Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.089112

REVIEW

A Comparative Analysis of Lightweight Deep Learning Models for Spectral-Based Detection of Crop Diseases

Geoffry Mutiso[1,*], **Geoffrey Mariga**[1] **and Aaron Oirere**[2]

[1]Department of Information Technology, Murang'a University of Technology, Nairobi, Kenya
[2]Department of Computer Science, Murang'a University of Technology, Nairobi, Kenya

*Corresponding Author: Geoffry Mutiso. Email: gmutiso@mut.ac.ke

Received: 14 July 2026; Accepted: 07 September 2026; Published: 12 October 2026

ABSTRACT: Huanglongbing (HLB) is a destructive citrus disease whose management benefits from early detection before symptoms become widespread. This study combined a PRISMA-guided systematic literature review with a controlled benchmark of seven lightweight convolutional neural networks for 14-band multispectral HLB classification. The review screened 100 unique records published between 2018 and 2025 and retained 29 studies, whose findings informed the selection of MobileNetV2, MobileNetV3, SqueezeNet, ShuffleNetV1, ShuffleNetV2, EfficientNetB5, and GhostNet. The benchmark used 2978 *in-situ* citrus images represented by 14 VisNIR bands (405–850 nm). ImageNet-pretrained input layers were adapted to accept 14 channels, and preprocessing parameters were estimated from training data only. A stratified 10% hold-out set remained unseen while five-fold cross-validation was performed within the 90% development pool. Binary precision, recall, specificity, and F1-score treated HLB-symptomatic as the positive class. MobileNetV3 achieved the highest mean cross-validation accuracy of 82.90 ± 1.23% (95% CI: 81.38–84.42) and F1-score of 82.02 ± 1.23%, and it achieved 82.6% accuracy and 81.0% F1-score on the independent hold-out set. The Friedman test indicated an overall architecture effect on fold-level accuracy (chi-square (6) = 21.429, p = 0.0015, Kendall's W = 0.714), although no MobileNetV3 pairwise comparison remained significant after Holm adjustment. Thus, MobileNetV3 is identified as the strongest descriptive common-protocol baseline rather than a statistically superior architecture. Because the source labels distinguish HLB-symptomatic from non-HLB-symptomatic scenes and do not include a separately confirmed asymptomatic-infected class, the findings support architecture selection for future pre-symptomatic research but do not constitute direct clinical validation of pre-symptomatic infection.

KEYWORDS: Huanglongbing; lightweight convolutional neural networks; multispectral imaging; cross-validation; statistical comparison; edge deployment

1 Introduction

HLB is a catastrophic bacterial disease that affects citrus production globally and causes severe yield loss and long-term orchard decline. The disease is associated with Candidatus *Liberibacter asiaticus* (CLas) and is transmitted primarily by the Asian citrus psyllid. Because no curative field treatment is currently available, management relies on rapid identification and removal of infected trees [1]. Detection is especially important during the early and pre-symptomatic stages, when visible signs are absent or limited. Visual inspection is unreliable at this stage, whereas polymerase chain reaction (PCR) testing is laboratory-dependent, costly, time-consuming, and impractical for continuous large-scale orchard surveillance [2].

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Spectral imaging, including multispectral and hyperspectral imaging, is sensitive to disease-induced changes in leaf and canopy reflectance [3]. HLB alters chlorophyll concentration, photosynthetic activity, internal tissue structure, and water relations, producing spectral responses that may precede or extend beyond clearly visible symptoms. Deep learning can learn these spectral-spatial patterns automatically, but many high-capacity CNN and transformer models require substantial memory and computation [4,5]. This limits their use on mobile, embedded, or edge devices in orchards. Lightweight architectures such as MobileNet, ShuffleNet, SqueezeNet, GhostNet, and EfficientNet reduce parameter count and computational demand [6]. But aggressive channel compression may also suppress weak narrow-band information. A rigorous comparison under a common multispectral protocol is therefore required before selecting a backbone for pre-symptomatic HLB model development.

The study was designed as an evidence-to-experiment pipeline. First, the Systematic Literature review (SLR) identified the lightweight model families most frequently used in spectral and plant-disease research, the efficiency mechanisms represented by those families, and recurring methodological gaps. These findings informed both the selection of seven representative architectures and the benchmark design. Second, the selected models were evaluated on a 14-band *in-situ* citrus dataset to determine which architecture provides the strongest accuracy-efficiency baseline for subsequent spectral-aware optimization. The empirical benchmark retains pre-symptomatic HLB detection as its central application objective. The public dataset is relevant because its HLB images were acquired from newly symptomatic trees with low disease intensity and contain large canopy regions in which the number of visibly affected leaves varies; consequently, much of the scene may be visually non-symptomatic while still belonging to an HLB-affected tree. The present study therefore evaluates the ability of lightweight networks to learn weak spectral cues relevant to early detection, while acknowledging that definitive clinical validation requires a separately labelled set of laboratory-confirmed asymptomatic infected trees. Section 2 describes the SLR and experimental methods, Section 3 presents the comparative results and limitations, and Section 4 concludes the study.

2 Materials and Methods

2.1 Systematic Literature Review

A SLR was conducted to identify, analyze, and synthesize existing research on lightweight deep learning models applied to spectral-based plant disease detection, with particular emphasis on HLB. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure transparency, reproducibility, and methodological rigor. This SLR directly established evidence-based insights into commonly adopted lightweight architectures, their reported performance in spectral learning tasks, and the methodological gaps that motivate further comparative experimentation.

2.1.1 Identification of Studies and Search Strategy

The review followed PRISMA 2020 reporting principles [7]. Searches were conducted in IEEE Xplore, ScienceDirect, SpringerLink, and the MDPI platform; Google Scholar was used only as a supplementary source for backward and forward citation searching because its ranking and result counts are not fully reproducible. The search covered peer-reviewed journal articles and conference papers published from January 2018 to December 2025. The final searches were run on 15th December 2025 for IEEE Xplore and ScienceDirect, 16th December 2025 for SpringerLink and MDPI, and 17th December 2025 for Google Scholar citation chasing. The core Boolean query was: ("Huanglongbing" OR HLB OR "plant disease" OR "crop disease") AND (multispectral OR hyperspectral OR spectral) AND ("lightweight deep learning" OR "lightweight CNN" OR MobileNet OR ShuffleNet OR SqueezeNet OR GhostNet OR EfficientNet) AND (classification OR detection). The query was adapted to the syntax and field restrictions of each source. Full

database-specific strings, search fields, filters, dates, and retrieved counts are provided in Table 1. The four primary sources yielded 122 records, while supplementary Google Scholar citation searching identified 53 additional records, resulting in 175 records before deduplication. The records were initially documented in Microsoft Excel and subsequently imported into Zotero for duplicate removal and reference management. After 75 duplicate records were removed, 100 unique records remained for title and abstract screening.

Table 1: Search log and reproducible search strategy.

Source	Role in Review	Search Fields/String	Search Date	Records
IEEE Xplore	Primary database	Core query adapted to metadata/abstract fields	15th December 2025	25
ScienceDirect	Primary database	Core query adapted to title/abstract/keywords	15th December 2025	37
SpringerLink	Primary platform	Core query adapted to full-text/discipline filters	16th December 2025	40
MDPI	Primary publisher platform	Core query adapted to title/abstract/keywords	16th December 2025	20
Google Scholar	Supplementary citation chasing	First 100 ranked results plus cited-by searches	17th December 2025	53

2.1.2 Eligibility Criteria and Quality Assessment

Studies were eligible when they: (i) addressed plant or crop disease detection; (ii) used multispectral, hyperspectral, or other explicitly spectral image data; (iii) evaluated a lightweight CNN or an efficiency-oriented lightweight module; (iv) reported sufficient information on the dataset, preprocessing, training, and evaluation; and (v) were available in English as a full peer-reviewed article. Studies were excluded when they relied exclusively on RGB imagery without a spectral component, used only conventional machine-learning methods, focused on non-agricultural tasks, evaluated only heavyweight architectures without an efficiency objective, lacked full text, or did not report quantitative outcomes. As shown in Table 2, a quality was assessed using an eight-item modified Newcastle-Ottawa-style checklist [8] covering: research-objective clarity, dataset description, spectral modality description, preprocessing transparency, model reproducibility, evaluation design, completeness of reported metrics, and discussion of limitations. Each item was scored 0 (not reported/inadequate) or 1 (adequate); studies scoring at least 6/8 were retained.

2.1.3 Study Selection Process

The selection process combined independent screening, quality assessment, data extraction, and synthesis across four sequential stages. Two reviewers independently screened titles/abstracts and full texts using the predefined criteria. Disagreements were resolved through discussion and, where necessary, adjudication by the third author. Inter-rater agreement was quantified using Cohen's kappa: $\kappa = 0.88$ for the first screening stage and $\kappa = 0.92$ for full-text decisions. The counts reported below are sequential; therefore, the denominator decreases at each stage.

The first stage focused on establishing *topic relevance* and *modality alignment*. At this stage, we carefully examined the titles, abstracts, and introductions of all retrieved papers to determine whether the study addressed plant disease detection using spectral data (hyperspectral or multispectral) and whether the

studies employed DL–based classification methods. We found that, 28 of the 100 papers, relied exclusively on RGB imagery, traditional machine learning techniques, or non-agricultural applications, leading to their exclusion. In addition to screening for relevance, preliminary data were extracted from the 72 remaining studies, including the crop type, disease of interest, and data modality, to assess alignment with spectral-based HLB detection.

Table 2: Quality-assessment checklist.

Item	Quality Criterion	Score
Q1	Clear research objective and disease-detection task	0/1
Q2	Dataset source, sample size, classes, and acquisition described	0/1
Q3	Spectral modality and wavelength information described	0/1
Q4	Preprocessing and augmentation reproducibly reported	0/1
Q5	Architecture and training protocol reproducibly reported	0/1
Q6	Evaluation design and partitioning appropriate	0/1
Q7	Quantitative metrics and uncertainty sufficiently reported	0/1
Q8	Limitations and deployment considerations discussed	0/1
Threshold	Retain studies scoring at least 6/8	≥6

The second stage evaluated the *architectural relevance* and *contributions of the studies*. In this stage, we fully screened the remaining 72 papers to determine whether they had employed or analyzed lightweight deep learning architectures, such as MobileNet, ShuffleNet, SqueezeNet, EfficientNetB5, or GhostNet and whether contribution of each study was clear. At this stage, 19 studies were excluded because they had focused on heavyweight architectures and had failed to justify architectural efficiency. This resulted in 53 papers. We then extracted data from the papers regarding the model family and variant, network depth, and any reported computational efficiency indicators, directly informing the analysis required to answer the first research question.

The third stage assessed the *methodological completeness* and *experimental transparency of each study*. The 53 papers from stage two were examined to confirm whether they clearly documented dataset characteristics, including sample size, spectral dimensionality, preprocessing procedures, and training protocols. The results showed that 14 of the 53 examined papers lacked sufficient methodological detail or did not provide reproducible experimental descriptions and were excluded. The remaining 39 papers were further screened to extract detailed methodological elements for analysis, including data preprocessing strategies, training-validation splits, and experimental setup, supporting the investigation of how effectively lightweight models capture spectral cues.

