

Benchmarking Convolutional Neural Networks and Vision Transformers for In-Field Crop Disease Diagnosis on Resource-Constrained Mobile Devices

Abstract—Accurate and timely identification of crop leaf diseases is critical to safeguarding global food security and mitigating agricultural economic losses. While deep learning has significantly advanced automated foliar diagnosis, practical field deployment on low-cost smartphones—such as those widely used by smallholder farmers in developing agricultural regions—remains hindered by severe computational constraints, uncontrolled illumination, complex backgrounds, and limited labeled data. This paper presents a structured review and comparative analysis of convolutional neural networks (CNNs), pure vision transformers (ViTs), and emerging hybrid architectures for in-field crop leaf disease detection. We systematically examine trade-offs across parameter footprint, computational latency, representation learning, and data efficiency. Traditional CNNs offer lightweight spatial inductive biases and low memory overheads, making them highly amenable to edge quantization and on-device deployment. Conversely, ViTs excel at modeling long-range contextual relationships and maintaining robustness against background clutter, yet they demand extensive pre-training and present high computational complexities that challenge mobile hardware. Emerging hybrid networks synthesize local convolutional feature extraction with global self-attention mechanisms, delivering superior diagnostic resilience under non-uniform field lighting and multi-organ occlusion. Drawing on recent empirical evidence across diverse agricultural datasets, we identify critical performance bottlenecks and outline optimization paradigms—including one-shot learning, structural pruning, and edge quantization—essential for transitioning automated disease diagnosis from laboratory benchmarks to accessible, offline smartphone applications for agricultural extension workers and resource-constrained farming communities.

Index Terms—crop disease diagnosis, convolutional neural networks, vision transformers, edge intelligence, mobile agricultural deployment, field imaging robustness

I. INTRODUCTION

Sustaining global agricultural productivity is an imperative challenged by escalating food demands, climate instability, and pervasive plant pathogen outbreaks. Projections by international bodies indicate that global food output must increase by approximately 70% by 2050 to sustain the expanding human population, even as roughly one-third of potential agricultural yields continue to be lost to plant diseases and pest infestations [1]. Foliar infections damage vital plant organs, severely degrading yield quality and inflicting substantial economic losses across cereal, vegetable, and cash-crop systems [2], [3], [4]. In smallholder-dominated agricultural economies such as India, timely and precise pathological identification remains paramount [5]. Conventional diagnostic paradigms rely heavily on manual inspection and advisory visits by botanical experts,

processes that are inherently laborious, expensive, subjective, and difficult to scale across millions of fragmented farming plots [3], [6]. Consequently, automating leaf disease detection through computer vision deployed on handheld mobile infrastructure represents an essential pathway for empowering rural smallholders and agricultural extension personnel [1], [7].

Recent breakthroughs in deep learning have accelerated the adoption of automated image-based diagnostics in precision agriculture [6], [5]. Early innovations centered on standard convolutional neural networks (CNNs), which leverage localized receptive fields and shared kernel weights to capture fine-grained spatial lesion patterns [4], [5], [8]. To reconcile these architectures with mobile hardware, researchers engineered compact convolutional variants, such as custom shallow networks and parameter-efficient backbones tailored for real-time edge processing [9], [10], [8]. More recently, the computer vision paradigm has expanded toward self-attention mechanisms and vision transformers (ViTs) [11], [12]. Transformers eliminate the rigid local inductive biases of convolutions, modeling long-range spatial dependencies across fragmented image patches [12], [2]. Furthermore, emerging architectural designs combine convolutional layers with self-attention modules into hybrid CNN-ViT pipelines, aiming to capture subtle localized lesion symptoms while concurrently encoding broad canopy-level context [11], [2], [7].

Despite these computational advancements, a pronounced operational gap persists between laboratory benchmark success and practical field execution [13], [14]. The vast majority of published disease-classification models are trained and validated on sanitized datasets captured under uniform artificial lighting against monochromatic backgrounds [13], [14]. When deployed in authentic agricultural environments, these systems confront severe performance degradation driven by non-uniform natural illumination, harsh shadows, intricate soil-and-weed foliage clutter, and dynamic camera angles [1], [13], [14]. This domain shift is exacerbated on low-cost Android smartphones commonly accessible to smallholder farmers in developing nations. Such devices operate with uncalibrated, noisy optical sensors, restricted random-access memory (RAM), limited battery capacities, and low-tier system-on-chip (SoC) processors [15], [9], [16]. Concurrently, smallholders and local extension services rarely possess large, curated, and expertly annotated localized disease datasets, creating a critical demand for data-efficient architectures capable of generalizing from small sample sizes [17], [18].

This divergence raises a fundamental research question: how do the structural design principles of CNNs, vision transformers, and hybrid networks balance the competing tensions of feature extraction fidelity under severe lighting variations, generalization under acute data scarcity, and computational feasibility on hardware-constrained smartphones? While lightweight CNNs offer inherent computational parsimony and low operational latency [9], [10], pure vision transformers demonstrate robust context parsing yet present formidable memory and training-data hurdles [11], [12]. Meanwhile, hybrid CNN-ViT configurations offer an appealing compromise, though their complex topological interactions often resist straightforward edge compilation and integer quantization [12], [2], [19]. A rigorous cross-architectural investigation is therefore essential to bridge the gap between abstract algorithmic performance and tangible agricultural field utility.

