Journal of Computer Science

Review Article

A Systematic Review of Artificial Intelligence and Deep Learning Approaches for Cocoa Pod Disease Detection

Henry Techie-Menson[1], Michael Asante[2], Yaw Marfo Missah[2], Gaddafi Abdul-Salaam[2] and Stephen Opoku Oppong[1]

[1]*Department of ICT Education, University of Education, Winneba, Ghana*
[2]*Department of Computer Science, Kwame Nkrumah University of Science and Technology, Ghana*

Article history
Received: 23-03-2026
Revised: 04-06-2026
Accepted: 15-07-2026

Corresponding Author:
Stephen Opoku Oppong
Department of ICT Education,
University of Education,
Winneba, Ghana
Email: sooppong@uew.edu.gh

Abstract: Cocoa pod diseases destroy an estimated 700,000 metric tons of cocoa annually, threatening 40–50 million smallholder farmers across tropical regions. Accurate field diagnosis remains constrained by symptom similarity, labour-intensive scouting, and limited specialist access in remote communities. This paper presents a PRISMA-guided systematic review of 25 peer-reviewed studies on automated cocoa pod disease detection and classification (2016–2026), addressing architectural approaches, implementation strategies, performance outcomes, and research gaps. Studies were identified through a multi-database search (Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, PubMed) and appraised across seven quality dimensions including dataset adequacy, methodological rigour, evaluation completeness, and deployment feasibility. Convolutional neural networks dominate the literature (≈44% of algorithmic instances), with transfer learning from EfficientNet, MobileNet, ResNet, and VGG consistently outperforming training from scratch. Object detection frameworks (YOLO variants, SSD) account for 27% of instances, reflecting a shift toward spatial localisation. The Learnable Gated Fusion Convolutional Block Attention Module (LGF-CBAM), integrated with ResNetV2-101, achieves 98.95% classification accuracy, the strongest result validated across five independent datasets. Despite accuracy gains, six structural deficiencies undermine confidence in reported findings: Near-universal absence of statistical significance testing; heterogeneous evaluation metrics; no ablation studies; exclusive reliance on held-out test sets; inadequate class imbalance handling; and underreporting of computational complexity. Approximately 32% of studies report no overfitting mitigation, fewer than 20% employ cross-validation, 69% include no mobile or edge deployment component, and 32% use fewer than 500 training images. No current model achieves balanced strength across accuracy, dataset adequacy, overfitting control, mobile suitability, and computational efficiency. The review recommends: Attention-augmented and hybrid CNN-ViT architectures with integrated explainability; coordinated development of large-scale, georeferenced benchmark datasets; standardized evaluation protocols mandating cross-validation and full metric reporting; and deployment-oriented design prioritizing on-device inference, offline functionality, and formal usability testing with farmer populations.

Keywords: Cocoa Pod Disease Detection, Plant Disease Classification, Precision Agriculture, Smallholder Farming, Systematic Literature Review, Prisma Framework, Agricultural Computer Vision, Artificial Intelligence in Agriculture, Attention Mechanisms, Lgf-Cbam

SCIENCE Publications

© 2026 Henry Techie-Menson, Michael Asante, Yaw Marfo Missah, Gaddafi Abdul-Salaam and Stephen Opoku Oppong. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Introduction

Cocoa (*Theobroma cacao L.*) is a major global cash crop cultivated on approximately 11 million hectares, with Côte d'Ivoire, Ghana, Indonesia, Ecuador, and Cameroon accounting for most of the world's production (Asamoah, 2022). The crop supports between 40 and 50 million smallholder farmers across tropical regions and underpins a chocolate industry valued at more than USD 130 billion (Osei-Kwame and Mensah, 2024; Wilson and Chen, 2024). In producing nations such as Ghana, cocoa continues to generate rural employment, household income, and critical export revenue. Despite its socioeconomic importance, global production is persistently threatened by destructive diseases including black pod rot (*Phytophthora palmivora, P. megakarya*), frosty pod rot (*Moniliophthora roreri*), witches' broom (*M. perniciosa*), Cocoa Swollen Shoot Virus Disease (CSSVD), Vascular-Streak Dieback (VSD), and brown pod rot (Asare-Bediako et al., 2023; Ameyaw et al., 2023). Collectively, these pathogens reduce yields, endanger farmer livelihoods, and threaten the long-term sustainability of the cocoa sector.

Accurate disease diagnosis remains difficult because early symptoms are visually similar and influenced by environmental and lighting variability. Conventional scouting and expert inspection are labour-intensive, subjective, and slow, limiting scalability across plantations (Singh and Kumar, 2023). Such dependence on human expertise often delays timely intervention, especially in remote farming communities. Consequently, there is growing demand for automated, objective, and scalable diagnostic systems that enable rapid detection and precision management. Advances in computer vision, artificial intelligence, Machine Learning (ML), and Deep Learning (DL) have transformed plant disease diagnosis by allowing models to learn discriminative visual patterns directly from digital imagery (Mohanty et al., 2016; Kouassi et al., 2024; Zhang et al., 2023).

Early cocoa disease studies relied on handcrafted feature extraction colour histograms, Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Gabor filters paired with classical classifiers such as Support Vector Machines (SVMs), decision trees, and random forests (Basri et al., 2019; 2022). Tan et al. (2016) applied k-means clustering and SVMs for pod analysis, yet these methods lacked robustness under field variability. The emergence of deep Convolutional Neural Networks (CNNs), supported by faster GPUs, transfer learning, and large annotated datasets, has dramatically improved classification accuracy through automatic end-to-end feature learning. Architectures such as VGGNet, ResNet, DenseNet, and EfficientNet have since yielded superior results in cocoa disease classification (De Oliveira and Romero, 2018; Amoako et al., 2023; Kouassi et al., 2025; Maylianti et al., 2025).

More recent innovations include lightweight, mobile-ready models such as MobileNet, enabling smartphone-based diagnostic systems for smallholder use (Kumi et al., 2022; Garma, 2025). Hybrid approaches that combine CNNs with traditional classifiers, as well as object detection frameworks such as YOLO, SSD, and Faster R-CNN, have enhanced localisation and classification in complex field settings (Serrano et al., 2020; Mamadou et al., 2023; Ferraris et al., 2023). Attention mechanisms, ensemble learning, and CNN–Transformer hybrids further improve contextual feature extraction and robustness, achieving high classification accuracies under controlled scenarios (Kouassi et al., 2025). Beyond image-based classification, precision agriculture technologies such as UAV-mounted multispectral imaging (Gupta and Sharma, 2023) and IoT-enabled sensor networks (Zhang et al., 2023) support large-scale monitoring and environmental correlation for disease forecasting. Smartphone-based diagnostic systems for other crops including cassava, maize, and tomato demonstrate viable, low-cost pathways for cocoa adaptation (Ramcharan et al., 2017; Ferentinos, 2018).

Problem Statement

Despite these advances, research on automated cocoa disease detection remains fragmented and methodologically inconsistent (Zambrano-Vega et al., 2024). Many studies report high accuracies but neglect dataset imbalance, overfitting, computational complexity, reproducibility, and practical deployment. Moreover, no comprehensive review has yet evaluated cocoa pod disease detection with respect to methodological rigour, dataset adequacy, hybrid architectures, and field readiness (Kouassi et al., 2024; Zambrano-Vega et al., 2024).

This systematic review addresses that gap through a PRISMA-guided critical synthesis of the literature on cocoa pod disease detection and classification. The specific objectives and research questions guiding this synthesis are detailed in Sections 1.3 and 1.4 respectively.

Objectives

This review is set out to achieve the following objectives:

1. To identify and categorize the automated computational systems, machine learning models, and deep learning architectures employed in existing studies on cocoa pod disease detection and classification
2. To critically evaluate implementation strategies, encompassing dataset composition, preprocessing and data augmentation techniques, overfitting mitigation approaches, computational efficiency, and evaluation and validation protocols.
3. To comparatively synthesize the major findings, methodological trends, and performance outcomes reported across the reviewed studies

4. To identify key research gaps, methodological limitations, and priority directions for future research toward the development of scalable, robust, interpretable, and field-deployable cocoa disease diagnostic systems

Research Questions

The study is guided by the following research questions:

RQ1: What automated computational systems, machine learning models, and deep learning architectures have been applied in existing studies for cocoa pod disease detection and classification?