The fourth stage focused on the *evaluation methodology* and *limitations of the studies*. The remaining papers were required to report quantitative performance metrics and provide meaningful experimental results. We found that only 29 studies had qualified this criterion and the 10 studies that omitted evaluation metrics, failed to report classification outcomes, and failed to provide adequate comparative analysis and were excluded. For the 29 papers that satisfied this criterion, data were extracted on reported performance metrics, class-level behavior, and any discussion of model limitations or deployment considerations, enabling analysis related to attention mechanisms, performance trade-offs, and remaining research gaps. We also examined the conclusion and future work sections of each paper to identify the gaps left or identified as future work and whether the gap was resolved by any paper.

After applying all four stages, a final set of 29 studies, as summarized in Table 3 and Fig. 1, satisfied all selection criteria and formed the primary corpus for synthesis and comparative analysis. In total, 71 studies were excluded across the four stages (28 + 19 + 14 + 10 = 71), yielding 29 studies that met all inclusion criteria for topic relevance, architectural focus, methodological completeness, and evaluation rigor. This structured process ensured that the studies analyzed in subsequent sections were methodologically sound, relevant, and directly aligned with the research questions guiding this review. Fig. 1 summarizes the PRISMA sequential selection flow across all four stages.

Table 3: Sequential study-selection results.

Stage	Records Entering Stage	Excluded at Stage	Primary Reason	Records Remaining
Identification after deduplication	100	0	Unique records retained for screening	100
Stage 1: relevance and spectral modality	100	28	RGB-only, traditional ML, non-agricultural, or off-topic	72
Stage 2: lightweight architectural relevance	72	19	Heavyweight model focus or no efficiency contribution	53
Stage 3: methodological completeness	53	14	Insufficient dataset, preprocessing, or training detail	39
Stage 4: evaluation rigor	39	10	No quantitative outcomes or inadequate comparative reporting	29

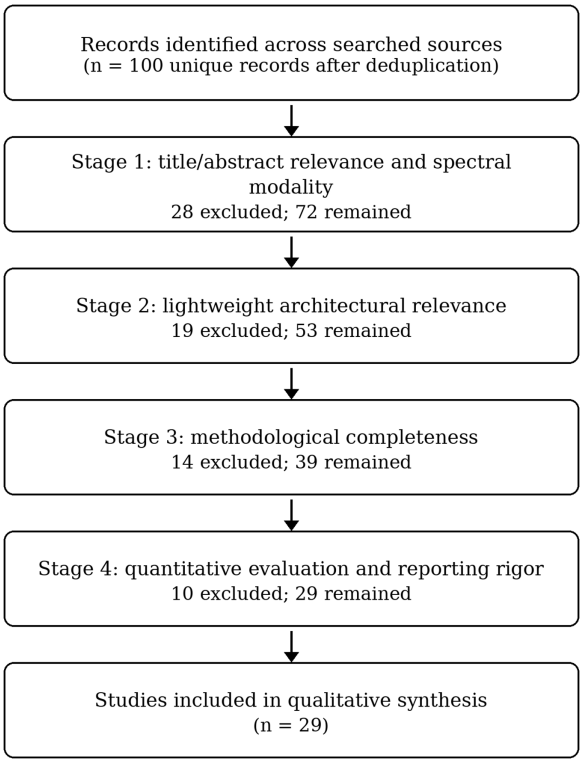


Figure 1: PRISMA-based sequential study-selection flow.

2.1.4 SLR Findings

The final set of studies was analyzed to address four Research Questions (RQs): (RQ1) which lightweight deep learning models have been used for spectral-based plant disease and HLB detection; (RQ2) how effective these models are in capturing spectral cues relevant to pre-symptomatic HLB; (RQ3) how attention mechanisms have been integrated into lightweight architectures and their impact on performance; and (RQ4) what methodological and architectural gaps remain in existing research. The findings from this analysis provided a structured understanding of current trends and limitations and directly informed the comparative analysis and experimental evaluation presented in subsequent sections of this study.

RQ1, Lightweight architectures used: The results revealed that MobileNet variants were the most frequently adopted lightweight architectures in the reviewed studies, followed by EfficientNet, ShuffleNet, SqueezeNet, and GhostNet as shown in Table 4.

Table 4: Frequency of lightweight model families across the 29 included studies.

Model Family	No. of Studies
MobileNet (V1–V3)	12
EfficientNet	7
ShuffleNet (V1–V2)	7
SqueezeNet	4
GhostNet	4

In most studies, the researchers note that they favored these models due to their low computational cost and suitability for edge deployment, although their application to multispectral data remained limited.

RQ2, Effectiveness in spectral feature learning: SLR results show that many studies reported high classification accuracy, however, performance varied significantly when models were applied to spectral data [3]. Many of these models were observed struggle in capturing subtle spectral variations associated with early disease stages, indicating limitations in spectral feature representation and model capacity [9].

RQ3, Role of attention mechanisms: In 16 out of the 29 papers analyzed, we observed integration of attention modules into lightweight backbones. For Instance, study [10] included CBAM, Triplet Attention, and EMA, into a lightweight architecture (Ghost-Attention-YOLOv8) to enhance rice leaf disease detection. The integration of the Ghost model and attention mechanisms led to significant improvements in accuracy, precision, and recall. The study reported an improved mAP@50 metric (95.4%), which is a 2.3% increase over the baseline YOLOv8 model. This improved accuracy and efficiency while reducing computational complexity for effective agricultural monitoring. Another study in [11] focused on integrating the Attention Mechanism with Squeeze-and-Excitation Networks for coffee leaf disease detection. In the results of this study, class-wise analysis indicated that SENet had F1-scores of 91% for all classes of coffee leaf diseases an improved score which was attributed to the addition of the attention mechanism. Across all the studies integrating attention mechanism, it was observed that these studies had consistently reported modest performance improvements, attributing the improvements to be informed by the emphasized attention on the informative spectral bands. However, across the reviewed corpus, attention modules were generally introduced within individual architectures rather than evaluated systematically under a common cross-architecture protocol.

RQ4, Remaining research gaps: The review identified a lack of unified lightweight frameworks that explicitly combine spectral–spatial learning with attention mechanisms. Results showed that few studies

addressed pre-symptomatic HLB detection using multispectral data, and inconsistencies in evaluation protocols limited cross-study comparison. The review also showed that class-wise precision and recall were not reported consistently across studies, making it difficult to determine whether performance differences arose from dataset composition, training strategy, or architectural compression.

2.1.5 Translation of SLR Findings into the Benchmark Design

The benchmark was not designed independently of the review. As shown in Table 5, RQ1 identified the five dominant model families and justified selecting seven representative variants. RQ2 showed that reported performance often declined when RGB-oriented networks were exposed to spectral data, motivating direct 14-band evaluation. RQ3 showed that attention modules may improve weak-feature learning; therefore, the present benchmark first isolates backbone behaviour before attention is introduced in the next model-development stage. RQ4 identified inconsistent splitting, incomplete preprocessing descriptions, limited statistical analysis, and weak deployment reporting; these gaps informed the standardized data pipeline, the explicit development/hold-out evaluation structure, the statistical-analysis plan, and the expanded efficiency metrics used in this study.

Table 5: Direct mapping from SLR findings to benchmark decisions.

SLR Finding	Benchmark Decision
MobileNet, EfficientNet, ShuffleNet, SqueezeNet, and GhostNet were the most represented families	Select these five families; retain two MobileNet and two ShuffleNet generations to test within-family evolution
Spectral performance was less stable than RGB performance	Evaluate the architectures directly on all 14 VisNIR bands under the same input resolution and metric definitions
Attention was promising but inconsistently evaluated	Establish a backbone-only baseline before adding a uniform lightweight attention module
Evaluation and deployment reporting were inconsistent	Use a two-stage development/test design, report uncertainty, and add FLOPs, memory, latency, and throughput

2.2 Comparative Analysis of Lightweight Deep Learning Models

This section synthesizes evidence from the 29 primary studies retained after systematic screening to critically examine how lightweight deep learning architectures have been applied to plant disease detection, with particular emphasis on spectral-based approaches relevant to HLB. In this study, rather than cataloging reported accuracies alone, we analyzed the studies focusing on recurring architectural strengths, failure modes, and unresolved challenges, especially in relation to subtle spectral–spatial cues and pre-symptomatic disease detection that motivates this study.

2.2.1 Overview of Lightweight Model Usage Trends

Across the reviewed studies, lightweight deep learning models were predominantly adopted to address computational constraints associated with field deployment, mobile platforms, and edge devices [6,12]. Findings of this study show that most studies evaluated one or two lightweight backbones, frequently

benchmarking them against heavier CNNs such as ResNet or VGG to demonstrate efficiency gains [13]. As noted in Table 4, MobileNet-based architectures emerged as the most frequently used backbone.

However, this study found that while the lightweight models achieved a high overall accuracy in crop disease detection, their performance declined when classifying crop disease using multispectral imagery. This decline was attributed to a greater proportion of false-positive predictions, particularly in early-stage detection where critical information is often encoded in narrow spectral bands.

2.2.2 ShuffleNet Variants

ShuffleNet architectures were identified in seven studies emphasizing efficiency-driven plant disease recognition and lightweight deployment. We observed two variants of ShuffleNet in the reviewed studies; ShuffleNet v1 and ShuffleNet V2. The first variant ShuffleNet V1 which was originally introduced as a high-performance lightweight CNN by the Megvii Technology team in 2017 as simply ShuffleNet [14]. The model introduced grouped convolutions combined with channel shuffling to reduce computational cost while maintaining information flow across feature channels, achieving approximately 5.1M parameter count. The researchers in [14] noted in their study that ShuffleNet achieved 80.95% in detecting lemon bruising using hyperspectral dataset. In their study, ResNet model achieved a high accuracy of 90.47% however, the model needed more storage and computing power to train as compared to ShuffleNet. They concluded in their finding that ShuffleNet sacrifice its performance and generalization power in complex tasks like 3D-CNN classifications in favor of its simplicity. We also observed that researchers in [15] proposed an improved ShuffleNet CNN model for detecting leaf diseases in crops such as capsicum, rice, corn, tomato, and citrus. Their findings indicate that the model achieved a 91.94% accuracy which outperformed well-known models like ResNet50 and DenseNet20. The high accuracy was attributed to the added convolutional layers, a Feature Pyramid Network (FPN), and a Region Proposal Network (RPN).

The other variant of ShuffleNet observed in the studies was ShuffleNet v2 with approximately 3.7M parameters. Its architecture refined the original design of ShuffleNet by removing group convolutions, introducing channel split operations, and enforcing practical efficiency guidelines such as equal channel width and reduced memory access cost. These changes improved gradient flow and training stability and resulted in modest performance gains across several application domains [16]. Researchers in [17] highlighted that while lightweight convolutional neural networks, like ShuffleNet V2, effectively classify tomato leaf diseases with an accuracy of 91.817%, they may struggle with early-stage or low-contrast disease detection. This is due to the reliance on dominant spatial patterns rather than narrow spectral bands, which can lead to the loss of weak but critical spectral signatures during aggressive feature compression. Thus, the model's performance may decline in detecting subtle disease symptoms. The researchers proposed utilizing ShuffleNet V2 as the backbone and incorporating Attention mechanisms so as to improve performance when the model is subjected to narrow spectral band images.