To address this challenge, this review delivers a comprehensive comparative analysis of deep visual diagnostic architectures tailored specifically to the operational realities of smallholder farming. The primary contributions of this work are threefold:

- We formulate a taxonomy categorizing standard CNNs, lightweight convolutional variants, pure vision transformers, and hybrid CNN-ViT models, evaluating their foundational mechanics regarding receptive fields, attention structures, and computational scaling [11], [12], [9], [10].
- We synthesize evidence on model sensitivity to challenging field conditions, including varying natural solar illumination, uncalibrated handheld smartphone lenses, complex soil and vegetative backgrounds, and sample-starved diagnostic classes [13], [15], [17], [14].
- We benchmark edge-deployment parameters—such as parameter counts, floating-point operations (FLOPs), structural pruning, and INT8 post-training quantization—to evaluate real-time, offline diagnostic execution on low-cost mobile hardware [12], [19], [16], [20].

The remainder of this paper is structured as follows. Section II reviews related literature, tracing the progression from classical CNNs to attention-based and hybrid architectural paradigms. Section III establishes the methodological framework, categorizing model families and outlining the evaluation dimensions: architectural complexity, edge deployment criteria, and environmental domain shifts. Section IV synthesizes comparative findings concerning feature discrimination, robustness to environmental variations, small-sample adaptability, and hardware metrics. Section V discusses practical design trade-offs, deployment strategies on budget mobile devices, and pathways toward robust uncertainty estimation in field tools. Finally, Section VI summarizes key architectural recommendations for agricultural engineering and extension applications.

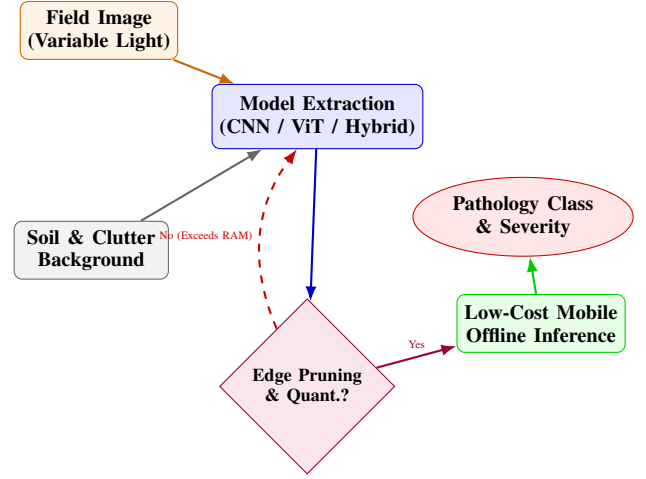


Fig. 1. Conceptual pipeline of edge-oriented leaf disease diagnosis under field constraints

II. RELATED WORK

A. Evolution of Lightweight CNN Architectures for Edge Deployment

Automated foliar pathology traditionally relied on standard convolutional neural networks (CNNs) optimized for laboratory benchmarks [3], [4]. However, translating these networks to mobile agricultural interfaces has driven substantial research toward parameter-efficient convolutional operators. To overcome the computational barriers that hinder standard deep networks from operating on portable, low-power hardware, specialized architectures have emerged. For instance, GrapeLeafNet [9] introduces a streamlined topological structure engineered to minimize floating-point operations while preserving diagnostic discriminability in viticulture. Similarly, CNN13 [8] employs a custom thirteen-layer structure with repeated pooling and convolutional blocks to classify citrus leaf conditions directly from unconstrained smartphone captures. In tomato disease management, the LightMixer network [10] demonstrates that lightweight convolutional designs can effectively suppress extraneous background noise while capturing primary lesion structures.

Concurrently, one-stage edge detection pipelines, such as MGA-YOLO [21] for apple foliar damage, YOLO-PLNet [19] for peanut leaves, and KFP-YOLO [16] for fruit and foliage anomalies, have embedded multi-scale convolutional attention to balance compact memory footprints with real-time inference constraints. While these specialized CNN frameworks attain high throughput and modest memory footprints suitable for edge deployment [19], [9], [16], their architectural reliance on rigid, localized receptive fields restricts their capacity to model long-range contextual relationships. Consequently, pure convolutional designs frequently degrade when confronting non-uniform shadows, complex soil-canopy boundaries, and widespread multi-target occlusions commonly encountered in field-scale photography [13], [22].

B. Vision Transformers and Global Attention in Plant Pathology

To address the localized perceptual constraints of convolutional kernels, attention-based models and vision transformers (ViTs) have been adapted for agricultural visual recognition. By utilizing multi-head self-attention mechanisms, ViTs process inputs as sequences of distinct visual tokens, enabling the network to establish long-range spatial dependencies across entire leaf surfaces. Recognizing that standard transformer backbones demand prohibitive computational and memory resources, recent efforts have focused on structural compression for mobile integration. PMVT [12] exemplifies this direction by introducing an efficient mobile vision transformer tailored specifically for plant disease identification under resource-constrained execution environments. In parallel, real-time transformer detectors, such as WMC-RTDETR [22], have demonstrated marked efficacy in handling complex tea plantation backgrounds and multi-target occlusions by decoupling cross-scale feature interactions through transformer encoder-decoder paradigms.

Despite their demonstrated capacity to decouple disease lesions from distracting background clutter, pure transformer designs present severe operational trade-offs. The absence of inductive spatial biases—such as translation equivariance and local pixel correlation—renders ViTs heavily reliant on large-scale annotated pre-training corpora [12], [23]. When deployed across domain-specific agronomic regimes characterized by small, heterogeneous sample sets, pure ViTs frequently exhibit slow convergence rates, high parameter counts, and excessive memory transfer overheads that exceed the computational boundaries of entry-level mobile processor units [12], [19].