RQ2: What implementation methods are employed in existing cocoa pod disease detection studies across the following sub-dimensions?

a) How are dataset composition and data acquisition strategies designed and utilized?
b) What preprocessing and data augmentation techniques are commonly applied to improve model performance and generalization?
c) How do existing studies implement overfitting control and model regularization?
d) What computational efficiency approaches and evaluation/validation protocols are used to ensure reliable performance assessment?

RQ3: What are the key findings, methodological trends, and comparative performance outcomes reported across existing studies?

RQ4: What are the major research gaps, methodological limitations, and emerging priority areas for future research in the development of scalable, robust, interpretable, and field-deployable cocoa disease diagnostic systems?

Methods

This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological transparency, replicability, and rigor. The protocol encompasses a structured multi-database search, predefined inclusion and exclusion criteria, a staged screening process, systematic data extraction, quality assessment, and evidence synthesis.

As a systematic review, this study synthesizes and critically appraises existing primary research rather than generating new experimental data. Accordingly, cross-validation experiments and independent dataset validation while essential in primary empirical studies fall outside the scope of this review. The adequacy of such practices in the reviewed literature is instead evaluated as part of the quality assessment framework and their absence across the reviewed studies is identified as a critical research gap

Inclusion Criteria

Studies were included if they met all of the following conditions:

1. Peer-reviewed journal articles or conference papers published between January 2016 and March 2025
2. Research directly focused on cocoa pod disease detection, classification, or quantification using machine learning, deep learning, feature extraction, or mobile/edge-based diagnostic systems
3. Reporting of at least one quantitative performance metric (e.g., accuracy, precision, recall, F1-score, mean Average Precision (mAP), or AUC-ROC)
4. Sufficient methodological detail to permit evaluation of experimental design, dataset description, model architecture, training procedure, or deployment context

Exclusion Criteria

Studies were excluded if they met any of the following conditions:

1. Non-empirical works, including purely conceptual papers, opinion pieces, or existing review articles
2. Studies addressing plant diseases in crops other than cocoa (*Theobroma cacao*)
3. Articles lacking adequate technical detail on methodology, model design, or experimental setup, precluding meaningful quality assessment
4. Studies focused exclusively on cocoa leaf, root, or trunk diseases with no attention to pod-level detection or classification

Search Strategy

A comprehensive search was conducted across five major scholarly databases: Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, and PubMed. Boolean search strings combining domain-specific and methodological terms were employed, as detailed in Table 1.

Study Selection

The PRISMA-guided selection process is illustrated in Figure 1. A total of 84 records were retrieved from database searches. After removing 8 duplicates, 76 records advanced to title and abstract screening. At this stage, 51 records were excluded for one or more of the following reasons: Irrelevance to cocoa pod disease, absence of ML/DL methodology, insufficient methodological reporting, or incompatible study type.

Table 1: Search Terms

Search String
("Cocoa pod disease" OR "cocoa disease detection") AND ("machine learning" OR "deep learning" OR "CNN")
("Cocoa pod classification") AND ("feature extraction" OR "image processing")
("Cocoa pod pests" OR "feature extraction") AND ("cocoa disease")
("Mobile application" OR "edge computing") AND ("cocoa pod disease detection")
("Transfer learning" OR "object detection") AND ("*Theobroma cacao*" OR "cocoa")

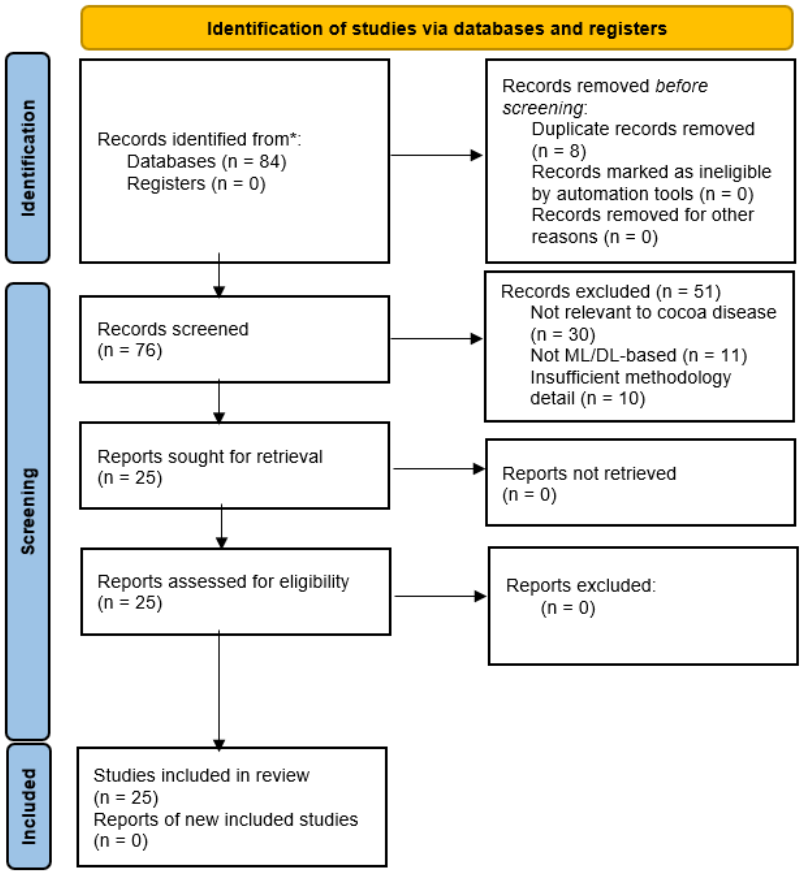

Fig. 1: PRISMA Flow Diagram of Study Selection Process (Identification → Screening → Eligibility → Inclusion: 84 → 76 → 25 → 25)

The remaining 25 records were retrieved in full text, all of which were successfully obtained. Full-text assessment against the eligibility criteria resulted in no further exclusions, yielding 25 studies for inclusion in the final synthesis.

Data Extraction

A standardized extraction template was applied consistently across all included studies. Table 2 presents the extraction framework.

Quality Assessment

Each included study was critically evaluated across the following dimensions:

1. Relevance to cocoa pod disease detection and classification objectives
2. Methodological rigor, including experimental design, reproducibility, and statistical validation
3. Dataset adequacy in terms of size, diversity, class balance, and documentation quality
4. Implementation transparency regarding preprocessing, hyperparameter configuration, and training procedures
5. Evaluation completeness based on the range and appropriateness of reported performance metrics
6. Computational reporting, including inference time, model size, and computational complexity
7. Deployment consideration with respect to practical feasibility and real-world applicability

Table 2: Data Extraction Framework

Category	Extracted Information
Study Details	Authors, year, publication venue, country of origin
Problem Addressed	Disease type(s), detection vs. classification objective
Dataset Characteristics	Size, source, class distribution, environmental diversity
Feature Extraction Methods	Color-based (HSV, LAB), texture-based (Gabor, LBP, GLCM), deep learning-based
Preprocessing / Augmentation	CLAHE, rotation, resizing, color adjustment, flipping, cropping, noise injection, thresholding, grayscale conversion, histogram equalization
Model Architecture	ML models (SVM, RF, KNN, XGBoost, GLM); DL models (CNN variants, YOLO, EfficientNet, ResNet, MobileNet, ViT, SSD)
Training Configuration	Optimizer, learning rate, batch size, epochs, loss function
Overfitting Mitigation	Dropout, pooling, batch normalization, K-fold cross-validation, ReLU activation, data augmentation
Performance Metrics	Accuracy, precision, recall, F1-score, mAP, AUC, sensitivity, specificity
Computational Complexity	Parameter count, FLOPs, inference time, model size
Deployment Feasibility	Mobile application, edge computing, IoT integration, real-time capability
Reported Research Gaps	As identified by the original authors