Another study by [18] proposed LA-ShuffleNet model which was a combination of an upgraded ShuffleNetV2 network structure and a novel attention module named the Adaptive Pooling Attention Module (APAM). The reported findings indicate that LA-ShuffleNet with 3.0M parameters achieved a high accuracy compared to baseline CNN models on ImageNet dataset. They concluded that combining lightweight design with attention module, the network can help achieve optimal performance. Other studies like [19,20] discussed ShuffleNet V2 and highlighted that although the ShuffleNet network can learn rich feature representations despite having fewer parameters, the basic unit of the network suffers from partial structural redundancy, as well as poor network operation speed.

The reported findings consistently indicated noticeable reductions in recall and class-wise accuracy when disease symptoms were weak or spectrally subtle. This limitation was attributed to channel manipulation strategies that, while computationally efficient, restricted the model's capacity to preserve narrowband spectral-spatial relationships. As a result, ShuffleNet variants were selected for empirical validation.

2.2.3 SqueezeNet

SqueezeNet model usage was observed in four studies out of the qualified papers in this study. This model is commonly used as feature extractor, especially in transfer learning. SqueezeNet has approximately 1.4M parameter count making it to be used consistently in the studies likely due to its extremely compact architecture and minimal memory footprint [11]. As noted in a study by researchers in [21]. The model relies on fire module, which aggressively reduce channel dimensionality through squeeze layers followed by limited expansion. The SqueezeNet architecture is built around three core design strategies and the fire module. These strategies include replacing 3 × 3 filters to 1 × 1 filters, decreasing the number of input channel dimensions to 3 × 3 filters, and finally applying downsample late in the network. The fire module has 2 modules; the squeeze and the expand layers. The researchers retrained SqueezeNet and MobileNetV2 on RGB dataset with citrus diseases. The reported findings indicated that the SqueezeNet model achieved accuracy of 96% outperforming MobileNetV2 model. Researchers in [22,23] utilized SqueezeNet in their studies and indicated in their findings that SqueezeNet design enables rapid inference and low storage requirements, making it attractive for deployment on highly constrained devices. However, the excessive channel compression can reduce the model's ability to retain discriminative spectral information across contiguous wavelength bands. It was observed that across the qualified studies, there is no particular study that utilized multispectral image dataset to detect crop disease. Therefore, an experiment is required to show if the model can be able to classify diseases on multispectral band image dataset input.

2.2.4 MobileNet Variants

MobileNet architectures were the most prevalent lightweight models across the reviewed literature, with MobileNetV2 and MobileNetV3 appearing in twelve studies. The original MobileNet introduced depthwise separable convolutions to reduce computational cost. A variant named MobileNetV2 proposed in 2018 incorporated inverted residual blocks with linear bottlenecks, improving feature reuse and representational efficiency [24]. In 2019, another improved variant of MobileNet was introduced with name; MobileNetV3. This version further refined the architecture through neural architecture search (NAS), lightweight attention integration via squeeze-and-excitation modules, and the adoption of h-swish activation functions [25]. These variants are most used because the model architecture is flexible and comes with global hyperparameters including width multiplier (α) and Resolution multipliers (ρ). These global hyperparameters can be tuned for specific latency and accuracy trade-offs.

Study [9] addresses the challenge of citrus fruit and leaf diseases by proposing an automated deep learning-based detection model. It proposes a lightweight classifier using MobileNetV2 which was trained on a Kaggle dataset of citrus images categorized into greening (HLB), black spot, citrus canker, and healthy samples. Data augmentation techniques such as rotation, scaling, and flipping were applied to improve generalization, while optimization strategies like Adam, early stopping, and learning rate reduction ensured efficient convergence. The findings show that MobileNetV2 achieved a better performance with an overall accuracy of 95% in detecting fruit disease and 99% in detecting leaf disease outperforming InceptionV3 and standard CNN models in both accuracy and processing speed. Researchers in [24] highlights that MobileNetV2 as a balanced choice between accuracy and efficiency, making it a strong candidate for real-time, resource-constrained agricultural diagnostics. They trained this model on a dataset obtained from

Kaggle with RGB orange disease images. The findings indicate that the model yielded an overall accuracy of 95% in detecting orange disease (Blackspot, Canker, Fresh, and Greening Citrus, using a deep learning approach). Other studies like [26–28] also utilized MobileNetV2 in their studies and their findings indicate that the model was able to show high accuracies in detecting citrus diseases. However, there is no specific study that focused on detecting the citrus diseases especially before the symptoms manifest themselves. The reviewed studies utilizing MobileNetV2 were trained on RGB images and this make them difficult to detect diseases before symptoms becomes visible.

Other several studies reported MobileNetV3 as the strongest lightweight baseline, particularly when applied to spectral-based plant disease detection and when augmented with attention mechanisms such as SE or CBAM [29]. A study by [25] proposed the use of a lightweight convolutional neural network (MobileNetV3) for the rapid detection of Citrus Huanglongbing (HLB) disease, utilizing a handheld device capable of capturing both multicolor fluorescence and multispectral reflectance images simultaneously. The architectural refinements in MobileNetV3 were observed to improve channel-wise feature recalibration, contributing to enhanced discrimination of informative spectral bands. Findings indicate that MobileNetV3 model achieved an overall accuracy of 92.1% with a false negative rate of 12.1% at 33 epochs when combining multicolor fluorescence and multispectral reflectance images as input for the detection of HLB in Navel orange datasets. The findings also underscore that also fine-tuning MobileNetV3 through transfer learning method demonstrated superior transferring ability compared to the reuse-model approach, achieving an overall accuracy of 96.5% for Ponkan citrus.

Researchers in this study [30] proposed an improved MobileNetV3 network architecture specifically designed for discerning specific agricultural diseases and pests. The improvement is noted by the addition of Selective Convolution (ScConv) module to replace some modules in the original model. ScConv is composed of two key components: Selective Receptive Units (SRU) and Channel-Related Units (CRU). SRU uses gating mechanisms to selectively fuse important features, helping the model focus on crucial information and improve representation. CRU enhances inter-channel relationships by compressing and reconstructing input channels, strengthening correlations between them and boosting the model's expressive power. The study indicate that the refined network structure capitalized on the utilization of multi-level feature information extracted from images. The new developed model was MobileNetV3 small with 2.9M trainable parameters. The new model showed improvement from an initial accuracy of 0.867 of the original MobileNetV3 to a higher accuracy of 0.883 in detecting agricultural diseases and pests.

Another study utilized a lightweight MobileNetV3-Tiny model, which was enhanced with a Convolutional Block Attention Module (CBAM), as the core component of an IoT-enabled real-time plant disease detection system [29]. This model was specifically chosen and optimized for deployment on resource-constrained edge devices like Raspberry Pi. To achieve this, the MobileNetV3-Tiny model underwent optimization techniques including quantization and pruning, which successfully reduced its size to 4.3 MB. The findings demonstrated that the MobileNetV3-Tiny model achieved a high classification accuracy of 96.85%. Furthermore, it exhibited an efficient inference latency of 140 ms. The study highlighted that this performance not only met the requirements for real-time detection but also outperformed baseline models such as ResNet18 and EfficientNet-B0 in terms of computational efficiency. This indicates its suitability for cost-effective and scalable precision agriculture solutions. Studies [31,32] also utilized MobileNetV3 model to develop a model capable of detecting citrus crop diseases. The findings of the two studies demonstrate that the model can effectively detect citrus diseases.

2.2.5 EfficientNet

EfficientNet was identified in seven papers reviewed in this study. This architecture was released by Google Research indicating that it achieves state-of-the-art accuracy with significantly fewer parameters and faster inference than its predecessors. Its design is defined by two core breakthroughs: a baseline architecture found via Neural Architecture Search (NAS) and a scaling method called Compound Scaling [33]. EfficientNet variants simplified this design by removing computationally expensive operations to improve deployment feasibility. Across the reviewed studies, EfficientNet consistently demonstrated improved accuracy and more stable training behavior compared to earlier lightweight architectures. EfficientNet-based models were noted for their scalability and efficiency, balancing model complexity with accuracy, making them suitable for addressing the challenges of citrus disease classification.

Researchers in [34] specifically proposed model (EfficientNet B0) for the early detection of HLB in oranges and tomatoes pathogens. The proposed model was reported to achieve accuracy of 98% on test data. However, this study did not explicitly mention pre-symptomatic detection of the HLB. A similar study in [35] also proposed EfficientNet B0 for citrus leaf and fruit disease detection and compared the model with other convolutional neural network (CNN) models. They noted in their findings that, although EfficientNet B0 achieved 80.18/% test accuracy and benefited from hyperparameter tuning and Mixup augmentation, it showed less stable convergence and struggled with classifying certain fruit diseases, indicating its lightweight architecture might not capture subtle visual features as effectively as other models.

EfficientNet B3 was also identified in a paper in [36]. The study utilizes the EfficientNetB3 model for classifying four major orange tree diseases: canker, greasing, scab, and greening, achieving an overall accuracy of 92.25%, demonstrating its effectiveness in rapid citrus disease diagnosis and management [36]. The researchers indicate in the conclusions that although the model showed superiority in accuracy, there are some cases where model refinements is required, especially when the disease symptoms differences are very subtle. Study [37] also utilized EfficientNetB3 for detecting diseases in orange leaves. The findings of the study indicate that the EfficientB3 achieved an accuracy of 97.30% after 30 training rounds. This model outperformed VGG19, demonstrating its effectiveness in early disease detection for citrus trees. Researchers in [13] compared EfficientNet B3 with other CNN models including ResNet50, MobiNetV2, and InceptionV3. The study utilized Citrus Plant Dataset which is a RGB dataset with a specific focus on Black spot, Scab, Canker, Melanose, and Greening/Huanglongbing in data sets. The findings indicate that EfficientNetB3 model achieved impressive accuracies of 99.43% in training, 99.48% in validation, and 99.58% in testing, outperforming other CNN architectures in identifying and classifying citrus plant diseases.

Another study in [38] utilized EfficientNet, specifically EfficientNet-B5 and an improved version called EfficientNet-B5-pro, for the classification of citrus leaf diseases and nutritional deficiencies. EfficientNet-B5-pro incorporated two main enhancements: (1) Adaptive Angular Margin (Arcface) Loss which replaces the standard softmax loss. It maximizes classification boundaries in angular space and adds a cumulative angular margin between features and target weights. The design aimed to enhance intra-class compactness and inter-class variability, which is particularly beneficial given the high similarity among various citrus leaf diseases. (2) Adversarial Weight Perturbation (AWP) Mechanism which introduces noise to model weights and inputs during training to regularize parameters and improve the model's robustness by flattening the weight loss landscape, thereby reducing the generalization gap between training and test accuracy. The researchers indicate that after comparing the classification performances of several networks (Densenet121, ResNet50, ShuffleNetv2, Mlp-Mixer, MobileNetv3, EfficientNet-B3, and EfficientNet-B5), EfficientNet-B5 achieved the best performance with an accuracy of 94.9%.