C. Emergence of Hybrid CNN-Transformer Topologies

The structural dichotomy between localized convolutional filtering and global token self-attention has motivated the synthesis of hybrid CNN-ViT architectures. By cascading convolutional layers with self-attention stages or fusing them via dual-branch modules, hybrid networks seek to capture fine-grained textural anomalies—such as minute necrotic spots and powdery mildews—while concurrently aggregating panoramic leaf contextual signals [11], [2]. In multispectral plant pathology, hybrid formulations leveraging convolutional front-ends with transformer-based back-ends have shown strong feature representation capabilities across complex multi-crop datasets [11]. Chen et al. [2] established that integrating convolutional layers with transformer modules provides enhanced diagnostic accuracy for tomato leaf diseases compared to standalone baselines, effectively disambiguating subtle morphological lesion variations.

Beyond standard dual-stream topologies, hybrid paradigms have been augmented with advanced routing mechanisms; for example, pulmonary pathology networks incorporating gated multi-scale fusion [23] confirm that explicit feature gating effectively mitigates the transformer’s inherent vulnerability to high-frequency textural loss. In specialized agricultural sectors, such as coffee cultivation, hybrid CNN-ViT models

coupled with counter-factual frameworks have demonstrated robust diagnostic resilience against intertwined physiological disorders [7]. Nevertheless, existing literature on hybrid architectures primarily highlights predictive gains on high-end hardware, frequently omitting rigorous evaluations of memory bandwidth bottlenecks, integer quantization compatibility, and latency degradation on constrained mobile chipsets.

D. Data-Efficient Learning and Field Generalization Challenges

A pervasive limitation across both convolutional and transformer-based agricultural diagnostic systems is their vulnerability to the substantial domain shifts between controlled laboratory benchmarks and unconstrained field conditions [1], [13]. Controlled glasshouse acquisitions typically present homogeneous backgrounds and stabilized illumination; conversely, in-situ field imagery involves non-uniform solar glare, complex weed-canopy interference, variable leaf orientations, and uncalibrated mobile optical sensors [13], [15]. Investigations into grape disease classification under authentic vineyard conditions highlight that predictive confidence deteriorates noticeably under intense ambient illumination variations and unstructured canopy backgrounds, underscoring the necessity of uncertainty-aware frameworks [14]. Similar environmental and severity distribution shifts have been documented in hierarchical coffee disease assessments [24] and multi-crop diagnostic trials encompassing maize and solanaceous crops [5].

To alleviate chronic training sample scarcity in specialized agricultural domains, data-efficient and parameter-sharing paradigms have emerged. Saad and Salman [17] investigated one-shot learning strategies utilizing field captures, demonstrating that metric-based embedding spaces can classify novel pathology profiles without large-scale retraining. Complementary, multi-model ensemble convolutional frameworks have been explored by Sahu et al. [18] to enhance diagnostic generalization across varied bean foliar diseases. Furthermore, edge-oriented deployment protocols on constrained embedded platforms—such as plug-tray monitoring systems [20]—emphasize that structural optimizations must account directly for the execution constraints of mobile edge hardware. Unlike prior studies that report accuracy metrics within isolated computational settings, this paper systematically contextualizes lightweight CNNs, pure ViTs, and hybrid architectures against the operational realities of field illumination shifts, dataset constraints, and the hardware limits of low-cost Android smartphones.

III. METHODOLOGY

A. Structured Review Scope and Evaluation Dimensions

To evaluate the feasibility of deploying deep neural networks on resource-constrained Android smartphones in rural farming contexts, we establish a multi-dimensional assessment matrix grounded in the operational demands of agricultural extension services. Field-captured leaf imagery deviates substantially from uniform benchmark imagery, exhibiting

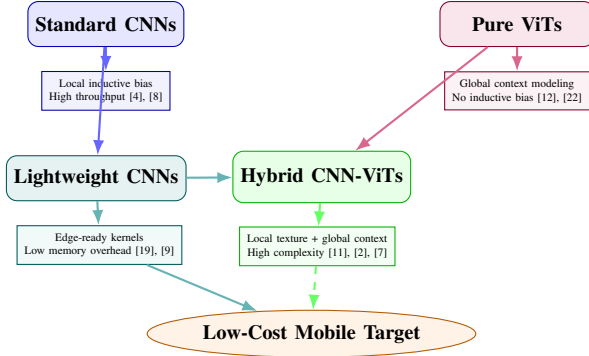


Fig. 2. Taxonomic progression of deep learning architectures applied to mobile crop disease detection

harsh shadows, non-uniform solar exposure, complex soil and weed backgrounds, and variable focal distances [1], [13], [14]. Consequently, models cannot be judged solely by top-1 classification accuracy. Our comparative framework assesses architectures across four interdependent axes: structural parameter complexity and floating-point operations (FLOPs), diagnostic resilience to lighting and environmental clutter, data efficiency under sample-scarce agricultural regimes, and operational edge deployability on entry-level mobile system-on-chip (SoC) architectures [12], [19], [17], [16].

B. Taxonomy of Deep Foliar Diagnostic Architectures

The surveyed models are categorized into three principal architectural paradigms: pure convolutional neural networks (CNNs), pure vision transformers (ViTs), and hybrid CNN-ViT frameworks. Standard and lightweight CNNs leverage translation invariance and localized receptive fields through stacked spatial convolutions, depthwise separable convolutions, and inverted residual bottlenecks [9], [10], [8]. These inductive biases allow CNNs to rapidly learn fine-grained lesion textures—such as fungal pustules, necrotic spots, and chlorotic halos—with modest parameter footprints and minimal operational latency [4], [5], [18].