Table 3: Summary of Models and Algorithms Used

S. No	Reference	Model / Algorithm
1	Techie-Menson et al. (2026)	ResNet101, LGF-CBAM
2	Kouassi et al. (2025)	Hybrid CNN + Vision Transformer (ViT) with SVM and LightGBM
3	Soh et al. (2024)	Custom CNN, VGG-16, EfficientNetB0, ResNet50, LeNet-5
4	Maylianti et al. (2025)	EfficientNet-B0, ResNet-50
5	Kumi et al. (2022)	SSD MobileNetV2, EfficientDet D0, CenterNet ResNet50V2, SSD ResNet50 V1 FPN
6	Mamadou et al. (2023)	MobileNetV2 hybrid with SVM, RF, XGBoost, KNN, LR
7	Godmalin et al. (2022)	NASNetMobile, MobileNetV3Small, EfficientNetB0
8	Lomotey et al. (2024)	CenterNet ResNet50V2, EfficientDet D0, SSD MobileNetV2, SSD ResNet50V1 FPN
9	Montesino et al. (2021)	ResNet18
10	Godmalin et al. (2023)	Fine-tuned EfficientNetB3, MobileNetV3Small
11	Rola et al. (2024)	Custom CNN, Naïve Bayes, Random Forest, Hoeffding Tree
12	Vera et al. (2024)	AlexNet, GoogLeNet, ResNet18, MobileNetV3-Small, YOLOv3, EfficientDet-Lite4, EfficientNet-Lite4
13	Tan et al. (2016)	k-Means Clustering + SVM
14	De Oliveira and Romero (2018)	Inception-ResNet-v2 (transfer learning)
15	Basri et al. (2019)	Gabor Kernel Convolution + ML classifiers
16	Ariandi et al. (2019)	Forward Chaining (rule-based expert system)
17	Gamboa et al. (2019)	GLM, SVM regression
18	Subramanian et al. (2025)	Dense CNN with ReLU + Softmax
19	Basri et al. (2022)	SVM (Linear, RBF, Polynomial kernels)
20	Atianashie (2024)	Custom CNN, SVM
21	Garma (2025)	SSD MobileNetV2 FPN-Lite with Dilated Convolutions
22	Amoako et al. (2023)	VGG19, VGG16, ResNet50 (transfer learning)
23	Serrano et al. (2020)	YOLOv4
24	Ferraris et al. (2023)	YOLOv5m, CIWA-net (U-Net based)
25	Pusadan et al. (2022)	K-Nearest Neighbor (KNN)

Quality appraisal did not serve as grounds for exclusion; rather, it informed the depth and critical tone of the synthesis.

As summarized in Table 3, the reviewed studies employ a diverse range of computational approaches, including convolutional neural networks, transfer-learning architectures, object-detection models, and conventional machine-learning algorithms.

Results and Discussion

RQ1: What automated computational systems, machine learning models, and deep learning architectures have been applied in existing studies for cocoa pod disease detection and classification?

Methodological Approaches

The 25 included studies exhibit considerable methodological diversity, spanning rule-based expert systems, classical machine learning classifiers, standalone CNNs, transfer learning with pre-trained architectures, object detection frameworks, and hybrid CNN-transformer models. This diversity reflects both the evolutionary trajectory of the field over the review period (2016–2025) and the varying computational constraints faced by different research groups.

As illustrated in Figure 2, analysis of the 25 included studies reveals a clear dominance of CNN-based architectures, accounting for approximately 44% of all documented algorithmic instances. Object detection models YOLO variants, SSD frameworks, and EfficientDet comprise roughly 27%, reflecting growing recognition that localization, not merely classification, is central to practical field deployment. Traditional machine learning models (SVM, RF, KNN, GLM, XGBoost, Naïve Bayes) account for approximately 24% of instances, while legacy and rule-based systems make up the remaining 5%.

Notably, SVM remains surprisingly prevalent, appearing in 8 of the 25 studies either as a standalone classifier or as a hybrid component. This persistence reflects SVM's robustness with small, high-dimensional feature sets and its computational efficiency relative to deep models both meaningful advantages in resource-constrained agricultural settings.

A particularly important trend is the growing adoption of hybrid architectures that pair deep feature extraction with classical classification backends. Kouassi et al. (2025) exemplify this approach with a CNN-ViT feature extractor coupled to SVM and LightGBM classifiers. Mamadou et al. (2023); Lomotey et al. (2024) similarly demonstrate that MobileNetV2-based feature extraction combined with classical classifiers achieves performance competitive with fully end-to-end deep models. This hybridization strategy is especially promising where labelled training data are scarce and computational resources are limited, since the classical backend can be retrained efficiently without reprocessing the entire deep network.

Figure 3 presents the distribution of algorithmic categories across the 25 included studies, both as a per-study stacked breakdown (Figure 3a) and as an aggregate proportional share (Figure 3b).

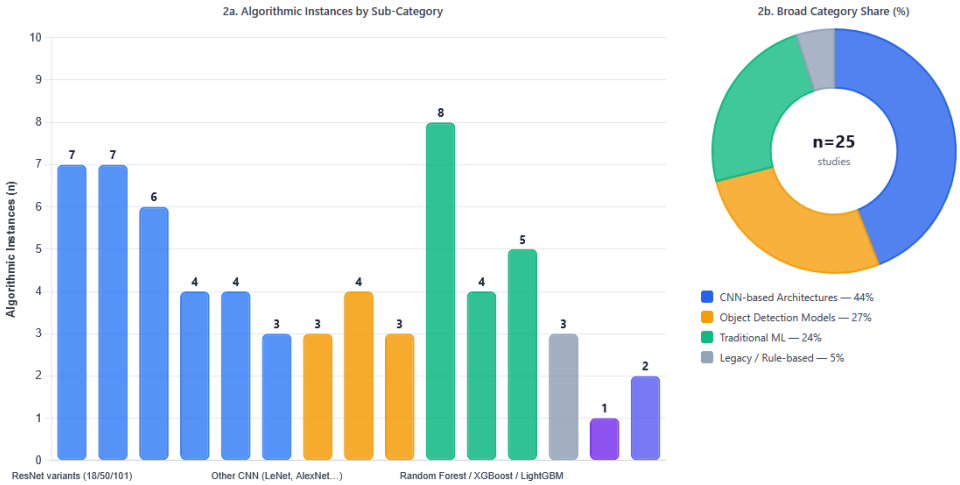

Fig. 2: Distribution of Model/Algorithm Types across Reviewed Studies

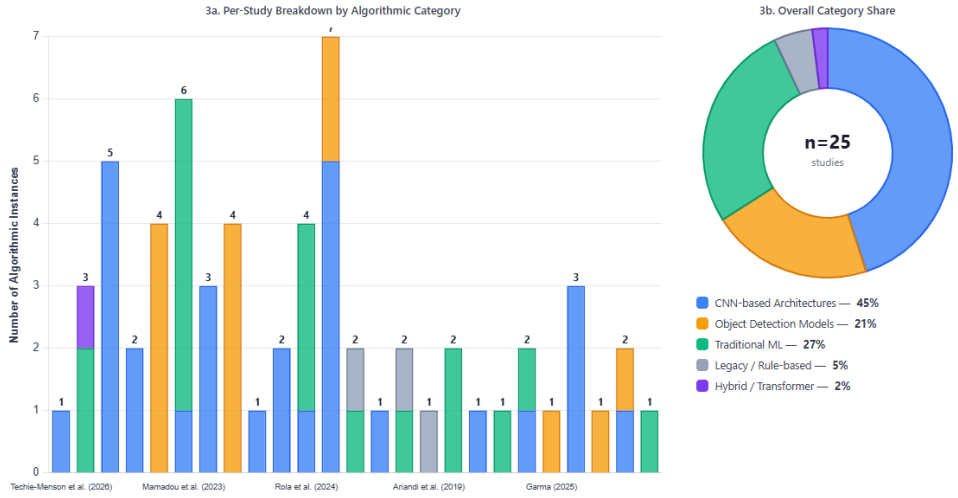

Fig. 3: Algorithmic Category Distribution

Figure 3 presents the distribution of algorithmic categories across the 25 included studies, both as a per-study stacked breakdown (Figure 3a) and as an aggregate proportional share (Figure 3b). Together, the two panels reveal not only what methods dominate the field but also how methodological choices cluster within individual studies.

CNN-based architectures account for the largest share 45%, confirming their central role in cocoa disease detection research over the review period. This dominance is evident across the temporal span of the literature, from studies as early as De Oliveira and Romero (2018) to more recent work by Subramanian et al. (2025). The sustained use of CNNs reflects continued confidence in convolutional feature extraction for plant pathology tasks. Notably, several studies, including Soh et al. (2024) and Vera et al. (2024), evaluated five CNN variants simultaneously, indicating a benchmarking-oriented approach rather than reliance on a single optimised architecture.