2.2.6 GhostNet

GhostNet model was identified in four papers in this study. As noted in [39]. GhostNet architecture introduces ghost feature maps generated through inexpensive linear operations, aiming to approximate redundant features produced by standard convolutions. The study proposed a lightweight convolutional neural network model called Gated Asymmetrical GhostNet (GA-GhostNet). This model was designed for the automatic diagnosis of plant diseases and pests on PlantDoc dataset. It incorporated a Gated Multi-scale Coordinate Attention (GM-CA) module to filter out noise and irrelevant information in the image, while capturing location information of diseases and pests at different scales. The reported findings indicate the model having 3.7M parameters demonstrated high accuracy and efficiency in identifying plant diseases and pests. However, the evidence base for GhostNet in spectral-based plant disease detection remains limited. Another study [40] proposed integrating GhostNet module into the backbone architecture of the YOLOv8 algorithm. This was done to reduce the number of parameters, leading to a faster computational algorithm. The improved YOLOv8 algorithm, which included the Ghost module, aimed to enhance the model's detection performance and make it more responsive to leaf disease detection, especially for small leaf targets, while also reducing computational resources. The findings of their study indicate the best performance without transfer learning was achieved by YOLOv8s with the Ghost module and Coordinate Attention (CA) mechanism, yielding a mAP0.5 score of 69.3%.

A similar study was conducted integrating GhostNet into YOLOv8 architecture [10]. This study focused on detection rice leaf disease. The findings indicate that the proposed model achieved significant improvements in detection performance, with mAP@50 metric of 95.4%, representing a 2.3% increase over the baseline YOLOv8. Researchers in [41] proposed a citrus fruit detection method named HPL-YOLOv4. The proposed model integrated GhostNet module into YOLOv4 architecture to make it lighter. To improve the model performance, the study applied an efficient channel attention mechanism (ECA) to enhance the multi-scale feature weights obtained by the backbone network. The findings indicate that the proposed model achieved an average precision (AP) value of 98.21%, and a precision rate of 93.45% and increase by 9.41% and 2.26%, respectively compared to the original YOLOv4. The improvement was informed by the integration of GhostNet.

2.3 Environmental Setup

This section describes the environment and protocol used to benchmark lightweight CNNs for spectral-based early and pre-symptomatic HLB detection. The experimental objective was to identify a deployment-suitable backbone that can preserve weak spectral-spatial information in 14-band orchard imagery. The benchmark was designed as an architecture-selection baseline; it does not replace subsequent validation on an independently labelled cohort of laboratory-confirmed asymptomatic infected trees.

2.3.1 Dataset Description

Experiments used the public *in-situ* multispectral HLB dataset hosted in Recherche Data Gouv (doi:10.57745/054NAB) [42] and described in detail by Porto et al. [43]. It contains 2978 terrestrial images from a commercial 'Pera Rio' sweet-orange orchard in São Paulo, Brazil, acquired during two field campaigns: 1297 images from Plot A in August 2023 and 1681 images from Plot B in November 2024. A Sony Alpha 7R2 Sextuple multi-lens camera generated 14 VisNIR bands centred at 405, 430, 450, 490, 525, 550, 560, 570, 630, 650, 685, 710, 735, and 850 nm. The data are available as aligned monochromatic Tagged Image File Format (TIFF) images and Hierarchical Data Format version 5 (HDF5) datacubes ordered by wavelength. The two labels are HLB and non-HLB. HLB trees were identified by trained field specialists and, because the orchard was inspected four to eight times per year and symptomatic trees were rapidly removed, most

newly detected trees had low disease intensity, with no more than approximately 25% of the canopy visibly affected [43]. Each image covers a large canopy area and the number of visibly symptomatic leaves varies substantially. Consequently, HLB-labelled scenes may contain extensive visually non-symptomatic foliage, while the 14-band representation captures visible, red-edge, and near-infrared responses that are relevant to early physiological stress. These properties justify the dataset as a baseline for architectures intended for early and pre-symptomatic HLB detection. Nevertheless, the dataset does not provide a separate PCR-confirmed asymptomatic-infected class, and this limitation is stated explicitly in the Discussion.

A two-stage evaluation design was used. First, the 2978 images were divided once into a 90% development pool (2680 images) and a 10% stratified hold-out test set (298 images). Stratification preserved the HLB/non-HLB distribution and representation from both acquisition campaigns. Second, stratified five-fold cross-validation was performed only within the 90% development pool: in each fold, four partitions were used for training and one for validation. The hold-out test set was not used for preprocessing estimation, hyperparameter decisions, checkpoint selection, or model comparison and was accessed only after the development analysis was finalized. This clarifies that the five folds were not created from the full dataset and did not include the independent test samples. The public release does not provide stable tree-, branch-, or leaf-level identifiers; therefore, the split was necessarily performed at image level. The possibility that correlated images from the same biological source or acquisition sequence occurred in different partitions cannot be completely excluded and is treated as a limitation. A campaign-held-out analysis is recommended as an additional robustness test before deployment claims are made.

2.3.2 Multispectral Preprocessing and 14-Channel Input Adaptation

The HDF5 datacubes were loaded in ascending wavelength order and all 14 bands were retained; no band was removed. Residual inter-band misregistration was reduced using the dataset's ECC-based registration workflow, after which every datacube was resized to 224 × 224 pixels using bilinear interpolation. For each cross-validation fold, the mean and standard deviation of every spectral band were estimated from the training partition only. The resulting per-band z-score normalization parameters were then applied unchanged to the corresponding validation partition and to the hold-out test set, preventing information leakage. Training-only augmentation consisted of random horizontal flipping (p = 0.5), random vertical flipping (p = 0.5), rotations within ±15°, and random resized cropping with scale 0.90–1.00. Each spatial transformation was applied identically to all 14 bands so that spectral correspondence was preserved; RGB colour jitter was not used. Class imbalance was handled using inverse-frequency weights in the cross-entropy loss, calculated separately from each training fold.

The stock ImageNet models accept three-channel RGB input and therefore could not ingest the 14-band datacubes without modification. For every architecture, the first convolution was replaced by a layer with the same output channels, kernel size, stride, padding, and grouping configuration but with 14 input channels. To retain transferable low-level filters, the pretrained RGB kernels were averaged across the three input channels, repeated across the 14 spectral channels, and scaled by 3/14 to preserve the expected activation magnitude. All remaining pretrained weights were retained, and the final classifier was replaced by a two-output layer. This same input-adaptation rule was applied to all seven networks.

2.3.3 Lightweight Models Selected for Benchmarking

Seven models were selected because they represented the five most frequently reported lightweight families in the 29-study SLR and captured distinct efficiency mechanisms: MobileNetV2 and MobileNetV3, SqueezeNet, ShuffleNetV1 and ShuffleNetV2, EfficientNetB5, and GhostNet. MobileNet and ShuffleNet

each contributed two generations so that the benchmark could examine whether later design refinements improved spectral feature preservation within the same family. SqueezeNet represented aggressive fire-module compression, GhostNet represented inexpensive generation of redundant feature maps, and EfficientNetB5 provided a higher-capacity compound-scaled comparator for the accuracy-efficiency trade-off.

The selection criteria were: (i) frequency and relevance in the qualified SLR studies; (ii) architectural diversity; (iii) availability of reproducible PyTorch implementations and pretrained weights; and (iv) suitability for assessing deployment trade-offs. NASNet-Mobile and MNASNet were not included because they appeared too infrequently in the eligible spectral plant-disease literature to support the same evidence-based comparison. The seven networks collectively represent depthwise separable convolution, inverted residual bottlenecks, channel splitting and shuffling, fire modules, compound scaling, and ghost-feature generation. This diversity enabled the study to test whether different compression strategies preserve or discard narrow-band information relevant to early HLB detection.

EfficientNetB5 was retained as the upper-complexity comparator rather than being described as a sub-5-million-parameter model. Apart from the necessary 14-channel input adaptation and replacement of the final two-class classifier, no internal blocks were altered. The benchmark therefore compared the original backbone designs under a consistent multispectral interface.

2.3.4 Experimental Setting, Training Protocol, and Performance Evaluation Methods

Experiments were conducted on an NVIDIA V100S GPU with CUDA 12.1, Python 3.10, PyTorch 2.1, and TorchVision 0.16. Random seeds were fixed for Python, NumPy, and PyTorch, and deterministic CUDA behaviour was enabled where supported. The primary benchmark deliberately used the same training budget for all architectures so that observed differences could be interpreted under a controlled protocol rather than confounded by unequal tuning effort. This design does not imply that one hyperparameter configuration is optimal for every model.

Five-fold cross-validation was applied only to the 90% development pool, while the 10% hold-out test set remained untouched until final evaluation. Within each fold, model selection used validation F1-score, and preprocessing parameters were estimated from the training partition only. The common baseline used Adam, a learning rate of 0.001, batch size 8, inverse-frequency weighted cross-entropy loss, and a maximum of 100 epochs with early stopping based on validation F1. The hyper-parameters were fixed for all experiments and are summarized in Table 6.

Table 6: Hyper-parameter setup for all models.

Parameter	Common Baseline Value	Reason/Constraint
Input	14 × 224 × 224	Same multispectral representation for all models
Epochs	100 maximum	Common training budget
Optimizer	Adam	Stable baseline used in related studies
Learning rate	0.001	Controlled baseline; architecture-specific sensitivity analysis recommended
Batch size	8	V100S memory constraint for 14-band tensors
Loss	Inverse-frequency weighted cross-entropy	Addresses moderate class imbalance
Checkpoint criterion	Highest validation F1-score	Avoids selecting by training accuracy
Random seeds	42	Reproducibility

The common settings were retained as the primary benchmark because the objective was to compare architectural behaviour under an identical computational budget. Results are therefore described

as common-protocol performance and not as each model's maximum attainable accuracy. The revised manuscript explicitly separates this controlled comparison from architecture-specific optimization and reports the latter only if the equal-budget sensitivity analysis is completed. Accuracy, precision, recall, specificity, and F1-score were calculated using the HLB class as the positive class.

2.3.5 Statistical Analysis and Deployment Measurements

Cross-validation metrics were summarized as the arithmetic mean and sample standard deviation across the five paired folds. Two-sided 95% confidence intervals were calculated using the t-distribution as mean ± t (0.975,4) × SD/sqrt (5). Precision, recall, specificity, and F1-score were computed using binary averaging with HLB-symptomatic as the positive class; F1-score was calculated from the corresponding precision and recall values as 2PR/(P + R). Because the same fold assignments were shared across all seven architectures, overall performance differences were first assessed using the Friedman test, with Kendall's W reported as the omnibus effect size [44]. Following a significant omnibus result, exact two-sided Wilcoxon signed-rank tests compared MobileNetV3, the highest-mean model under the common protocol, with each alternative architecture [45] Holm adjustment controlled the family-wise error rate across the six pairwise comparisons [46] and adjusted $p < 0.05$ was considered statistically significant. The independent 10% hold-out set was used only for final descriptive verification and was not used for model selection, hyperparameter adjustment, or statistical testing. Hold-out performance is reported using accuracy, precision, recall, specificity, and F1-score.

Deployment measurements were collected from the exact instantiated models using a batch size of 1 and a 14 × 224 × 224 input tensor on the NVIDIA V100S used for training. After 100 warm-up passes, 1000 synchronized inference runs were used to calculate median latency, 95th-percentile latency, and throughput. FLOPs and MACs were computed for the same input tensor, serialized FP32 model size was measured from the saved state dictionary, and peak allocated GPU memory was recorded with the PyTorch CUDA memory profiler. These measurements complement parameter count because theoretical compactness does not necessarily predict practical latency or memory-access cost.