In contrast, pure vision transformers discard localized inductive biases in favor of global multi-head self-attention mechanisms [12], [2]. An input leaf image $X \in \mathbb{R}^{H \times W \times C}$ is partitioned into non-overlapping patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, linearly projected to dimension D , and augmented with learnable positional embeddings before entering alternating layers of multi-head self-attention (MHSA) and multi-layer perceptron (MLP) blocks [11], [2]:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

While MHSA empowers pure ViTs to model long-range spatial dependencies across the entire foliar canopy, it scales quadratically with token sequence length, inducing substantial computational overhead and poor sample efficiency when trained on small agricultural datasets [11], [12], [23].

To resolve these architectural tensions, hybrid CNN-ViT designs integrate convolutional feature extractors into early

stages to capture high-frequency textural cues, subsequently passing downsampled token representations into lightweight transformer layers to model contextual relationships across leaves and background elements [11], [2], [7], [23]. Furthermore, lightweight object detection frameworks derived from single-stage architectures incorporate optimized attention backbones to simultaneously classify and localize multi-scale foliar lesions under occlusion [19], [22], [16], [21].

C. Field Heterogeneity and Data Efficiency Metrics

A foundational challenge in foliar pathology is the persistent domain shift between artificial laboratory benchmarks—characterized by excised leaves against monochrome backgrounds—and unconstrained field photography [1], [13], [6], [24]. In operational field settings, models must withstand acute specular highlights, partial shadowing under variable solar angles, dynamic motion blur, and structural occlusion from surrounding non-target foliage or soil [13], [22], [14], [8]. Moreover, acquiring exhaustive, expert-annotated pathological datasets for regional crop varieties across developing farming ecosystems is frequently cost-prohibitive [1], [17], [5].

Our analytical protocol assesses candidate frameworks against dataset scarcity and distributional shift. We contrast models pre-trained on generic image corpuses and fine-tuned on limited crop images against specialized data-efficient methodologies, including few-shot and one-shot distance-metric approaches designed to generalize from sparse exemplars [17], [18]. Diagnostic reliability under field uncertainty is further calibrated by examining uncertainty-aware estimation techniques and confidence thresholding mechanisms that flag out-of-distribution or ambiguous field samples [15], [14].

D. Edge-Readiness Profiling and Hardware Constraints

To ground architectural utility in the economic reality of smallholder agriculture in regions such as rural India, we define an edge-readiness pipeline that profiles deep models under the hardware limitations of low-cost Android smartphones [19], [9], [16]. Target devices typically possess restricted random-access memory (frequently limited to 2 GB to 4 GB shared memory), modest octa-core ARM architectures without high-end neural processing units (NPUs), and severe thermal throttling constraints under sustained field usage [12], [19], [20].

The optimization workflow mandates post-training structural pruning and 8-bit integer (INT8) uniform quantization to diminish disk storage footprints and runtime memory access latency [19], [25], [16]. Models are converted into mobile-optimized inference formats compatible with lightweight runtime frameworks [12], [20]. Under these deployment assumptions, we establish evaluation criteria prioritizing architectures that exhibit parameter counts below 10 million parameters, operational sizes below 50 megabytes post-compression, and single-frame inference latencies under 200 milliseconds on ARM-based mobile CPUs, ensuring real-time feedback for uncalibrated smartphone captures [12], [19], [15], [9].

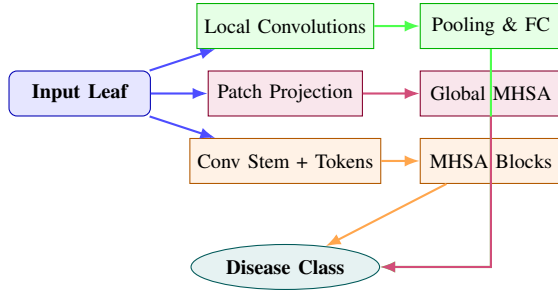


Fig. 3. Comparative structural workflow of CNN, pure ViT, and hybrid CNN-ViT foliar diagnosis pipelines

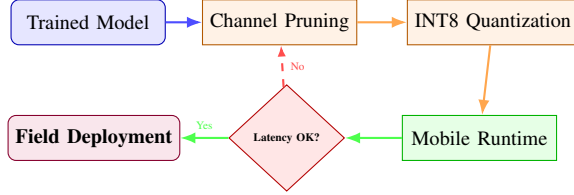


Fig. 4. End-to-end edge deployment pipeline from model optimization to uncalibrated field inference

IV. RESULTS

A. Comparative Architectural Efficacy and Feature Discrimination

Empirical investigations across the reviewed literature demonstrate marked differences in how convolutional networks, vision transformers, and hybrid topologies extract discriminative pathological features from foliar imagery. Standard and lightweight CNNs consistently excel at isolating fine-grained surface anomalies—such as early-stage chlorotic spotting, necrotic lesions, and rust pustules—due to the strong local inductive biases inherent in spatial convolutions [4], [9], [5], [8]. Specialized architectures such as GrapeLeafNet [9] and the 13-layer CNN13 [8] demonstrate that compact convolutional layers with repeated pooling mechanisms can effectively distinguish subtle foliar damage across crops including grape, citrus, maize, and potato [9], [5], [8]. However, prior studies report that purely convolutional representations tend to overfit to localized high-frequency artifacts, rendering them vulnerable when lesion margins faded into chlorotic tissue or when leaf textures are disrupted by non-target visual noise [2], [10], [23].

