications for application design, including user guidance on image acquisition and the integration of image quality checks.

More broadly, the results support the feasibility of integrating AI-based disease detection into extension services and digital agriculture platforms. Lightweight models can enable rapid, on-device diagnostics, reducing reliance on connectivity and improving access to timely decision support. At the same time, the observed sensitivities underscore the need to complement automated predictions with user awareness and, where possible, additional sources of agronomic knowledge.

Future Directions

Future work could explore the development of mobile diagnostic applications that integrate deep learning models with Indigenous Knowledge practices used by smallholder farmers. Such an approach would combine algorithmic predictions with culturally embedded diagnostic cues, potentially enhancing usability, promoting farmer trust, and improving the relevance of diagnostic tools in practice.

In addition, further research could focus on improving model robustness under diverse field conditions, including variations in lighting, occlusion, and image quality, as well as evaluating performance across different geographical regions and sorghum cultivars. Advances in explainability and context-aware modeling may also provide deeper insight into model decision-making and help mitigate reliance on unintended visual cues.

Conclusion

This study presented a comprehensive benchmark of three deep learning architectures, MobileNetV3-Large, EfficientNet-B0, and ResNet-50, for sorghum leaf disease detection under a unified training framework. All models achieved high accuracy, demonstrating the effectiveness of modern convolutional neural networks for this task.

Despite comparable predictive performance, clear differences in computational efficiency were observed, with MobileNetV3-Large providing the most favorable balance between accuracy, model size, and inference speed. These findings highlight the importance of considering deployment constraints, rather than accuracy alone, when selecting models for real-world agricultural applications. Beyond baseline performance, the additional analyses revealed important dependencies on image quality and contextual information. In particular, the substantial performance decline observed under background ablation indicates that models rely not only on localized disease features but also on broader contextual cues present in field images. This behavior aligns with emerging evidence that deep learning models may exploit shortcut features or spurious correlations, which can limit generalization under changing environmental conditions. These findings underscore the importance of robustness evaluation and context-aware model design in the development of reliable plant disease detection systems. Future work should therefore focus on improving model generalization across diverse field conditions, incorporating explainability techniques to understand model decision-making better, and exploring approaches that balance the use of both symptom-level and contextual information.

This study establishes a reproducible, deployment-oriented benchmark while providing new insights into the role of contextual information in deep learning-based plant disease detection, thereby supporting the development of robust, field-ready diagnostic systems for resource-constrained environments.

Acknowledgment

The author gratefully acknowledges Dr. William Vambe and Dr. Courage Matobobo for their insightful reviews and constructive feedback. Their comments and suggestions significantly contributed to improving the clarity, rigor, and overall quality of this manuscript.

Funding Information

The author acknowledges Walter Sisulu University for providing financial support toward the publication fees of this manuscript.

Ethics

This study uses only publicly available secondary data, involves no human or animal subjects, and contains no personal, sensitive, or identifiable information. Ethical approval was therefore not required.

During the preparation of this manuscript, the author utilized ChatGPT (OpenAI, GPT-5.1) to support tasks such as content organization, refinement of wording, and improvement of clarity, coherence, and readability. All AI-generated suggestions were critically reviewed, verified, and extensively edited by the author to ensure accuracy, impartiality, and alignment with the author's own analysis, interpretation, and original contributions. The author takes full responsibility for the content of this article and confirms that AI tools were used in a manner that protects data privacy, adheres to intellectual property requirements, and maintains the integrity of the scholarly process.

References

Ahn, E., Fall, C., Botkin, J., Curtin, S., Prom, L. K., & Magill, C. (2023). Inoculation and Screening Methods for Major Sorghum Diseases Caused by Fungal Pathogens: Claviceps africana, Colletotrichum sublineola, Sporisorium reilianum, Peronosclerospora sorghi and Macrophomina phaseolina. *Plants*, *12*(9), 1906. https://doi.org/10.3390/plants12091906

Bebber, D. P. (2015). Range-Expanding Pests and Pathogens in a Warming World. *Annual Review of Phytopathology*, *53*(1), 335–356. https://doi.org/10.1146/annurev-phyto-080614-120207

Bheemanahalli, R., Wang, C., Bashir, E., Chiluwal, A., Pokharel, M., Perumal, R., Moghimi, N., Ostmeyer, T., Caragea, D., & Jagadish, S. V. K. (2021). Classical phenotyping and deep learning concur on genetic control of stomatal density and area in sorghum. *Plant Physiology*, *186*(3), 1562–1579. https://doi.org/10.1093/plphys/kiab174

Bisong, E. (2019). Google Colaboratory. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9(13), 1-73.