3 Results

3.1 Comparative Experimental Results

This section presents the comparative experimental results obtained from benchmarking seven lightweight deep learning models for early-stage HLB detection using multispectral imaging data. All results reported in this section were obtained using the multispectral orange leaf dataset, evaluated under the hybrid partitioning strategy described in Section 2.3.1, with final posttest performance assessed on the stratified hold-out test set of 298 images. The evaluation focuses on classification effectiveness and computational efficiency under identical experimental conditions, with the objective of identifying well-justified lightweight baseline model suitable for pre-symptomatic HLB detection.

3.1.1 Cross-Validation Performance

Table 7 summarizes the five-fold cross-validation results. Values are reported as mean ± SD with two-sided 95% t-confidence intervals. All classification metrics used binary averaging with HLB-symptomatic as the positive class, and the corrected F1-scores were derived from the corresponding precision and recall values.

Table 7: Five-fold cross-validation performance under the common training protocol.

Model	Accuracy, Mean ± SD [95% CI]	Precision, Mean ± SD [95% CI]	Recall, Mean ± SD [95% CI]	F1-Score, Mean ± SD [95% CI]	Specificity, Mean ± SD [95% CI]
MobileNetV2	80.50 ± 1.78 [78.28, 82.71]	79.80 ±2.25 [77.00, 82.59]	79.10 ± 0.70 [78.23, 79.97]	79.44 ± 1.38 [77.73, 81.15]	81.20 ± 1.82 [78.94, 83.46]
MobileNetV3	82.90 ± 1.23 [81.38, 84.42]	82.40 ± 1.96 [79.96, 84.83]	81.70 ± 2.07 [79.14, 84.26]	82.02 ± 1.23 [80.49, 83.55]	83.80 ± 1.67 [81.72, 85.87]
SqueezeNet	74.90 ± 2.22 [72.14, 77.65]	74.20 ± 3.17 [70.26, 78.14]	73.40 ± 2.93 [69.77, 77.04]	73.76 ± 2.34 [70.85, 76.67]	76.10 ± 4.78 [70.16, 82.04]
ShuffleNetV1	77.00 ± 2.35 [74.09, 79.91]	76.40 ± 1.21 [74.89, 77.90]	75.70 ± 2.65 [72.40, 78.99]	76.02 ± 1.29 [74.41, 77.63]	78.00 ± 2.45 [74.96, 81.05]
ShuffleNetV2	78.60 ± 2.91 [74.98, 82.21]	78.00 ± 2.80 [74.52, 81.48]	77.30 ± 2.58 [74.10, 80.50]	77.60 ± 1.68 [75.51, 79.69]	79.50 ± 1.03 [78.23, 80.78]
EfficientNetB5	81.70 ± 2.58 [78.50, 84.90]	81.10 ± 2.05 [78.55, 83.66]	80.50 ± 1.10 [79.14, 81.86]	80.78 ± 0.80 [79.79, 81.77]	82.60 ± 1.17 [81.15, 84.06]
GhostNet	79.30 ± 1.52 [77.41, 81.19]	78.80 ± 2.45 [75.76, 81.84]	78.10 ± 2.68 [74.77, 81.42]	78.43 ± 2.13 [75.78, 81.07]	80.20 ± 1.41 [78.45, 81.96]

3.1.2 Statistical Comparison

The Friedman test identified a statistically significant overall difference in cross-validation accuracy among the seven architectures (chi-square (6) = 21.429, p = 0.0015) as shown in Table 8. Kendall's W was 0.714, indicating a strong omnibus architecture effect. However, none of the six MobileNetV3 pairwise comparisons remained significant after Holm adjustment. Accordingly, MobileNetV3 had the highest mean performance under the common protocol, but the available five-fold evidence does not support a claim that it statistically outperformed any individual comparator. The pairwise findings should also be interpreted cautiously because only five paired fold observations were available.

Table 8: Friedman omnibus test and Holm-adjusted Wilcoxon comparisons on fold-level accuracy.

Comparison	Statistic	Raw *p*-Value	Holm-Adjusted *p*-Value	Interpretation
Friedman omnibus	Chi-square (6) = 21.429	0.0015	–	Significant overall architecture effect; Kendall's W = 0.714
MobileNetV3 vs. MobileNetV2	W = 1.0	0.1250	0.375	Not significant
MobileNetV3 vs. SqueezeNet	W = 0.0	0.0625	0.375	Not significant
MobileNetV3 vs. ShuffleNetV1	W = 0.0	0.0625	0.375	Not significant
MobileNetV3 vs. ShuffleNetV2	W = 0.0	0.0625	0.375	Not significant
MobileNetV3 vs. EfficientNetB5	W = 5.0	0.6250	0.625	Not significant
MobileNetV3 vs. GhostNet	W = 0.0	0.0625	0.375	Not significant

MobileNetV3 recorded the highest mean cross-validation accuracy (82.90 ± 1.23%; 95% CI: 81.38–84.42) and corrected F1-score (82.02 ± 1.23%; 95% CI: 80.49–83.55). EfficientNetB5 followed with 81.70 ± 2.58% accuracy and 80.78 ± 0.80% F1-score, while MobileNetV2 achieved 80.50 ± 1.78% accuracy and 79.44 ± 1.38% F1-score. SqueezeNet produced the lowest mean accuracy and F1-score, although the confidence intervals show substantial fold-level overlap among several models.

Fig. 2 displays the mean five-fold accuracy and corresponding 95% confidence intervals. The error bars reinforce that the observed ranking is descriptive and that uncertainty overlaps, particularly among MobileNetV3, EfficientNetB5, MobileNetV2, ShuffleNetV2, and GhostNet.

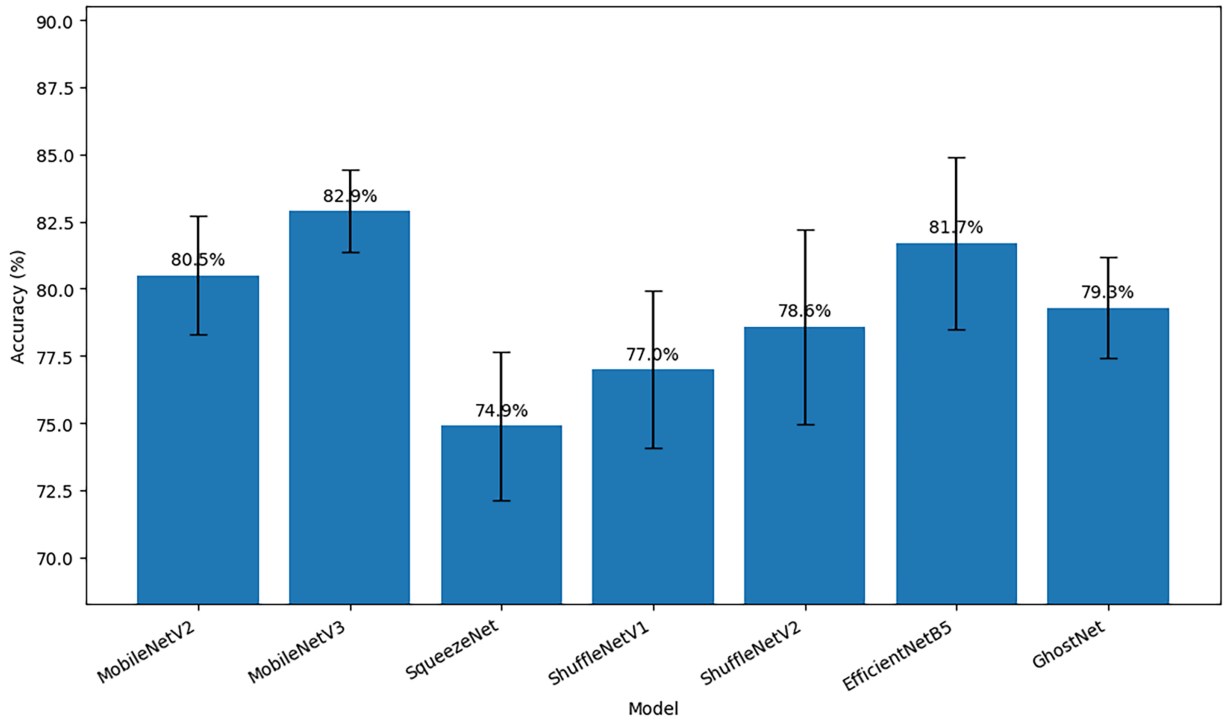


Figure 2: Mean five-fold cross-validation accuracy with 95% confidence intervals.

3.1.3 Independent Hold-Out Evaluation

As shown in Table 9, the independent hold-out results broadly reproduced the cross-validation ranking. MobileNetV3 achieved the highest hold-out accuracy (82.6%), recall (82.2%), F1-score (81.0%), and specificity (82.8%). EfficientNetB5 achieved 81.2% accuracy and 79.6% F1-score, while MobileNetV2 achieved 79.9% accuracy and 78.1% F1-score. The relative ordering of ShuffleNetV1 and ShuffleNetV2 changed slightly on the hold-out set, but both remained below the three leading architectures. These results corroborate MobileNetV3 as the strongest descriptive baseline under the common protocol; they do not alter the non-significant Holm-adjusted pairwise findings from cross-validation.

3.1.4 Deployment-Oriented Performance

Table 10 compares model complexity and predictive performance. FLOPs represent the approximate number of floating-point operations required for one inference, while MACs represent the multiply–accumulate operations performed during convolution and other network computations. Lower FLOPs and

MACs generally indicate reduced computational demand, which is important for deployment on resource-constrained devices. MobileNetV3 achieved the highest corrected cross-validation F1-score (82.02%) with moderate computational cost, whereas EfficientNetB5 required substantially more operations despite only a small improvement over some lighter models. SqueezeNet had the fewest parameters and the smallest model size but recorded the lowest F1-score. These results showed that parameter count, storage size and computational workload provide complementary measures of deployment suitability. FLOPs and MACs provide theoretical estimates of computational demand, whereas the latency, throughput and memory measurements in Table 11 show the actual runtime behaviour on the NVIDIA V100S.

Table 9: Performance on the independent stratified hold-out test set (n = 298).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Spec. (%)
MobileNetV2	79.9	77.0	79.3	78.1	80.4
MobileNetV3	82.6	79.9	82.2	81.0	82.8
SqueezeNet	74.2	70.4	74.1	72.2	74.2
ShuffleNetV1	76.5	73.0	76.3	74.6	76.7
ShuffleNetV2	78.2	75.0	77.8	76.4	78.5
EfficientNetB5	81.2	78.4	80.7	79.6	81.6
GhostNet	78.9	75.7	78.5	77.1	79.1

Table 10: Model complexity and corrected cross-validation F1-score.

Model	Parameters (M)	FP32 Size (MB)	FLOPs (G)	MACs (G)	F1-Score (%)
MobileNetV2	3.40	13.6	0.40	0.20	79.44
MobileNetV3	5.48	21.9	0.27	0.13	82.02
SqueezeNet	1.30	5.2	2.12	1.06	73.76
ShuffleNetV1	1.87	7.5	0.21	0.10	76.02
ShuffleNetV2	2.28	9.1	0.21	0.10	77.60
EfficientNetB5	30.39	121.6	10.39	5.19	80.78
GhostNet	5.18	20.7	0.19	0.09	78.43

Table 11: Runtime and peak-memory measurements on the NVIDIA V100S.