In contrast, pure vision transformers (ViTs) and their mobile variants, such as PMVT [12], process entire leaves by mapping visual patches through global multi-head self-attention. This architectural formulation empowers transformer backbones to capture long-range contextual relationships and holistic foliar geometry [11], [12]. Nevertheless, medical and agricultural pathology analyses indicate that pure ViTs struggle to retain high-frequency textural details critical for identifying minute fungal or bacterial lesions, often requiring multi-scale feature enhancement or convolutional assistance to avoid losing diagnostic cues [2], [23]. Consequently, hybrid models combining

convolutional front-ends with transformer back-ends—such as those evaluated by Chen et al. [2] for tomato leaf diseases, De Silva and Brown [11] for multispectral plant pathology, and Selvaraj et al. [7] for coffee disorders—consistently demonstrate superior feature extraction by concurrently preserving fine spatial lesion structures and global leaf context [11], [2], [7].

B. Robustness to Field Illumination, Background Clutter, and Occlusion

When transitioning from uniform laboratory photography to uncontrolled natural environments, severe performance degradation is widely reported across candidate paradigms [1], [13], [14]. Field acquisitions gathered under direct solar radiation encounter harsh specular reflections, deep shadows, and variable ambient color temperatures that perturb pixel-level values [13], [15], [8]. Research examining wheat and grape leaf diseases emphasizes that models validated exclusively on controlled glasshouse imagery experience acute drops in diagnostic accuracy and predictive confidence when tested against authentic field environments characterized by soil clutter, weed growth, and canopy interference [13], [14]. Under these conditions, standard CNN architectures exhibit pronounced sensitivity to background distractors, mistaking reflective soil or overlapping foliage for diseased margins unless augmented with dedicated background-filtering mechanisms such as those introduced in LightMixer [10] or specialized segmentation networks [6].

Pure attention networks and hybrid formulations provide measurably enhanced diagnostic resilience in the presence of complex backgrounds and structural occlusions [2], [22]. Lightweight vision transformer architectures—such as the real-time detection framework WMC-RTDETR engineered for tea plantations [22] and the Mobile Vision Transformer PMVT [12]—demonstrate an innate capacity to suppress irrelevant canopy clutter and disentangle multiple overlapping leaf targets [12], [22]. By routing visual features across self-attention maps, these models decouple lesion symptoms from background noise more effectively than purely convolutional baselines [2], [22]. Furthermore, integrating uncertainty-aware and confidence calibration mechanisms into deep architectures enables models to flag ambiguous predictions caused by intense outdoor illumination changes, mitigating false positives under uncalibrated field captures [15], [14].

C. Sample Scarcity and Few-Shot Adaptation Dynamics

A major impediment to developing broad-spectrum agricultural diagnostic tools in regional settings such as rural India is the chronic scarcity of curated, expert-annotated leaf datasets [1], [17], [5]. The literature indicates that purely convolutional networks, assisted by generic transfer learning or multi-model ensemble strategies, demonstrate reasonable baseline generalization on modest sample sizes [5], [8], [18]. For example, Sahu et al. [18] observe that ensembled CNNs can capture diverse foliar symptoms in beans despite limited class instances. Conversely, the absence of spatial inductive

bias in vision transformers manifests as acute data hunger; pure ViT architectures trained from scratch on constrained agricultural datasets exhibit pronounced vulnerability to overfitting and sluggish convergence rates [11], [12], [23].

To circumvent annotation bottlenecks in real-world agromonic contexts, data-efficient learning frameworks have shown substantial promise. Saad and Salman [17] demonstrate that one-shot metric learning frameworks utilizing Siamese or metric-embedding strategies can accurately categorize diverse field-acquired crop diseases from only a single or few reference exemplars, bypassing the requirement for expansive re-annotation cycles [17]. When applied in hierarchical classification settings—such as coffee foliar disease identification and visible symptom severity grading under unstructured field conditions [24]—deep models benefit significantly from structured inductive pipelines that separate disease identity from severity stages, reducing the total sample volume needed to achieve robust diagnostic separation [17], [24].

D. Edge Profile and On-Device Hardware Feasibility

The physical deployment of deep learning models on low-cost Android smartphones hinges on strict computational and memory budgets [12], [19], [9]. Edge-tailored convolutional detectors, such as YOLO-PLNet for peanut foliar diseases [19], MGA-YOLO for apple leaves [21], and KFP-YOLO for pear anomalies [16], show that pruning, lightweight backbone engineering, and channel reduction can compress parameter counts well within edge thresholds while maintaining real-time frame rates [19], [16], [21]. Similarly, agricultural deployment benchmarks on portable embedded hardware, such as maize seedling monitoring using MPE-YOLO [20] and compact grape leaf networks like GrapeLeafNet [9], illustrate that lightweight convolutional models can achieve low inference latencies on CPU-bound or low-power embedded processors without relying on expensive hardware accelerators [9], [20].