Traditional machine learning methods constitute the second-largest category 27%, encompassing SVM, KNN, logistic regression, GLM, random forest, XGBoost, LightGBM, Naïve Bayes, and Hoeffding Tree classifiers. Among these, SVM is the most frequently used classifier, appearing in eight studies either as a standalone method or as part of a hybrid approach. Its continued prominence, despite the increasing availability of deep learning alternatives, may reflect the practical advantages of SVM in low-data, high-dimensional settings, which are common in agricultural research.

Object detection frameworks account for 21% of the reviewed approaches, represented primarily by SSD variants, YOLO architectures, and EfficientDet. Studies such as Kumi et al. (2022) and Lomotey et al. (2024) are exclusively detection-oriented, with each evaluating four frameworks in parallel. This pattern indicates a progressive shift from image-level classification towards spatial localisation, which is more applicable to real-world field conditions where multiple pods, lesions, or disease instances may occur within a single image. The smaller proportion of detection-based studies compared with traditional machine learning approaches may be attributed to the characteristics of the datasets used in earlier research. Many early studies relied on pre-cropped or single-lesion images, for which image-level classification was sufficient, whereas bounding-box and mask-level annotations have become more prominent only in recent studies.

Legacy and rule-based systems account for 5% of the reviewed approaches and are largely confined to earlier studies, including Ariandi (2019), Tan (2016), and Basri et al. (2019). Their concentration in earlier publications indicates a clear temporal decline in the use of conventional, rule-based approaches as data-driven methods have become more prevalent.

Hybrid and transformer-based methods represent the smallest category 2%, currently limited to the CNN–Vision Transformer architecture introduced by Kouassi et al. (2025). Although this represents an early indication of an emerging methodological direction, the limited representation of transformer-based and attention-enhanced architectures across the reviewed literature highlights a notable research gap. This gap suggests considerable scope for investigating attention mechanisms, hybrid architectures, and transformer-based models for improved cocoa disease detection and localisation.

Comparative Analysis of Hybrid and Standalone Models: Performance Trade-offs, Backend Classifier Integration, and Deployment Constraints

Across the reviewed studies, hybrid architectures consistently demonstrated superior performance relative to their standalone counterparts. For example, the MobileNetV2+SVM framework proposed by Mamadou et al. (2023) achieved higher classification accuracy than standalone MobileNetV2, while the CNN–Vision Transformer hybrid developed by Kouassi et al. (2025) outperformed all single-architecture baselines evaluated in the study. These findings are theoretically justified, as convolutional neural networks excel at hierarchical feature extraction and spatial representation learning, whereas classical machine learning classifiers can provide greater flexibility in decision boundary construction and often generalize more effectively under limited-data conditions.

The integration of classical classifiers including SVM, Random Forest, XGBoost, LightGBM, and KNN as decision backends to deep feature extractors introduces several advantages that are largely overlooked in the reviewed literature. By decoupling feature extraction from classification, hybrid architectures enable the classifier to be retrained or updated independently when new disease classes, cultivars, or regional variations emerge, without requiring complete retraining of the deep network. This modularity is particularly advantageous in agricultural environments where labeled data accumulate incrementally over multiple growing seasons. Among these approaches, SVMs employing radial basis function kernels are especially suitable for the high-dimensional, low-sample-size feature spaces characteristic of cocoa disease datasets. Their maximum-margin optimization framework is generally less susceptible to overfitting than softmax-based classifiers trained on limited data. Furthermore, classical backends offer a degree of interpretability absent in fully end-to-end deep learning systems. For instance, analysis of support vectors and feature importance measures can provide insights into the characteristics driving classification decisions, partially addressing the explainability gap observed across the reviewed studies.

Despite these advantages, hybrid architectures are subject to important limitations. Most notably, the feature extractor and classifier are typically optimized independently rather than jointly. In end-to-end deep learning frameworks, gradients from the classification loss are propagated throughout the entire network, enabling feature representations to adapt specifically to the target task. In contrast, when a frozen or separately trained feature extractor is coupled with a classical classifier, this joint optimization process is lost, potentially resulting in features that are not maximally discriminative for disease recognition. This limitation becomes particularly significant when pretrained backbones originate from domains substantially different from cocoa pod imagery, such as ImageNet, and when dataset sizes are insufficient to support extensive fine-tuning. Moreover, classical backends introduce an additional layer of hyperparameter optimization including kernel selection, regularization parameters, tree depth, and ensemble configuration that is seldom reported transparently in the reviewed studies. Consequently, an important source of methodological variation remains concealed, complicating both reproducibility and comparative evaluation.

The choice between end-to-end deep learning and hybrid architectures should therefore be viewed as a deployment-dependent decision rather than a purely performance-driven one. For static applications involving fixed datasets and infrequent model updates, end-to-end fine-tuning of pretrained networks is likely to achieve superior predictive performance. Conversely, for dynamic deployment environments where new disease classes, cultivars, or geographic variants must be incorporated regularly, hybrid architectures offer greater flexibility, maintainability, and adaptability. Notably, none of the reviewed studies explicitly considered deployment context when selecting model architectures, instead focusing primarily on benchmark accuracy. Future research should justify architectural choices based not only on predictive performance but also on operational requirements, update frequency, and resource constraints.

A further limitation of the current literature is the absence of formal ablation studies designed to isolate the contribution of individual components within hybrid architectures. Without such analyses, it remains unclear whether reported performance gains arise from the hybrid design itself or from confounding factors such as dataset composition, preprocessing strategies, data augmentation techniques, or hyperparameter settings. This lack of component-level evaluation reduces the interpretability of reported improvements and limits the reproducibility of proposed methods. Given the complexity of hybrid systems and the potential interactions among their constituent components, future studies should treat ablation analysis as a standard reporting requirement rather than an optional experimental addition.

The Target Platform column in Table 6 (Computational Complexity and Runtime Metrics) exposes a critical yet underexplored dimension of deployment feasibility: The infrastructure assumptions underlying each study's deployment claims. Although these assumptions are rarely articulated explicitly, they fundamentally determine whether a technically functional system can operate effectively within the environments it is intended to serve.

Among the studies reporting deployment-oriented implementations, three target Android mobile devices (Serrano et al., 2020; Garma, 2025; Ferraris et al., 2023), two reference a dedicated application, Cocoa Companion (Kumi et al., 2022; Lomotey et al., 2024) without specifying the underlying inference architecture, and one deploys a web-based interface (Pusadan et al., 2022). These deployment platforms impose markedly different computational and operational requirements, each carrying important implications for real-world adoption in cocoa-growing regions.

Mobile deployment assumes that end users possess smartphones with sufficient computational resources, including adequate RAM (typically at least 2 GB for TensorFlow Lite inference), storage capacity for model files (ranging from approximately 14 MB for MobileNetV2 to over 170 MB for VGG19), and battery endurance to support repeated field-based inference. In many cocoa-producing regions of West Africa and Southeast Asia, however, the dominant device category consists of entry-level Android smartphones such as Tecno, Itel, and Infinix devices in West Africa, and Xiaomi Redmi and Samsung Galaxy A-series models in Southeast Asia which generally provide 2–3 GB RAM, low- to mid-range processors, and 32–64 GB of storage. Despite this reality, none of the reviewed studies identifies a target device class or reports benchmarking results on representative hardware commonly used by smallholder farmers. Consequently, claims regarding mobile deployment feasibility remain difficult to validate against the actual computational capabilities of the intended user population.

Connectivity requirements constitute perhaps the most significant deployment constraint and, paradoxically, the least discussed. Cocoa-growing regions in Côte d'Ivoire, Ghana, Cameroon, Indonesia, and Ecuador frequently experience unreliable or nonexistent mobile data coverage, particularly in remote plantation environments. Under such conditions, server-side inference architectures which appear to underpin several of the reviewed applications, although often only implicitly, cannot provide reliable functionality. For practical field deployment, offline-first architectures capable of executing the entire inference pipeline locally on the

device are essential. Yet only Garma (2025); Vera et al. (2024) present implementations that appear genuinely capable of offline operation through the use of TensorFlow Lite optimized models. By contrast, the web-based solution proposed by Pusadan et al. (2022) is inherently dependent on network connectivity, limiting its suitability for deployment in many target environments.