Chen, J., Chen, J., Zhang, D., Sun, Y., & Nanehkaran, Y. A. (2020). Using deep transfer learning for image-based plant disease identification. *Computers and Electronics in Agriculture*, *173*, 105393. https://doi.org/10.1016/j.compag.2020.105393

Desai, P., Motagi, A., Srithan, M. S., Anmol, A., Thushara, S., & Varur, S. (2025). Severity Detection of Fall Armyworm Infestation in Sorghum Using Deep Learning. *Data Science and Applications*, 85–97. https://doi.org/10.1007/978-981-96-1188-1_7

Dunjana, N., Dube, E., Chauke, P., Motsepe, M., Madikiza, S., Kgakatsi, I., & Nciizah, A. (2022). Sorghum as a household food and livelihood security crop under climate change in South Africa: A review. *South African Journal of Science*, *118*(9/10), 311–318. https://doi.org/10.17159/sajs.2022/13340

Ferentinos, K. P. (2018). Deep Learning Models for Plant Disease Detection and Diagnosis. *Computers and Electronics in Agriculture*, *145*, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Genze, N., Ajekwe, R., Güreli, Z., Haselbeck, F., Grieb, M., & Grimm, D. G. (2022). Deep learning-based early weed segmentation using motion blurred UAV images of sorghum fields. *Computers and Electronics in Agriculture*, *202*, 107388. https://doi.org/10.1016/j.compag.2022.107388

Ghosal, S., Zheng, B., Chapman, S. C., Potgieter, A. B., Jordan, D. R., Wang, X., Singh, A. K., Singh, A., Hirafuji, M., Ninomiya, S., Ganapathysubramanian, B., Sarkar, S., & Guo, W. (2019). A Weakly Supervised Deep Learning Framework for Sorghum Head Detection and Counting. *Plant Phenomics*, *2019*, 1525874. https://doi.org/10.34133/2019/1525874

Gonzalez, E. M., Zarei, A., Calleja, S., Christenson, C., Rozzi, B., Demieville, J., Hu, J., Eveland, A. L., Dilkes, B., Barnard, K., Lyons, E., & Pauli, D. (2024). Quantifying leaf symptoms of sorghum charcoal rot in images of field-grown plants using deep neural networks. *The Plant Phenome Journal*, *7*(1), e20110. https://doi.org/10.1002/ppj2.20110

Grijalva, I., Spiesman, B. J., & McCornack, B. (2023). Computer vision model for sorghum aphid detection using deep learning. *Journal of Agriculture and Food Research*, *13*, 100652. https://doi.org/10.1016/j.jafr.2023.100652

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.1109/cvpr.2016.90

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching for MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324. https://doi.org/10.1109/iccv.2019.00140

Khalid, F., & Romle, A. A. (2023). Herbal Plant Image Classification using Transfer Learning and Fine-Tuning Deep Learning Model. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *35*(1), 16–25. https://doi.org/10.37934/araset.34.3.1625

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *ArXiv*, 1–15. https://doi.org/10.48550/arXiv.1412.6980

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kumar, A., Monga, H. P., Brahma, T., Kalra, S., & Navas. (2025a). Mobile-Friendly Deep Learning for Plant Disease Detection: A Lightweight CNN Benchmark Across 101 Classes of 33 Crops. *ArXiv Preprint*.

Kumar, D., Pawar, P. P., Addula, S. R., Meesala, M. K., Oni, O., Cheema, Q. N., Haq, A. U., & Sajja, G. S. (2025b). AI-Powered Security for IoT Ecosystems: A Hybrid Deep Learning Approach to Anomaly Detection. *Journal of Cybersecurity and Privacy*, *5*(4), 90. https://doi.org/10.3390/jcp5040090

Kutyauripo, I., Rushambwa, M., & Palaniappan, R. (2025). Applications of Artificial Intelligence in sorghum and millet farming. *Circular Agricultural Systems*, *5*(1), e003. https://doi.org/10.48130/cas-0025-0001

Li, H., Wang, P., & Huang, C. (2022). Comparison of Deep Learning Methods for Detecting and Counting Sorghum Heads in UAV Imagery. *Remote Sensing*, *14*(13), 3143. https://doi.org/10.3390/rs14133143

Lin, Z., & Guo, W. (2020). Sorghum Panicle Detection and Counting Using Unmanned Aerial System Images and Deep Learning. *Frontiers in Plant Science*, *11*, 534853. https://doi.org/10.3389/fpls.2020.534853