While lightweight vision transformers such as PMVT [12] succeed in reducing theoretical parameter counts and computational operations, on-device execution on budget smartphone chipsets presents practical bottlenecks. Vision transformers and hybrid topologies incorporate non-linear self-attention matrices and memory-intensive token routing mechanisms, which frequently induce memory bandwidth saturation on mobile system-on-chip architectures with limited shared RAM [12], [19], [25]. Furthermore, deploying deep models on uncalibrated mobile devices often demands uniform INT8 post-training quantization [19], [25], [16]. Although convolutional layers undergo INT8 quantization with negligible degradation in diagnostic accuracy [19], [25], quantizing multi-head attention blocks without accuracy loss remains computationally delicate [12], [25]. Consequently, current evidence suggests that highly compact CNNs and selectively pruned lightweight hybrid architectures represent the most viable operational compromise for offline execution on low-cost smartphones [12], [19], [9], [10].

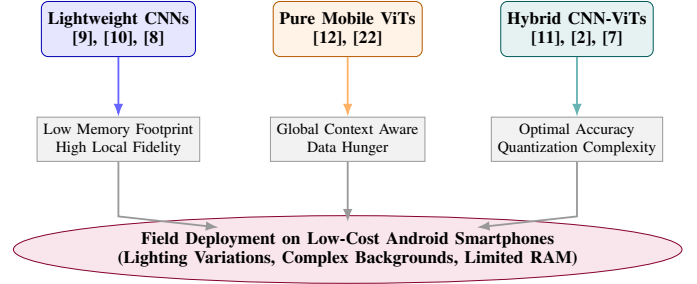


Fig. 5. Multi-attribute operational trade-offs across CNN, pure ViT, and hybrid models for field smartphone deployment

V. DISCUSSION

A. Operational Trade-offs Between Representation Power and Edge Deployability

The empirical synthesis reveals a persistent operational friction between the expressive contextual modeling of attention-based architectures and the computational parsimony required for low-cost mobile deployment [12], [2], [9]. Pure vision transformers and emerging hybrid CNN-ViT configurations demonstrate clear diagnostic advantages when disentangling overlapping foliar structures, multi-target lesions, and complex canopy clutter under variable outdoor conditions [11], [2], [22]. However, this elevated representational capacity comes with severe execution penalties on the hardware typical of low-cost Android smartphones in rural developing regions [12], [19], [15]. Entry-level mobile systems-on-chip lack high-throughput neural processing units, possessing narrow memory bus bandwidths and constrained random-access memory that struggle with the dynamic tensor reshaping and non-linear Softmax bottlenecks inherent to self-attention mechanisms [12], [25]. Consequently, while theoretical parameter counts of compressed transformers such as PMVT [12] approach those of compact convolutional networks, runtime memory footprints and latency scaling on ARM CPUs continue to favor lightweight convolutional topologies like GrapeLeafNet [9], LightMixer [10], and pruned one-stage detectors [19], [16], [21]. For agricultural extension deployments where immediate, offline feedback is vital, the marginal accuracy gains of unpruned transformer blocks are frequently outweighed by latency penalties and thermal throttling on entry-level devices.

B. Robustness to Optical Variations and Uncalibrated Smartphone Sensors

Deploying computerized foliar diagnostics in authentic agricultural environments introduces critical domain shifts that challenge standard visual feature extraction [1], [13], [14]. In-field captures by smallholder farmers are inevitably produced using budget smartphones equipped with uncalibrated CMOS optical sensors, differing color balances, compression artifacts, and inconsistent dynamic ranges [15], [8]. Under acute natural solar illumination, specular highlights and severe foliar shadows perturb pixel intensity distributions, causing conventional convolutional networks to trigger false positive

detections on reflective leaf margins or confuse desiccated leaf edges with necrotic fungal lesions [13], [14], [10]. To build dependable diagnostic applications, deep vision systems must transcend reliance on curated benchmarks and incorporate explicit environmental adaptation mechanisms [13], [14], [24]. Promising engineering remedies include integrating lightweight background suppression modules [10], leveraging multispectral or color-invariant feature projections [11], [15], and applying confidence calibration or uncertainty estimation [14]. Embedding uncertainty-aware thresholds allows mobile applications to identify out-of-distribution or ambiguous field samples, prompting users to reposition the camera rather than returning hazardous, false high-confidence misdiagnoses [15], [14].

C. Mitigating Chronic Small-Data Constraints in Smallholder Contexts

A foundational impediment to democratizing deep learning across diverse crop ecosystems—such as those found across fragmented farming plots in India—is the systemic shortage of large-scale, expertly annotated pathology datasets for localized cultivars [1], [17], [5]. Pure vision transformers are especially disadvantaged under these operational constraints due to their lack of spatial inductive bias, requiring extensive generic pre-training corpuses that may fail to capture subtle agronomic morphological cues [11], [12], [23]. In contrast, lightweight CNNs demonstrate greater baseline resilience on constrained sample distributions, particularly when bolstered by multi-model ensembling [18] or metric-based one-shot learning strategies [17]. The demonstrated success of metric embeddings in identifying novel foliar pathologies from minimal reference instances [17] suggests that agricultural extension frameworks should prioritize architectures capable of few-shot edge adaptation over brute-force overparameterized models. Hierarchical classification paradigms that decouple foliar organ localization, disease identification, and symptom severity grading [24] provide an additional structural mechanism to maximize diagnostic yield from limited training imagery, allowing practitioners to optimize lightweight backbones for specific diagnostic tiers [9], [24].