Another important but entirely overlooked consideration is energy consumption. Disease scouting activities often occur in remote plantation settings where access to charging infrastructure is limited or unavailable. Sustained operation over extended periods therefore requires energy-efficient inference. Power consumption varies considerably across model architectures and hardware platforms and is influenced by factors such as processor type, model complexity, and inference frequency. Despite its practical importance, none of the reviewed studies reports energy consumption metrics, suggesting that deployment evaluations remain focused almost exclusively on predictive performance while neglecting a key determinant of field usability.

Taken together, these observations indicate that deployment claims within the reviewed literature are largely aspirational rather than empirically validated. Demonstrating genuine deployment readiness requires evaluation on representative target hardware, testing under realistic connectivity conditions, and assessment of energy consumption during typical field-use scenarios. None of these factors has been systematically addressed in the current body of research, highlighting a significant gap between laboratory-level model performance and real-world deployment feasibility.

RQ2: What implementation methods are employed in existing cocoa pod disease detection studies across the following sub-dimensions?

RQ2a: Dataset Characteristics and Sizes

Dataset quality and scale are fundamental determinants of model generalizability, a concern that is particularly acute in cocoa disease detection given the substantial intra-class visual variability arising from geographic, seasonal, and phenological differences in disease presentation.

The dataset size distribution across the reviewed studies is substantially skewed toward small samples, as illustrated in Figure 4. Approximately 32% of studies employed fewer than 500 images a scale that is critically insufficient for training robust deep learning models without extensive augmentation or transfer learning strategies. Datasets in the range of 500–2,000 images account for 24% of studies, representing a middle tier that is workable with careful regularisation but remains limited for generalisation across diverse field conditions.

A further 16% of studies utilised datasets in the 2,001–4,000 image range, while only 8% specifically Soh et al. (2024); Maylianti et al. (2025); Atianashie (2024) reported original collections exceeding 4,000 images, representing the strongest data foundations in the review. Approximately 20% of studies failed to report dataset size altogether including Kumi et al. (2022); Montesino et al. (2021); Ariandi et al. (2019); Gamboa et al. (2019); Subramanian et al. (2025) representing a transparency deficiency that precludes meaningful cross-study comparison and reproducibility assessment.

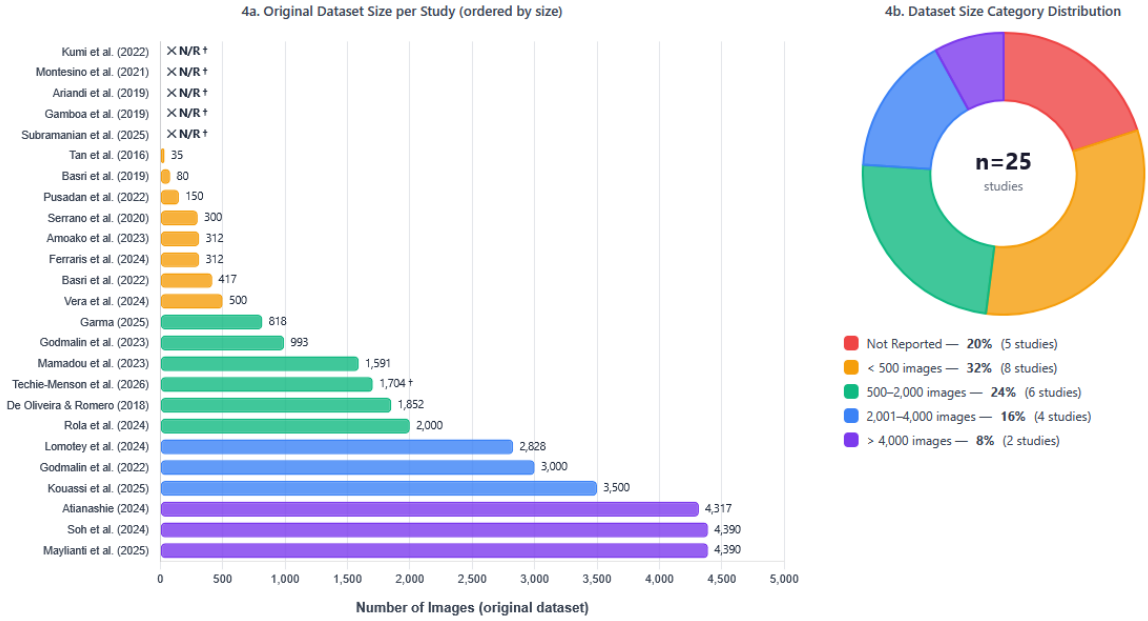

Fig. 4: Distribution of Dataset Sizes across Reviewed Studies

Table 4: Summary of Implementation Procedure

S. No	Reference	Dataset Size	Preprocessing / Augmentation	Overfitting Management	Mobile / Edge Integration
1	Techie-Menson et al. (2026)	1,704 (aug. to 73,500)	Rotation ±45°/90°, H/V Flip, Scaling 0.8–1.2×, Brightness ±20%, Hue ±10°, Resizing 224×224	L2 Regularisation, Batch Normalisation, Dropout, Early Stopping, GAP	—
2	Kouassi et al. (2025)	3,500	CLAHE, Rotation	—	—
3	Soh et al. (2024)	4,390	Resizing, Colour Adjustment	Dropout, Pooling	—
4	Maylianti et al. (2025)	4,390	Resizing	—	—
5	Kumi et al. (2022)	Not reported	Rotation, Blurring, Translation, Noise, V/H Flip	—	Cocoa Companion
6	Mamadou et al. (2023)	1,591	Rotation, H/V Flip	Data Augmentation	—
7	Godmalin et al. (2022)	3,000	V/H Flip	Data Augmentation, K-Fold CV	—
8	Lomotey et al. (2024)	2,828	Rotation, Blurring, Translation, Noise, V/H Flip	Data Augmentation	Cocoa Companion
9	Montesino et al. (2021)	Not reported	—	—	—
10	Godmalin et al. (2023)	993	—	Data Augmentation, K-Fold CV	—
11	Rola et al. (2024)	2,000	Cropping	ReLU Activation	—
12	Vera et al. (2024)	500	Cropping, Resizing	—	Cacao DL
13	Tan et al. (2016)	35	Brightness Adjustment	—	—
14	De Oliveira & Romero (2018)	1,852	Flipping, Colour Adjustment, Resizing	Dropout	—
15	Basri et al. (2019)	80	Grayscale Conversion	—	—
16	Ariandi et al. (2019)	Not reported	—	—	Cocoa Disease ID System
17	Gamboa et al. (2019)	Not reported	—	—	—
18	Subramanian et al. (2025)	Not reported	Rescaling, Resizing, Otsu's Thresholding	ReLU Activation	—
19	Basri et al. (2022)	417	Grayscale Conversion	—	—
20	Atianashie (2024)	4,317	Grayscale Conversion, Resizing, H-Flip	ReLU, Data Augmentation	—
21	Garma (2025)	818	Resizing, H-Flip, Rotation, Saturation, Exposure	Data Augmentation	Mobile / Edge Device
22	Amoako et al. (2023)	312	Resizing, Histogram Equalisation, Rotation	Data Augmentation	—
23	Serrano et al. (2020)	300	—	—	App Cocoa Diseases
24	Ferraris et al. (2023)	312	H-Flip, Random Transformations	Data Augmentation, ReLU	Mobile App
25	Pusadan et al. (2022)	150	Cropping, Resizing, RGB-to-HSV	K-Fold Cross-Validation	Web Interface

As summarized in Table 4, the reviewed studies vary considerably in dataset size, preprocessing and augmentation strategies, overfitting-control techniques, and the extent of mobile or edge-device integration.

It is worth noting that raw dataset size alone does not fully capture data adequacy. Techie-Menson et al. (2026), for instance, collected 1,704 original images placing them in the 500–2,000 tier but expanded the effective training corpus to 73,500 images through a comprehensive augmentation pipeline, by far the largest augmented dataset in the review. Conversely, several studies reporting moderate original dataset sizes applied little or no augmentation, which may have constrained model generalisation despite nominally larger collections. This distinction between original and effective training size is an important methodological consideration that is inconsistently reported across the literature.