Miao, C., Guo, A., Thompson, A. M., Yang, J., Ge, Y., & Schnable, J. C. (2021). Automation of leaf counting in maize and sorghum using deep learning. *The Plant Phenome Journal*, *4*(1), e20022. https://doi.org/10.1002/ppj2.20022

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*, *7*, 1419. https://doi.org/10.3389/fpls.2016.01419

Mwamahonje, A., Mdindikasi, Z., Mchau, D., Mwenda, E., Sanga, D., Garcia-Oliveira, A. L., & Ojiewo, C. O. (2024). Advances in Sorghum Improvement for Climate Resilience in the Global Arid and Semi-Arid Tropics: A Review. *Agronomy*, *14*(12), 3025. https://doi.org/10.3390/agronomy14123025

NVIDIA Corporation. (2018). NVIDIA Tesla T4: Product Specifications. *NVIDIA Developer / Product Documentation*.

NVIDIA Corporation. (2025). CUDA C++ Programming Guide Release 13.0 NVIDIA Corporation. *NVIDIA Documentation / CUDA Toolkit*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems. Curran Associates, Inc*, 8024–8035.

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. *Frontiers in Plant Science*, *8*, 01852. https://doi.org/10.3389/fpls.2017.01852

Rossum, G. van, & Drake, F. L. Jr. (2009). *Python 3 reference manual*.

Rother, C., Kolmogorov, V., & Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, *23*(3), 309–314. https://doi.org/10.1145/1015706.1015720

Sandip, T., & Kailas, P. (2024). Sorghum Disease Image Dataset. *Mendeley Data*. https://doi.org/10.17632/fgb3cmfrg2.1

Santiago, G. N., Magalhaes, P. H. C., Carcedo, A. J. P., Marziotte, L., Mayor, L., & Ciampitti, I. A. (2024). Deep Learning Methods Using Imagery from a Smartphone for Recognizing Sorghum Panicles and Counting Grains at a Plant Level. *Plant Phenomics*, *6*, 0234. https://doi.org/10.34133/plantphenomics.0234

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The Global Burden of Pathogens and Pests on Major Food Crops. *Nature Ecology and Evolution*, *3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Senthil, A. K., Kesavan, S., Neeraj, K., Sharath, N. B., Sasikala, K., & Addisu, B. (2024). Detection of Leaf Blight Disease in Sorghum Using Convolutional Neural Network. *Intelligent Computing for Sustainable Development*, 123–134. https://doi.org/10.1007/978-3-031-61298-5_10

Shah, S. R., Qadri, S., Bibi, H., Shah, S. M. W., Sharif, M. I., & Marinello, F. (2023). Comparing Inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A Case Study on Early Detection of a Rice Disease. *Agronomy*, *13*(6), 1633. https://doi.org/10.3390/agronomy13061633

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. In *Journal of Big Data* (Vol. 6, Issue 1, pp. 1–48). https://doi.org/10.1186/s40537-019-0197-0

Sithole, A., & Olorunfemi, O. D. (2024). The Adoption of Sustainable Farming Practices by Smallholder Crop Farmers: Micro-Level Evidence from North-Eastern South Africa. *Agriculture*, *14*(12), 2370. https://doi.org/10.3390/agriculture14122370

Steinmann, D., Divo, F., Kraus, M., Wüst, A., Struppek, L., Friedrich, F., & Kersting, K. (2024). Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *ArXiv*, 05152. https://doi.org/10.48550/arXiv.2412.05152

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

Tolosa, A. A., Dadi, D. K., Mirkena, L. W., Erena, Z. B., & Liban, F. M. (2023). Impacts of Climate Variability and Change on Sorghum Crop Yield in the Babile District of Eastern Ethiopia. *Climate*, *11*(5), 99. https://doi.org/10.3390/cli11050099

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A Comparative Study of Fine-Tuning Deep Learning Models for Plant Disease Identification. *Computers and Electronics in Agriculture*, *161*, 272–279. https://doi.org/10.1016/j.compag.2018.03.032

Varur, S., Maidur, T., Sharad, C. K., Maladkar, S. M., & Bhajantri, S. H. (2024). Severity Calculation of Tar Disease in Sorghum Using Deep Learning. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/i2ct61223.2024.10543459

Waldamichael, F. G., Debelee, T. G., Schwenker, F., Ayano, Y. M., & Kebede, S. R. (2022). Machine Learning in Cereal Crops Disease Detection: A Review. *Algorithms. MDPI*, *15*(3), 75. https://doi.org/10.3390/a15030075