D. Threats to Validity and Practical Implementation Constraints

A rigorous interpretation of the reviewed evidence requires acknowledging significant methodological limitations and threats to external validity across published literature [1], [13]. Chief among these is the pervasive heterogeneity of experimental protocols: published benchmarks vary widely in dataset composition, image acquisition hardware, ambient illumination controls, and background conditions, preventing direct, apples-to-apples statistical comparisons across distinct studies [13], [6], [14]. Many high-accuracy benchmarks continue to rely on artificial indoor datasets or sanitized imagery with excised leaves, failing to reflect authentic in-situ field conditions where soil, weeds, and multi-organ occlusions dominate [1], [13], [6], [24]. Furthermore, edge-deployment

feasibility is often reported purely through theoretical floating-point operations rather than real-world profiling on constrained mobile hardware undergoing INT8 post-training quantization and memory bandwidth pressure [19], [25], [16], [20]. Agricultural practitioners must interpret reported laboratory accuracies with appropriate caution and prioritize empirical on-device stress testing using unconstrained, locally acquired field captures prior to regional deployment.

VI. CONCLUSION

This review has systematically investigated the architectural paradigms, environmental robustness, and edge feasibility of deep visual frameworks for crop leaf disease diagnosis, contextualized within the operational realities of smallholder agriculture and budget Android hardware. By cross-examining traditional and lightweight convolutional neural networks against vision transformers and emerging hybrid formulations, we have clarified the fundamental trade-offs governing in-field pathological classification. Standard and compact CNNs remain exceptionally competitive in computational parsimony and deployment ease, leveraging localized spatial priors to isolate minute necrotic structures while operating efficiently within constrained mobile memory bounds [9], [10], [8]. However, their performance degrades when confronted with severe solar glare, complex soil-canopy boundaries, and structural foliage occlusions typical of unconstrained field photography [13], [14].

Conversely, self-attention mechanisms and vision transformers present powerful capabilities for long-range spatial contextualization and distractor suppression, effectively disentangling overlapping leaf margins and background foliage clutter [12], [22]. Despite these advantages, pure ViTs suffer from severe data hunger, convergence difficulties on small domain-specific datasets, and non-linear memory access patterns that strain the shared random-access memory and thermal budgets of entry-level mobile system-on-chip architectures [11], [12], [23]. Hybrid topologies that combine convolutional shallow feature extractors with lightweight transformer bottlenecks emerge as the most resilient diagnostic paradigm, capturing high-frequency lesion textures while maintaining global canopy contextual fidelity [11], [2], [7]. Nevertheless, deploying these multi-stage networks onto low-cost smartphones mandates careful architectural co-design, structural pruning, and hardware-aligned INT8 quantization to avoid latency bottlenecks during on-device execution [12], [19], [25].

For agricultural engineering practitioners, extension agencies, and mobile tool developers targeting resource-limited farming communities, lightweight convolutional frameworks and selectively compressed hybrid architectures currently provide the most practical operational balance [12], [19], [9]. Until mobile runtime engines and low-tier hardware accelerators offer seamless native acceleration for dynamic attention operators, highly optimized convolutional backbones remain the primary vehicle for real-time, offline diagnostic applications on entry-level Android devices [19], [16], [8]. Bridging the remaining gap between laboratory benchmark metrics and

sustained field reliability requires moving beyond raw top-1 accuracy to prioritize calibration, uncertainty quantification, and data efficiency across uncalibrated consumer optics [15], [17], [14].

A. Future Work

To advance automated foliar pathology toward robust, accessible edge deployment in developing agricultural regions, future research should pursue the following concrete directions:

- **Quantization-Aware Hybrid Topologies:** Developing unified integer quantization frameworks tailored specifically for mobile hybrid CNN-ViT backbones, mitigating the accuracy loss typically encountered when quantizing multi-head self-attention matrices for low-tier ARM NEON and mobile CPU runtime environments [12], [19], [25].
- **Uncertainty-Calibrated Mobile Decision Support:** Integrating computationally lightweight epistemic and aleatoric uncertainty estimation directly into mobile diagnostic pipelines, enabling handheld tools to reject out-of-distribution imagery, dynamic motion blurs, and extreme lighting artifacts before presenting recommendations to farmers [15], [14].
- **Metric-Based Few-Shot Adaptation:** Designing generalized one-shot and few-shot metric learning models that allow agricultural extension workers to register emerging, localized crop disease variants in the field through a handful of reference photos without demanding extensive network retraining or cloud-based parameter fine-tuning [17], [18].
- **Hierarchical Multi-Task Symptom Parsing:** Formulating integrated edge models that simultaneously execute foliar disease detection, visible severity scoring, and organ segmentation under complex in-situ occlusion, providing actionable severity-graded advisories adapted to rural field constraints [6], [16], [24].