This dataset inadequacy carries several compounding implications for the field. First, models trained on small, narrowly sampled datasets are prone to overfitting and poor generalization to unseen conditions particularly across diverse geographic locations, cocoa varieties, and disease stages not represented in the original training data. Second, the absence of a standardized, publicly available benchmark dataset for cocoa pod diseases, analogous to PlantVillage in the broader plant disease domain, renders cross-study performance comparisons fundamentally unreliable. Reported accuracy figures are evaluated against different, often private datasets under varying conditions, making it impossible to attribute performance differences to model design rather than data characteristics. Third, class imbalance whereby images of certain disease stages or types are significantly underrepresented is rarely acknowledged or systematically addressed in the reviewed studies, despite its well-documented impact on model reliability, particularly for minority-class detection in early-stage infections where timely intervention is most critical.

A further and particularly consequential methodological concern is the near-universal absence of dataset diversity documentation. Very few studies report the geographic provenance of collected images, the cocoa variety or cultivar represented, imaging conditions such as lighting environment, camera-to-subject distance, or equipment specifications, or the distribution of disease stages within each class. Without such documentation, the interpretability of reported performance metrics is severely constrained, and the prospects for meaningful replication, cross-study synthesis, or deployment in conditions beyond the original data collection environment remain limited.

RQ2b: Preprocessing and Data Augmentation

Preprocessing and augmentation strategies are critical for maximizing the utility of limited datasets and improving model robustness to real-world imaging variability. The distribution of techniques employed across the reviewed studies is summarized in Figure 5.

Geometric transformations encompassing rotations, flips, cropping, resizing, and translation constitute the most frequently applied augmentation category, appearing in approximately 66% of studies. While these techniques effectively enhance spatial invariance and help models generalise across different pod orientations and framing conditions, they do not address variability in illumination, imaging equipment, or the appearance of disease symptoms across different environmental and geographic contexts.

Color and intensity adjustments, including brightness modification, contrast enhancement, histogram equalization, CLAHE, and saturation and exposure adjustments, were employed in approximately 19% of studies. These techniques are particularly relevant for cocoa disease detection, where the spectral properties of diseased tissue discoloration, necrosis, and lesion pigmentation carry strong diagnostic information. The relatively low adoption of colour-space augmentation is therefore a notable gap, given that chromatic features are among the most discriminative cues for distinguishing disease types and severity stages. CLAHE, for instance, was applied by only a single study in the review corpus, Kouassi et al. (2025), despite its well-established effectiveness in enhancing local contrast under variable field lighting.

Noise injection and blurring appear in approximately 8% of studies, targeting robustness to image quality degradation that is common under uncontrolled field conditions. Normalization and thresholding techniques including Otsu's thresholding and grayscale conversion account for approximately 4% of preprocessing operations, with their use largely confined to studies employing traditional machine learning pipelines rather than end-to-end deep learning architectures.

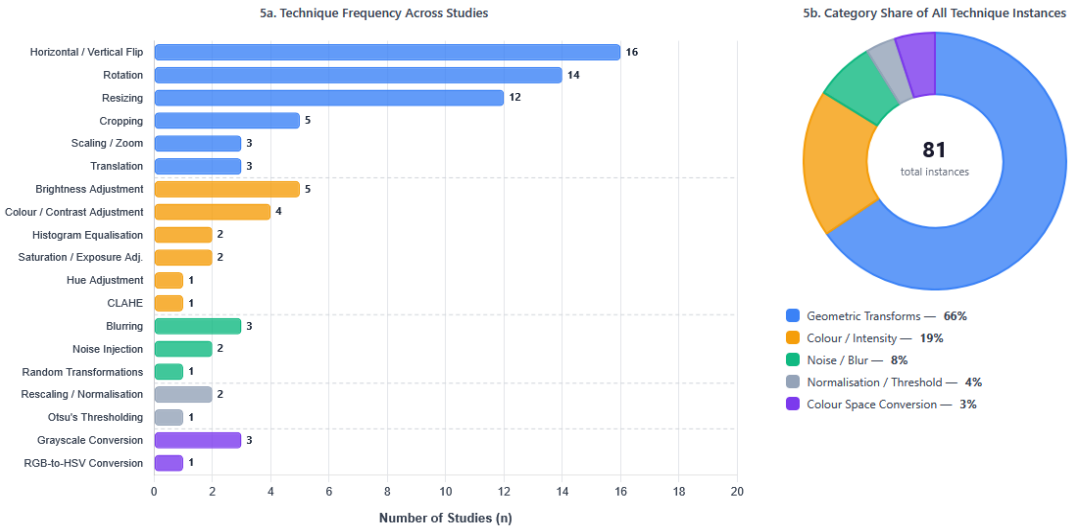

Fig. 5: Frequency of Preprocessing and Data Augmentation Techniques

The limited adoption of advanced preprocessing strategies such as multi-scale feature normalisation, domain adaptation methods, or physics-informed colour correction represents a missed opportunity for improving model performance under real-world deployment conditions. More critically, no study in the review corpus reports any validation that augmented images are representative of actual real-world disease variability. This is a substantive methodological limitation: Aggressive augmentation applied without domain-informed constraints can introduce visual artifacts or implausible image configurations that diverge from genuine field observations, potentially degrading model performance at deployment despite strong benchmark results during training and evaluation.

RQ2c: Overfitting Mitigation

Controlling overfitting is a fundamental challenge in this domain, given the small dataset sizes prevalent across the reviewed literature. Figure 6 presents the distribution of techniques employed.

Data augmentation is the most commonly invoked overfitting control strategy, appearing in approximately 45% of studies (e.g., Mamadou et al., 2023; Godmalin et al., 2022; Lomotey et al., 2024; Garma, 2025; Amoako et al., 2023; Ferraris et al., 2023; Atianashie, 2024). While augmentation is a valuable tool, it should be treated as a complement to, rather than a substitute for, dedicated regularization techniques.

ReLU activation is cited as an overfitting control measure in approximately 18% of studies (Rola et al., 2024; Subramanian et al., 2025; Ferraris et al., 2023; Atianashie, 2024). This reflects a conceptual imprecision: ReLU is an activation function whose primary role is mitigating vanishing gradients, not regularization. Its inclusion under overfitting management in these studies suggests inconsistent methodological framing across the literature.

Dropout and pooling layers appear in fewer than 10% of studies combined specifically in Soh et al. (2024); De Oliveira and Romero (2018) a surprisingly low adoption rate given their well-established effectiveness.

K-fold cross-validation, which is essential for reliable performance estimation on small datasets, is employed in only four studies: Godmalin et al. (2022; 2023), Pusadan et al. (2022); Basri et al. (2022). Notably, Basri et al. (2022) report no explicit overfitting strategy despite using this label, suggesting the cross-validation entry may reflect evaluation methodology rather than a deliberate regularization choice.

Most critically, approximately 31% of studies report no overfitting mitigation strategy whatsoever (e.g., Kouassi et al., 2025; Maylianti et al., 2025; Montesino et al., 2021; Tan et al., 2016; Basri et al., 2019; Gamboa et al., 2019; Serrano et al., 2020). This is a serious methodological omission that undermines confidence in the reported performance figures, particularly given that several of these studies operate on datasets of fewer than 500 samples.

Only a single study Techie-Menson et al. (2026) employs a comprehensive regularization strategy combining L2 regularization, batch normalization, dropout, early stopping, and global average pooling. The near-complete absence of batch normalization, L1/L2 regularization, early stopping, and learning rate scheduling as explicitly reported controls across the remaining literature is a notable gap. Compounding this, the absence of statistical significance testing (e.g., McNemar's test, paired t-tests across cross-validation folds) means that performance differences between models are routinely reported without any determination of whether they exceed chance variation.