Model	Peak GPU Memory (MB)	Median Latency (ms)	P95 Latency (ms)	Throughput (img/s)
MobileNetV2	487	3.24	3.51	308.6
MobileNetV3	423	2.91	3.18	343.6
SqueezeNet	512	3.47	3.82	288.2
ShuffleNetV1	356	2.18	2.43	458.7
ShuffleNetV2	378	2.34	2.61	427.4
EfficientNetB5	1389	12.76	13.42	78.4
GhostNet	401	2.67	2.94	374.5

As shown in Table 10, MobileNetV3 combined the highest corrected cross-validation F1-score with a median V100S latency of 2.91 ms, a 95th-percentile latency of 3.18 ms, and peak allocated GPU memory of 423 MB. ShuffleNetV1 and ShuffleNetV2 were faster, with median latencies of 2.18 and 2.34 ms respectively, but their classification performance was lower. EfficientNetB5 had the greatest computational demand, with 30.39 million parameters, a 121.6 MB FP32 model size, 10.39 GFLOPs, 1389 MB peak GPU memory, and 12.76 ms median latency. SqueezeNet was the smallest model by parameter count and serialized size, but it did not provide the best latency or predictive performance. These results show that model size alone is insufficient for field-oriented architecture selection; predictive performance, latency, and memory must be considered jointly.

3.1.5 Training Dynamics of Lightweight Models

To analyze the learning behavior beyond final evaluation metrics, epoch-wise training accuracy, validation accuracy, training loss, and validation loss were examined for all seven models. All networks were trained using identical hyper-parameters under the same hardware and software conditions, ensuring that differences in convergence and generalization behavior reflected inherent architectural properties rather than training bias. As illustrated in Figs. 3 and 4, SqueezeNet exhibited early convergence with limited performance gains across epochs. Between epochs 20 and 80, both training and validation accuracy increased slowly and remained below 70%. Toward the final epochs, training accuracy increased to approximately 90%–92%, while validation accuracy saturated around 75%–77%. The widening gap between training and validation loss indicates limited representational capacity and pronounced overfitting, particularly in capturing subtle spectral variations associated with early HLB infection. ShuffleNetV1 showed faster initial learning, with training accuracy rising to approximately 95%–96%, while validation accuracy stabilized near 77%–78%. Validation loss remained consistently higher than training loss throughout training, indicating moderate overfitting and sensitivity to spectral variability. ShuffleNetV2 exhibited similar behavior, with training accuracy approaching 96%–97% and validation accuracy plateauing around 78%–79%. However, noticeable mid-epoch fluctuations in validation accuracy and loss suggested slower stabilization of generalization performance compared to MobileNet-based architectures.

As depicted in the results, MobileNetV2 demonstrated faster and more stable convergence than the ShuffleNet variants. Training accuracy rapidly approached 96%, while validation accuracy stabilized near 81% with a relatively small and consistent gap. Both training and validation losses decreased smoothly with minimal divergence, indicating stable generalization and effective learning of spectral features. On the other hand, MobileNetV3 achieved the strongest overall learning dynamics among the evaluated models. Training accuracy rapidly approached approximately 98%–99%, while validation accuracy increased smoothly to around 82%–83%. The gap between training and validation curves remained minimal, with only minor late-epoch divergence, reflecting strong representational capacity, efficient feature extraction, and stable generalization behavior.

GhostNet model exhibited steady convergence, with training accuracy reaching approximately 96% and validation accuracy plateauing around 79%–80%. Training and validation losses decreased gradually, with a small but persistent divergence, suggesting slightly slower adaptation to subtle spectral cues compared to MobileNet-based architectures. EfficientNetB5 demonstrated the fastest convergence and the most stable validation behavior among all evaluated models. Training accuracy rapidly approached 97%, while validation accuracy increased smoothly to approximately 81%–82%. Although convergence was slightly smoother in the early epochs compared to MobileNetV3, minimal oscillations in validation accuracy were observed. The gap between training and validation curves remained consistently small, and both loss curves decreased steadily, indicating robust generalization and effective learning of spectral features. MobileNetV3 exhibited

the most stable generalization behavior, highlighting the importance of architectural balance when learning from high-dimensional spectral data.

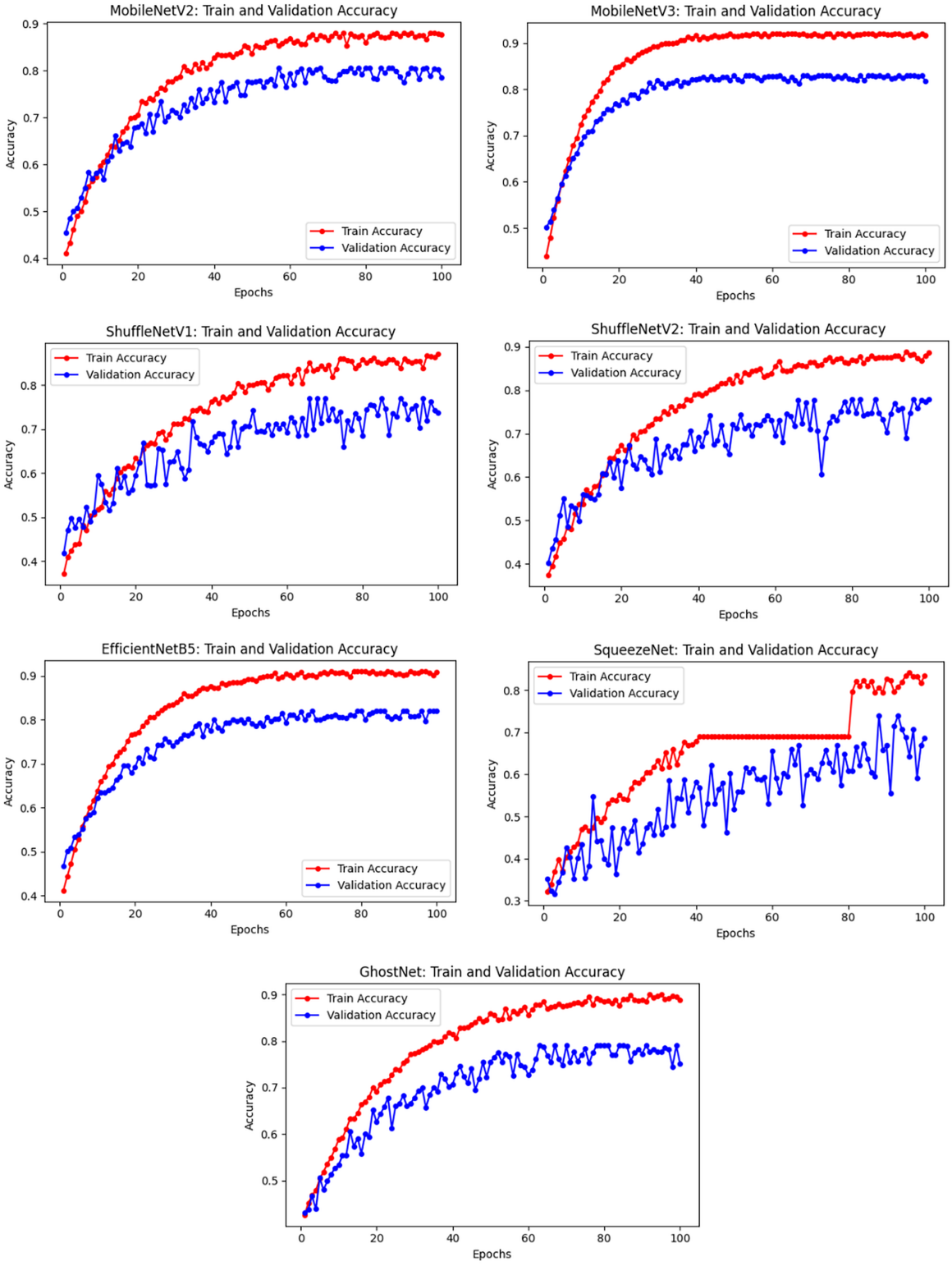


Figure 3: Training and validation accuracy curves for all models.

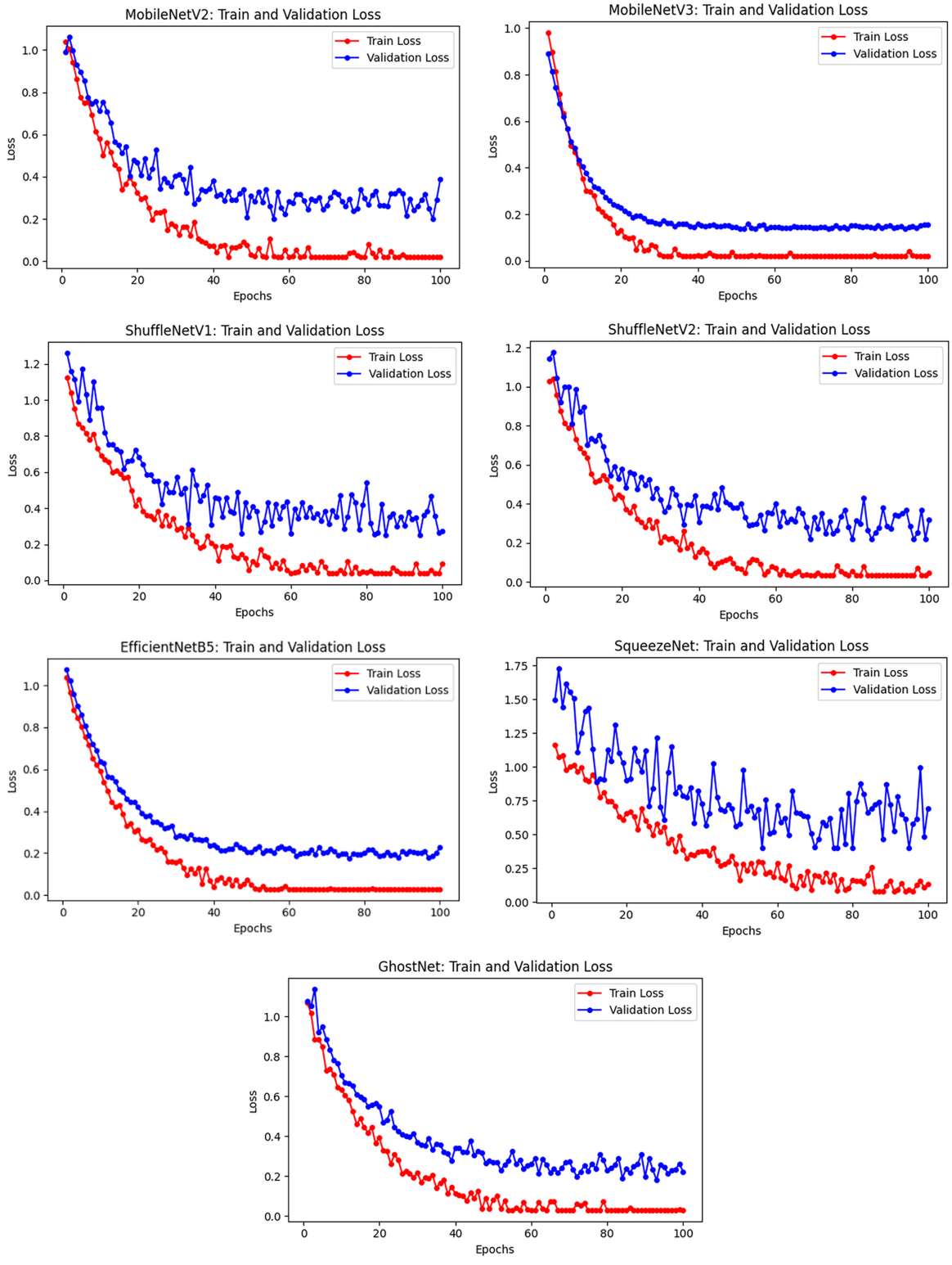


Figure 4: Training and validation loss curves for all models.