REFERENCES

- [1] D. Krishana, "A Dataset of Field Plant Images for Plant Disease Detection and Classification with Deep Learning," *International Journal Of Scientific Research In Engineering And Management*, 2024, doi: 10.55041/ijserm32034.
- [2] Z. Chen, G. Wang, T. Lv, and X. Zhang, "Using a Hybrid Convolutional Neural Network with a Transformer Model for Tomato Leaf Disease Detection," *Agronomy*, 2024, doi: 10.3390/agronomy14040673.
- [3] D. Pantha, "Fruits Leaf Disease Detection Using Convolutional Neural Network," *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 2024, doi: 10.55529/jaiml.42.1.13.
- [4] S. Guo, "Leaf Disease Detection by Convolutional Neural Network (CNN)," *Highlights in Science, Engineering and Technology*, 2023, doi: 10.54097/aex9r523.
- [5] V. S. Mandlik and L. V. Birgale, "A Convolutional Neural Network (CNN) Based Classification Framework for Multi-Crop Disease Detection Using Leaf Images," *Current Agriculture Research Journal*, 2026, doi: 10.12944/carj.13.3.15.
- [6] M. Shoaib et al., "Deep Learning-Based Segmentation and Classification of Leaf Images for Detection of Tomato Plant Disease," *Frontiers in Plant Science*, 2022, doi: 10.3389/fpls.2022.1031748.
- [7] K. Selvaraj, R. Selvanarayanan, S. K. G. Venkatesan, and S. Rajendran, "Future-Proof Coffee Plant Disease Detection Based on Counter-Factual Recommendation with a Hybrid Vision Transformer and Convolutional Neural Network Model," *Agrociencia*, 2025, doi: 10.47163/agrociencia.v59i4.3385.
- [8] C. Yang, Z. Teng, C. Dong, Y. Lin, R. Chen, and J. Wang, "In-Field Citrus Disease Classification via Convolutional Neural Network from Smartphone Images," *Agriculture*, 2022, doi: 10.3390/agriculture12091487.
- [9] M. Aslan, "GrapeLeafNet: A Lightweight and High-Performance Convolutional Neural Network for Grape Leaf Disease Detection," *Agronomy*, 2026, doi: 10.3390/agronomy16100976.
- [10] Y. Zhong, Z. Teng, and M. Tong, "LightMixer: A Novel Lightweight Convolutional Neural Network for Tomato Disease Detection," *Frontiers in Plant Science*, 2023, doi: 10.3389/fpls.2023.1166296.
- [11] M. De Silva and D. Brown, "Multispectral Plant Disease Detection with Vision Transformer-Convolutional Neural Network Hybrid Approaches," *Sensors*, 2023, doi: 10.3390/s23208531.
- [12] G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang, "PMVT: A Lightweight Vision Transformer for Plant Disease Identification on Mobile Devices," *Frontiers in Plant Science*, 2023, doi: 10.3389/fpls.2023.1256773.
- [13] M. Long, M. Hartley, R. J. Morris, and J. K. M. Brown, "Classification of Wheat Diseases Using Deep Learning Networks with Field and Glasshouse Images," *Plant Pathology*, 2023, doi: 10.1111/ppa.13684.
- [14] E. Ergün, H. Okumus, O. E. Cinar, and N. Batan, "Confidence and Uncertainty Aware Deep Learning for Reliable Grape Leaf Disease Diagnosis Under Real World Field Conditions," *Frontiers in Plant Science*, 2026, doi: 10.3389/fpls.2026.1839652.
- [15] A. Chakraborty, Y. Thota, C. Luca, and I. van der Linde, "Explainable Deep Learning for Neonatal Jaundice Classification Using Uncalibrated Smartphone Images," *Machine Learning and Knowledge Extraction*, 2025, doi: 10.3390/make7040136.
- [16] Z. Xu et al., "KFP-YOLO: A Lightweight Detection Model for Korla Fragrant Pear Disease and Pest Detection Toward Edge Deployment," *Agriculture*, 2026, doi: 10.3390/agriculture16151631.
- [17] M. H. Saad and A. E. Salman, "A Plant Disease Classification Using One-Shot Learning Technique with Field Images," *Multimedia Tools and Applications*, 2023, doi: 10.1007/s11042-023-17830-4.
- [18] P. Sahu, A. Chug, and A. P. Singh, "Crop Leaf Disease Detection for Beans Using Ensembled-Convolutional Neural Networks," *International Journal of Food Engineering*, 2023, doi: 10.1515/ijfe-2023-0055.
- [19] J. Sun, Z. Feng, J. Han, F. Xu, H. Zhang, and Y. Guo, "YOLO-PLNet: A Lightweight Real-Time Detection Model for Peanut Leaf Diseases Based on Edge Deployment," *Frontiers in Plant Science*, 2025, doi: 10.3389/fpls.2025.1707501.
- [20] L. Xia et al., "MPE-YOLO: An Efficient and Lightweight Model for Maize Plug-Tray Empty-Cell Detection and iOS Deployment," *Frontiers in Plant Science*, 2026, doi: 10.3389/fpls.2026.1886272.
- [21] Y. Wang, Y. Wang, and J. Zhao, "MGA-YOLO: A Lightweight One-Stage Network for Apple Leaf Disease Detection," *Frontiers in Plant Science*, 2022, doi: 10.3389/fpls.2022.927424.
- [22] Y. Zhang, J. Song, X. Yu, and X. Ji, "WMC-RTDETR: A Lightweight Tea Disease Detection Model," *Frontiers in Plant Science*, 2025, doi: 10.3389/fpls.2025.1574920.
- [23] O. Chibuike and X. Yang, "Convolutional Neural Network-Vision Transformer Architecture with Gated Control Mechanism and Multi-Scale Fusion for Enhanced Pulmonary Disease Classification," *Diagnostics*, 2024, doi: 10.3390/diagnostics14242790.
- [24] L. A. AlFrhan and A. Almaleh, "A Hierarchical Deep Learning Framework for Coffee Leaf Disease Detection and Visible Severity Classification Under Saudi Arabian Field Conditions," *Applied Sciences*, 2026, doi: 10.3390/app16126109.
- [25] R. Pandey, S. Rajput, and A. K. Agrahari, "Improved and Lightweight Diabetic Retinopathy Detection: Overcoming Class Imbalance and Advancing Mobile-Ready Clinical Deployment," *Journal of Electronic Imaging*, 2026, doi: 10.1117/1.jei.35.2.023017.