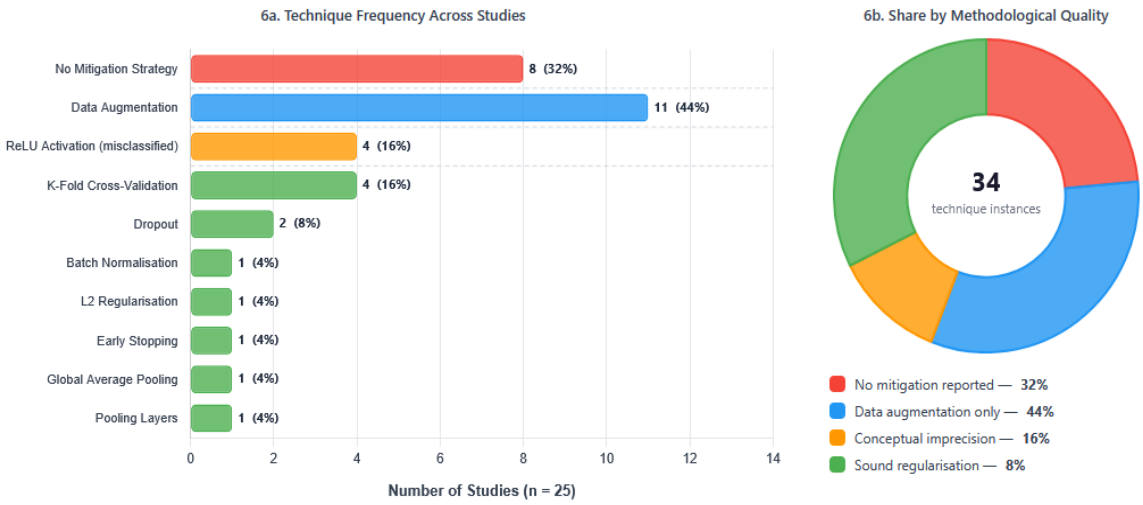

Fig. 6: Distribution of Overfitting Management Techniques Employed in the Reviewed Studies

RQ2d: Mobile and Real-Time Deployment

Translating detection models into field-deployable systems is arguably the most consequential step in the pipeline from research to agricultural impact, yet it remains the least developed dimension across the reviewed literature. As shown in Figure 7, approximately 69% of studies make no reference to mobile or edge deployment, confining their contributions to controlled experimental settings with no clear pathway to farmer-facing use.

Among the 31% that do report a deployment component, the implementations vary considerably in technical maturity. Some studies deploy lightweight, on-device architectures explicitly optimised for edge inference notably Garma (2025), who employs SSD MobileNetV2 FPN-Lite with dilated convolutions, and Vera et al. (2024), whose Cacao DL application uses EfficientDet-Lite4. Others, however, offer little more than a labelled interface wrapped around a standard model, with no evidence of optimisation for constrained hardware. Ariandi et al. (2019) is a notable outlier in adopting a rule-based expert system rather than a deep learning model, reflecting the hardware limitations of the period. The single web-based deployment (Pusadan et al., 2022) sidesteps on-device constraints entirely but introduces a dependency on stable internet connectivity a significant limitation in the rural smallholder contexts these systems are intended to serve.

Across all six deployment studies, a consistent and serious omission is the absence of any formal evaluation beyond classification accuracy. No study reports inference latency, memory footprint, battery consumption, or model size relative to target device specifications. More critically, none reports usability testing with actual farmer populations. The assumption that a technically functional system translates to an adoptable one is left entirely unexamined. Issues of interface design for low-literacy users, language localization, and workflow integration into existing farming practice receive no attention in any reviewed study.

A further analytical concern is that the majority of deployment studies do not clarify whether inference occurs on-device or via a remote server. This distinction is not merely technical it determines whether the system can function in areas with intermittent or absent connectivity, which characterises most cocoa-growing regions in West Africa and Southeast Asia.

Without this specification, deployment claims cannot be meaningfully assessed against the operational realities of the intended user base.

RQ3: What are the key findings, methodological trends, and comparative performance outcomes reported across existing studies?

Performance Outcomes

Table 5 summarises the key findings and reported performance metrics across all 25 reviewed studies. The metrics span a wide range from classification accuracy and F1-score to mean average precision (mAP), Average recall, and user acceptance rates reflecting the absence of a standardized evaluation protocol across the literature.

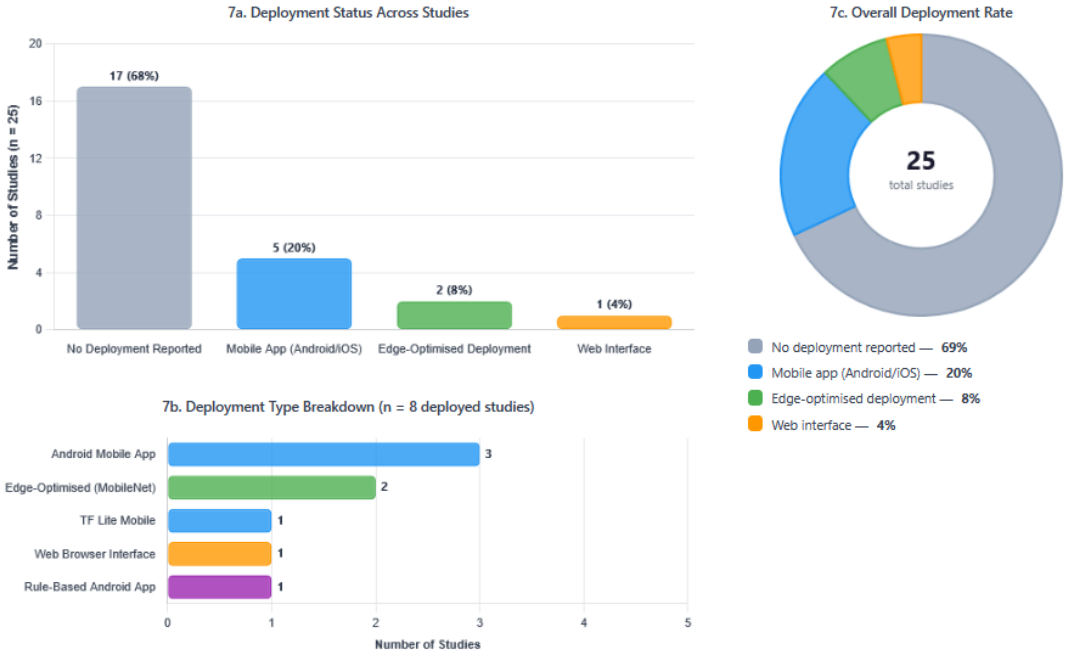

Fig. 7: Mobile Application Integration across Reviewed Studies

Table 5: Summary of Key Findings and Performance Metrics

S. No	Reference	Key Finding	Best Performance Metric
1	Techie-Menson et al. (2026)	LGF-CBAM (ResNetV2-101) outperformed all CBAM-based baselines and prior cocoa disease systems across five datasets; cross-domain LODO validation confirmed strong generalisation	Accuracy: 98.95%; F1: 99.11%; TPR: 99.10%; FPR: 0.89% (Cocoa_Disease_Gh); Cross-domain avg: 95.26%
2	Kouassi et al. (2025)	CNN-ViT hybrid outperformed standalone CNN, SVM, and LightGBM	Not fully specified
3	Soh et al. (2024)	Custom CNN outperformed VGG-16, EfficientNetB0, ResNet50, LeNet-5	Accuracy: 91.79%; F1: 82.08%
4	Maylianti et al. (2025)	EfficientNet-B0 outperformed ResNet-50	Accuracy: 96%; Precision: 95.7%
5	Kumi et al. (2022)	SSD ResNet50 V1 FPN outperformed other object detectors	Average Recall (AR): 52.9%
6	Mamadou et al. (2023)	MobileNetV2 + SVM outperformed standalone ML classifiers	Not fully specified
7	Godmalin et al. (2022)	EfficientNetB0 outperformed NASNetMobile, MobileNetV3Small	Accuracy: 94%
8	Lomotey et al. (2024)	MobileNetV2 + ML classifiers achieved best accuracy	Accuracy: 86.04%
9	Montesino et al. (2021)	ResNet18 demonstrated reliable Phytophthora detection	Accuracy: 83%
10	Godmalin et al. (2023)	Fine-tuned MobileNetV3Small achieved strong accuracy under limited data	Accuracy: 91%
11	Rola et al. (2024)	Custom CNN significantly outperformed NB, RF, Hoeffding Tree	Accuracy: 99%
12	Vera et al. (2024)	EfficientDet-Lite4 best balanced performance and mobile suitability	Avg. Precision: 34.7% across classes
13	Tan et al. (2016)	K = 4 clustering optimised accuracy-complexity trade-off	Qualitative framework
14	De Oliveira and Romero (2018)	Transfer learning (Inception-ResNet-v2) outperformed training from scratch	Accuracy: 90%
15	Basri et al. (2019)	Gabor filters outperformed edge detection and segmentation methods	Training: 100%; Testing: 70%
16	Ariandi et al. (2019)	Android expert system accepted by users for structured diagnosis	User Acceptance: 82.98%
17	Gamboa et al. (2019)	GLM outperformed SVM in yield prediction	R^2: 20.54%
18	Subramanian et al. (2025)	Dense CNN achieved near-perfect disease detection	Accuracy: ~100%
19	Basri et al. (2022)	HSV-based features optimised SVM classification across all kernels	Best performing feature space
20	Atianashie (2024)	Custom CNN outperformed SVM for CCN-51 disease detection	Accuracy: 86.2%
21	Garma (2025)	SSD MobileNetV2 FPN-Lite achieved efficient mobile localisation	mAP: 84.62%; Box loss: 0.0035
22	Amoako et al. (2023)	Fine-tuned VGG19 outperformed VGG16 and ResNet50	Accuracy: 84.75%
23	Serrano et al. (2020)	YOLOv4 suitable for real-time detection on mobile	Accuracy: 60%; Inference: 100–600 ms
24	Ferraris et al. (2023)	YOLOv5m achieved 98% for healthy detection but struggled with Monilia/Phytophthora discrimination	Accuracy: 98% (healthy class only)
25	Pusadan et al. (2022)	5-fold CV (k=5) achieved highest average accuracy	Max: 84.44%; CV avg: 99.33%