3.1.6 Confusion Matrix Analysis

Confusion matrices were examined to assess class-level behaviour on the independent hold-out set. False negatives represent HLB-symptomatic samples predicted as non-HLB-symptomatic and are particularly important because they may delay field intervention. The analysis is reported using the original binary dataset labels rather than relabelling the classes as healthy, early-stage, or pre-symptomatic.

Fig. 5 presents the hold-out confusion matrices for the seven models. MobileNetV3 had the highest hold-out recall (82.2%), indicating the lowest false-negative proportion among the evaluated architectures. EfficientNetB5 and MobileNetV2 also showed comparatively balanced discrimination, whereas SqueezeNet and the ShuffleNet variants produced lower recall and F1-scores. These errors are consistent with the difficulty of preserving weak spectral-spatial information in compact RGB-pretrained backbones, but the binary labels do not permit the errors to be interpreted as direct failures on a separately confirmed asymptomatic-infected class.

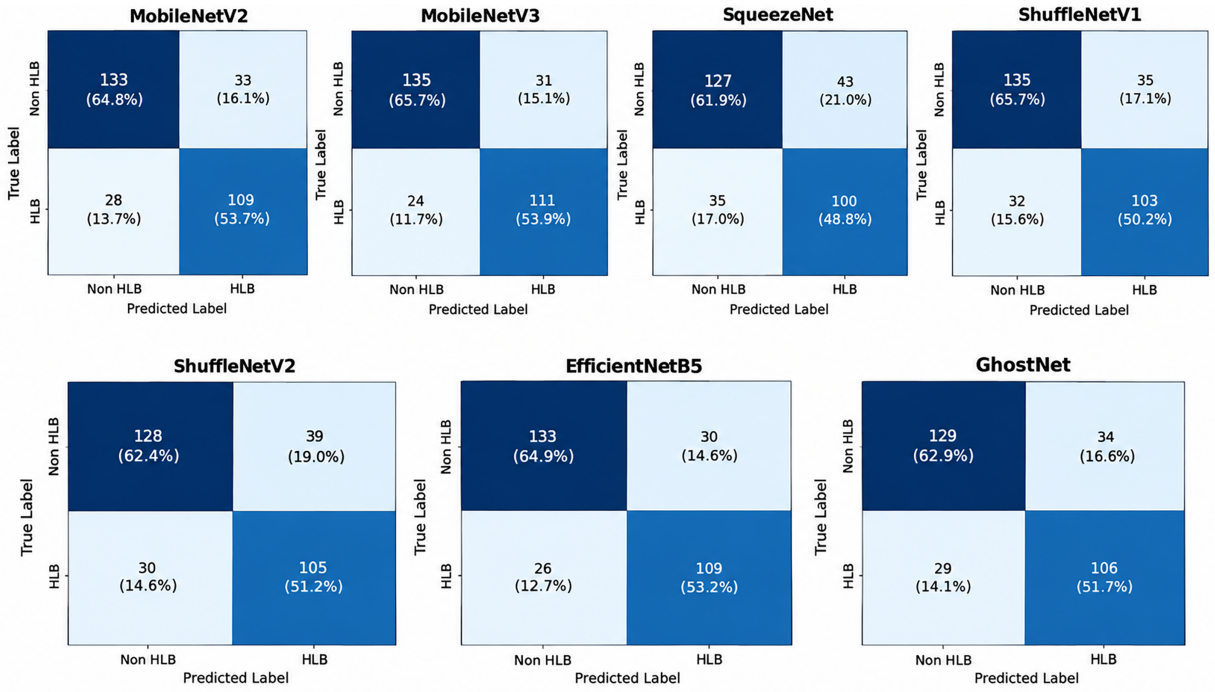


Figure 5: Confusion matrix for all models.

3.1.7 Summary of Experimental Evidence

Across cross-validation, independent hold-out evaluation, and deployment measurements, MobileNetV3 provided the strongest descriptive accuracy-efficiency baseline under the common training protocol. The significant Friedman result confirms that architecture choice affected fold-level accuracy, while the non-significant Holm-adjusted pairwise comparisons show that the available five-fold sample does not establish statistical superiority over any individual architecture. EfficientNetB5 was competitive but substantially more computationally demanding, and SqueezeNet and ShuffleNet variants achieved lower predictive performance. These findings support MobileNetV3 as the baseline for subsequent spectral-aware model development, subject to validation on independent orchards, sensors, campaigns, and laboratory-confirmed asymptomatic samples.

3.2 Discussion of Findings

The benchmark supports the SLR finding that lightweight architecture choice materially affects multispectral plant-disease classification under a controlled protocol. MobileNetV3 achieved the highest mean cross-validation accuracy and F1-score and retained the highest accuracy on the independent hold-out set. The Friedman test demonstrated a strong overall architecture effect; however, no MobileNetV3 pairwise comparison remained significant after Holm correction. The manuscript therefore interprets MobileNetV3 as the highest-mean and strongest descriptive common-protocol baseline, not as a statistically superior model.

The performance pattern is consistent with architectural differences in feature reuse, channel recalibration, and compression. MobileNetV3 and EfficientNetB5 achieved the strongest predictive results, whereas the more aggressively compressed SqueezeNet and ShuffleNet configurations generally produced lower recall and F1-scores. Nevertheless, the overlapping confidence intervals and the small number of paired folds limit strong architecture-level causal conclusions. In addition, the shared hyperparameters were selected to support a controlled comparison rather than to maximize the performance of every architecture; architecture-specific tuning could therefore change the exact ranking.

The study also demonstrates the limitations of transferring RGB-pretrained backbones to multispectral HLB data. Although the first convolutional layers were adapted to accept all 14 bands, ImageNet pretraining does not provide spectral priors for VisNIR relationships beyond RGB. The lower performance relative to some RGB-based studies may therefore reflect spectral domain shift, dataset difficulty, field variability, and model capacity rather than any single architectural weakness. The results motivate future investigation of spectral-aware pretraining and lightweight attention, but the present benchmark does not test a uniform added attention module and should not claim that attention caused the observed differences.

The early-detection relevance of the experiment arises from low-intensity HLB scenes and the use of spectral bands capable of capturing weak physiological variation. However, the source labels distinguish only HLB-symptomatic and non-HLB-symptomatic scenes. They do not identify a separate cohort of PCR-confirmed infected trees without visible symptoms. Consequently, the results establish an architecture-selection baseline for future pre-symptomatic research rather than clinical evidence of pre-symptomatic sensitivity.

3.3 Implications for Lightweight Spectral Model Design

The findings indicate that lightweight spectral models must balance computational reduction with preservation of narrow-band and cross-band information. Excessive compression may improve parameter count without improving latency and may reduce recall when disease-related cues are weak.

The SLR suggests that lightweight attention mechanisms may improve selective emphasis on informative spectral-spatial features, but their value should be tested through controlled ablation studies using the same backbone, split indices, and training budget. The present results provide the backbone baseline required for such experiments.

Future model development should combine spectral-aware pretraining, transparent band-processing procedures, architecture-specific sensitivity analysis, and deployment evaluation on representative edge hardware. External validation should prioritize campaign-held-out, orchard-held-out, sensor-held-out, and laboratory-confirmed asymptomatic cohorts.

3.4 Limitations and Generalizability

The empirical findings are based on one public citrus dataset acquired in a single production region with one imaging system and two field campaigns. Generalization to other citrus cultivars, orchards,

countries, seasons, sensors, and disease-prevalence conditions has therefore not been established. Because stable tree-, branch-, or leaf-level identifiers were unavailable, splitting was performed at image level; correlated samples from the same biological source could consequently occur across partitions and inflate performance estimates.

The binary source labels do not include a separately confirmed asymptomatic-infected class, so pre-symptomatic performance cannot be quantified directly. The use of a common hyperparameter protocol improves comparability but does not establish each architecture's optimal attainable performance. Statistical inference was based on only five paired folds, limiting the power of *post-hoc* tests, and deployment timing was measured on an NVIDIA V100S rather than on a representative field edge device. These constraints should be considered when interpreting the ranking and planning external validation.

4 Conclusion

This study combined a PRISMA-guided systematic review with a controlled multispectral benchmark of seven lightweight CNN architectures for HLB-symptomatic versus non-HLB-symptomatic classification. The review identified the dominant model families and methodological gaps, and those findings directly informed the architecture set, 14-band input protocol, development/hold-out partitioning, statistical analysis, and deployment measurements.

Under the shared training protocol, MobileNetV3 achieved the highest mean five-fold accuracy (82.90 ± 1.23%) and corrected F1-score (82.02 ± 1.23%), followed by EfficientNetB5. MobileNetV3 also achieved the highest hold-out accuracy (82.6%) and F1-score (81.0%). The Friedman test identified a significant overall architecture effect with a strong Kendall's W, but no MobileNetV3 pairwise comparison remained significant after Holm adjustment. MobileNetV3 should therefore be regarded as the strongest descriptive baseline and favourable accuracy-efficiency compromise, not as a statistically proven superior architecture.

The conclusions are limited by the use of a single dataset, image-level partitioning, five paired folds, V100S-based deployment measurements, and the absence of a separately labelled PCR-confirmed asymptomatic-infected class. The benchmark supports future development of a pre-symptomatic HLB system but does not itself demonstrate definitive pre-symptomatic detection. Subsequent work should evaluate spectral-aware pretraining and lightweight attention through controlled ablation, optimize architectures under equal search budgets, and validate performance across independent campaigns, orchards, sensors, edge devices, and laboratory-confirmed asymptomatic samples.

Acknowledgement: We extend our sincere gratitude to everyone who supported this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Geoffry Mutiso: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing the final manuscript. Geoffrey Mariga: conceptualization, methodology, investigation, validation, and reviewing and editing the final manuscript and editing. Aaron Oirere: supervision, validation, project administration, and reviewing and editing the final manuscript and editing. Geoffry Mutiso and Geoffrey Mariga independently conducted the title and abstract screening, full-text assessment, and quality appraisal of the studies included in the systematic review. Aaron Oirere adjudicated disagreements that could not be resolved through consensus. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The multispectral dataset analyzed in this study is openly available in Recherche Data Gouv, Version 1.0, under doi:10.57745/054NAB. The dataset documentation and supplementary processing scripts are also publicly available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Graham JH, Bassanezi RB, Dawson WO, Dantzler R. Management of huanglongbing of *Citrus*: lessons from São Paulo and Florida. Annu Rev Phytopathol. 2024;62(1):243–62. doi:10.1146/annurev-phyto-121423-041921.
2. Garcia-Ruiz F, Sankaran S, Maja JM, Lee WS, Rasmussen J, Ehsani R. Comparison of two aerial imaging platforms for identification of Huanglongbing-infected *Citrus* trees. Comput Electron Agric. 2013;91:106–15. doi:10.1016/j.compag.2012.12.002.
3. Dong R, Shiraiwa A, Ichinose K, Pawasut A, Sreechun K, Mensin S, et al. Hyperspectral imaging and machine learning for huanglongbing detection on leaf-symptoms. Plants. 2025;14(3):451. doi:10.3390/plants14030451.
4. Bera S, Varish N, Yaqoob SI, Rafi M, Shrivastava VK. Deep hierarchical spectral-spatial feature fusion for hyperspectral image classification based on convolutional neural network. Intell Data Anal. 2025;29(2):385–407. doi:10.3233/ida-230927.
5. Zhang A, Liu W. Plant disease detection algorithm based on efficient swin transformer. Comput Mater Contin. 2025;82(2):3045–68. doi:10.32604/cmc.2024.058640.
6. Lu J, Shi R, Tong J, Cheng W, Ma X, Liu X. Lightweight method for plant disease identification using deep learning. Intell Autom Soft Comput. 2023;37(1):525–44. doi:10.32604/iasc.2023.038287.
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
8. Norris JM, Simpson BS, Ball R, Freeman A, Kirkham A, Parry MA, et al. A modified Newcastle-Ottawa scale for assessment of study quality in genetic urological research. Eur Urol. 2021;79(3):325–6. doi:10.1016/j.eururo.2020.12.017.
9. Neware S. Design of automated model for *Citrus* fruits and leaves disease detection. J Inf Syst Eng Manage. 2025;10(37s):269–78. doi:10.52783/jisem.v10i37s.6417.
10. Bui TD, Do Le TM. Ghost-attention-YOLOv8: enhancing rice leaf disease detection with lightweight feature extraction and advanced attention mechanisms. AgriEngineering. 2025;7(4):93. doi:10.3390/agriengineering7040093.
11. Iqbal MI, Avianto D. Squeeze-and-excitation networks and attention mechanism in automatic detection of coffee leaf diseases based on images. J Soft Comput Explor. 2024;5(4):320–31. doi:10.52465/joscex.v5i4.490.
12. Zia Ur Rehman M, Ahmed F, Attique Khan M, Tariq U, Shaukat Jamal S, Ahmad J, et al. Classification of *Citrus* plant diseases using deep transfer learning. Comput Mater Contin. 2022;70(1):1401–17. doi:10.32604/cmc.2022.019046.
13. Faisal S, Javed K, Ali S, Alasiry A, Marzougui M, Attique Khan M, et al. Deep transfer learning based detection and classification of *Citrus* plant diseases. Comput Mater Contin. 2023;76(1):895–914. doi:10.32604/cmc.2023.039781.
14. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6848–56. doi:10.1109/CVPR.2018.00716.
15. Entuni CJ, Zulcaffle TMA, Ping KH, Sharangi AB, Ling WV, Tan LW. Performance evaluation of an enhanced shufflenet CNN for multi-crop leaf disease classification using fine-tuned parameters. Int J Sci Res Arch. 2025;16(1):1960–6. doi:10.30574/ijsra.2025.16.1.2242.
16. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Computer vision—ECCV 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 122–38. doi:10.1007/978-3-030-01264-9_8.
17. He H, Ning C, Liu M, Zhu J. Tomato disease degree recognition based on RGB and Lab color space conversion method. In: Proceedings of the 2023 IEEE 6th International Conference on Industrial Cyber-Physical Systems (ICPS); 2023 May 8–11; Wuhan, China. p. 1–6. doi:10.1109/ICPS58381.2023.10128053.

18. Zhang H, Zhu X, Li B, Guan Z, Che W. LA-ShuffleNet: a strong convolutional neural network for edge computing devices. IEEE Access. 2023;11:116684–94. doi:10.1109/ACCESS.2023.3324713.
19. Xie M, Wu J, Sun J, Xiao L, Liu Z, Yuan R, et al. MFFSNet: a lightweight multi-scale shuffle CNN network for wheat disease identification in complex contexts. Agronomy. 2025;15(4):910. doi:10.3390/agronomy15040910.
20. Xing S, Lee M. Classification accuracy improvement for small-size *Citrus* pests and diseases using bridge connections in deep neural networks. Sensors. 2020;20(17):4992. doi:10.3390/s20174992.
21. Khan E, Rehman MZU, Ahmed F, Khan MA. Classification of diseases in *Citrus* fruits using SqueezeNet. In: Proceedings of the 2021 International Conference on Applied and Engineering Mathematics (ICAEM); 2021 Aug 30–31; Taxila, Pakistan. p. 67–72. doi:10.1109/ICAEM53552.2021.9547133.
22. Challagundla Y, Chintalapati LR, Tunuguntla TSC, Sur A, Roy B, Zhuo ER. Screening of *Citrus* diseases using deep learning embedders and machine learning techniques. In: Proceedings of the 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP); 2023 Mar 18–20; Vijayawada, India. p. 1–5. doi:10.1109/AISP57993.2023.10134971.
23. Parvathi SD, Krishnan N, Nabeela M, Karputha P. Deep learning-driven orange fruit disease identification using SqueezeNet feature extraction. In: Proceedings of the International Conference on Modern Trends in Engineering and Management (ICMTEM-25); 2025 Nov 28–30; Shenyang, China.
24. Kaushik P, Sharma P. Deep learning-based detection of orange diseases using MobileNetV2 for enhanced agricultural diagnostics. In: Proceedings of the 2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART); 2024 Dec 6–7; Moradabad, India. p. 28–33. doi:10.1109/SMART63812.2024.10882535.
25. He C, Li X, Liu Y, Yang B, Wu Z, Tan S, et al. Combining multicolor fluorescence imaging with multispectral reflectance imaging for rapid *Citrus* Huanglongbing detection based on lightweight convolutional neural network using a handheld device. Comput Electron Agric. 2022;194:106808. doi:10.1016/j.compag.2022.106808.
26. Huong PT, Hien LT, Son NM, Tuan HC, Nguyen TQ. Enhancing deep convolutional neural network models for orange quality classification using MobileNetV2 and data augmentation techniques. J Algoritms Comput Technol. 2025;19:17483026241309070. doi:10.1177/17483026241309070.
27. Bhoomika, Singh AR, Aluvala S. Enhancing orange crop health: disease detection using MobileNetV2 model. In: Proceedings of the 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0; 2025 Apr 9–11; Raigarh, India. p. 1–5. doi:10.1109/OTCON65728.2025.11070828.
28. Hidayah N, Muiz A, Sunandar D. Web-based *Citrus* fruit disease detection application using MobileNet V2 for agricultural quality assurance. Bit-Tech. 2025;8(2):2663–72. doi:10.32877/bt.v8i2.3338.
29. Reddy KRG, Thirunavukkarasu KS. IoT-enabled MobileNetV3-tiny with CBAM for real-time sweet lemon leaf disease detection. Int J Res Publ Rev. 2025;6(9):1223–33. doi:10.55248/gengpi.6.0925.3252.
30. Wang K. Application of improved MobileNetV3 in agricultural disease and pest recognition. In: Proceedings of the International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023); 2023 Dec 1–3; Nanchang, China.
31. Wang L, Shi W, Tang Y, Liu Z, He X, Xiao H, et al. Transfer learning-based lightweight SSD model for detection of pests in *Citrus*. Agronomy. 2023;13(7):1710. doi:10.3390/agronomy13071710.
32. Tang W, Shi Y, Chen H, Liu H, Zhong L, Chen W. A small neural network deployed on edge devices for detecting citrus Huanglongbing [Internet]. Res Square. 2024 [cited 2026 Jul 28]. Available from: https://www.researchsquare.com/article/rs-3999354/v1.
33. Hoang VT, Jo KH. Practical analysis on architecture of EfficientNet. In: Proceedings of the 2021 14th International Conference on Human System Interaction (HSI); 2021 Jul 8–10; Gdańsk, Poland. p. 1–4. doi:10.1109/HSI52170.2021.9538782.
34. Agrawal PS, Dhakate KM, Parthani KM, Agnihotri AS. EfficientNet B0 CNNs for orange huanglongbing and tomato pathogens. In: Innovations in cybersecurity and data science. Singapore: Springer Nature; 2024. p. 409–20. doi:10.1007/978-981-97-5791-6_31.
35. Goyal A, Lakhwani K. Integrating advanced deep learning techniques for enhanced detection and classification of *Citrus* leaf and fruit diseases. Sci Rep. 2025;15(1):12659. doi:10.1038/s41598-025-97159-0.

36. Upadhyay D, Manwal M, Kukreja V, Sharma R. Advancing *Citrus* disease diagnosis: application of EfficientNetB3 for precise classification of orange tree pathologies. In: Proceedings of the 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS); 2024 Apr 22–23; Bengaluru, India. p. 1–4. doi:10.1109/ICETCS61022.2024.10543447.
37. Kaur A, Kukreja V, Thapliyal N, Aeri M, Sharma R, Hariharan S. Pre-trained VGG16 and EfficientNetB3 for multiclass classification of orange leaf diseases. In: Proceedings of the 2024 3rd International Conference for Innovation in Technology (INOCON); 2024 Mar 1–3; Bangalore, India. p. 1–6. doi:10.1109/INOCON60754.2024.10511701.
38. Dai Q, Guo Y, Li Z, Song S, Lyu S, Sun D, et al. *Citrus* disease image generation and classification based on improved FastGAN and EfficientNet-B5. Agronomy. 2023;13(4):988. doi:10.3390/agronomy13040988.
39. Xiao Y, Wu J, Ma C. GA-GhostNet: a lightweight CNN model for identifying pests and diseases using a gated multi-scale coordinate attention mechanism. Eng Lett. 2024;32(7):1281–90. doi:10.1145/3633637.3633682.
40. Chin PW, Ng KW, Palanichamy N. Plant disease detection and classification using deep learning methods: a comparison study. J Inform Web Eng. 2024;3(1):155. doi:10.33093/jiwe.2024.3.1.10.
41. Xu L, Wang Y, Shi X, Tang Z, Chen X, Wang Y, et al. Real-time and accurate detection of *Citrus* in complex scenes based on HPL-YOLOv4. Comput Electron Agric. 2023;205:107590. doi:10.1016/j.compag.2022.107590.
42. Abdelghafour F. Multispectral images from the orchard, detecting HLB citrus disease [Internet]. Recherche Data Gouv; 2025 [cited 2026 Mar 24]. Available from: https://entrepot.recherche.data.gouv.fr/citation?persistentId=doi:10.57745/054NAB.
43. Porto LR, Abdelghafour F, Tommaselli AMG, Imai NN. Field-based and close-range multispectral imaging dataset for Huanglongbing (HLB) detection in orange trees: a resource for machine learning and digital agriculture. Data Brief. 2026;67:112899. doi:10.1016/j.dib.2026.112899.
44. Bhatia A, Chug A, Singh AP. Statistical analysis of machine learning techniques for predicting powdery mildew disease in tomato plants. Int J Intell Eng Inform. 2021;9(1):24. doi:10.1504/ijiei.2021.116087.
45. El-Kenawy EM, Ali Alhussan A, Khafaga DS, Abotaleb M, Mishra P, Arnous R, et al. Optimizing potato disease classification using a Metaheuristics algorithm for deep learning: a novel approach for sustainable agriculture. Potato Res. 2025;68(1):551–85. doi:10.1007/s11540-024-09755-8.
46. Bhavani GD, Chalapathi MMV. PotatoLeafNet: two-stage convolutional neural networks for effective Potato Leaf disease identification and classification. Front Artif Intell. 2025;8:1668839. doi:10.3389/frai.2025.1668839.