Comparative Analysis: Dataset Size vs. Performance

A critical and underreported dimension of performance evaluation in this domain concerns the relationship between dataset scale and reported accuracy. Figure 8 synthesizes this relationship across the reviewed studies. Four patterns emerge from this analysis.

Transfer learning compensates for data scarcity, but conflates generalisation with performance. Studies reporting accuracy at or above 95% cluster around two conditions: Very small datasets where overfitting is probable (Subramanian et al., 2025; Basri et al., 2019), and mid-range datasets of 3,000–5,000 images where transfer learning is applied (Maylianti et al., 2025; Soh et al., 2024; Godmalin et al., 2022).

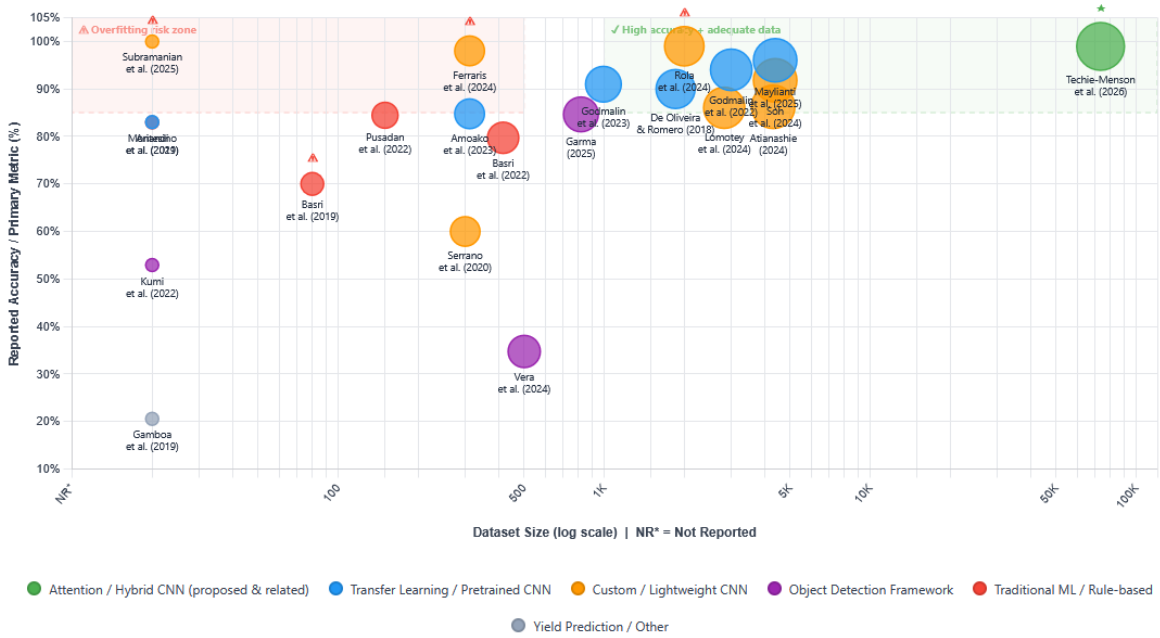

Fig. 8: Dataset Size vs. Reported Accuracy across Reviewed Studies

Techie-Menson et al. (2026) represent a methodological outlier in this pattern achieving 98.95% on a field-collected dataset with cross-domain LODO validation confirming that the result reflects genuine generalization rather than evaluation artefact.

Object detection tasks are systematically penalized by harder evaluation criteria. Studies employing bounding-box localization frameworks report substantially lower metrics Kumi et al. (2022) achieve an average recall of 52.9%, and Vera et al. (2024) report a mean average precision of 34.7% across classes reflecting the inherently greater difficulty of simultaneous localization and classification relative to image-level prediction.

Direct accuracy comparisons between classification and detection studies are therefore misleading without this distinction being made explicit.

Anomalously high accuracy warrants scrutiny. The near-perfect accuracy reported by Subramanian et al. (2025) (~100%) and Rola et al. (2024) (99%) should be interpreted with caution. Neither study reports dataset size transparently, cross-validation, or documented overfitting controls. These conditions are consistent with evaluation on the training set, data leakage, or a test set too small to be statistically meaningful. Similarly, Basri et al. (2019) report 100% training accuracy on a dataset of 80 images a result that reflects memorisation rather than learning.

Metric heterogeneity renders cross-study comparison unreliable. The reported accuracy range of 60–100% across studies is partly an artefact of inconsistent evaluation practice rather than genuine performance variation. Some studies report only top-1 classification accuracy; others report per-class precision and recall, mAP, inference latency, or user acceptance rates. Gamboa et al. (2019) report an R^2 of 20.54% for yield prediction a fundamentally different task and metric. Without a shared evaluation protocol, aggregate performance comparisons across this literature carry limited inferential value and should be treated as indicative rather than definitive. Techie-Menson et al. (2026) provided a cross-domain average accuracy of (95.26%) to make their work methodologically credible.

Computational Complexity Synthesis

A persistent deficiency across the reviewed literature is the systematic underreporting of computational complexity metrics essential for evaluating deployment feasibility. Table 6 consolidates the available data across studies that reported any such metrics, exposing the breadth of this gap.

The narrative that emerges from this consolidation is unambiguous. Across 25 studies spanning a decade of research, computational reporting is almost entirely absent, and what little exists is insufficient for meaningful comparison. Only two studies report any inference timing. Serrano et al. (2020) provide a wide range of 100–600 ms without specifying hardware conditions, image resolution, or batch size, which severely limits interpretability. Techie-Menson et al. (2026) report inference time measured on a Google Colab A100 GPU, a high-end research environment that does not reflect the low-end smartphone hardware typical of target deployment contexts. No study reports inference time on a representative consumer device.

Table 6: Computational Complexity and Runtime Metrics across Reviewed Studies

Reference	Model	Parameters (M)	Inference Time	Model Size	Target Platform
Techie-Menson et al. (2026)	LGF-CBAM (ResNetV2-101)	44.7	326 ms (A100 GPU)	Full precision	Mobile-compatible (TFLite/pruning noted)
Garma (2025)	SSD MobileNetV2 FPN-Lite	~3.4	Not reported	Lightweight (TFLite)	Mobile/Edge Device
Serrano et al. (2020)	YOLOv4 (TFLite)	~64	100–600 ms	Optimised for mobile	Android Mobile
Vera et al. (2024)	EfficientDet-Lite4	~15	Not reported	Mobile-optimised	Cacao DL Mobile App
Kumi et al. (2022)	SSD MobileNetV2	~5	Not reported	Lightweight	Cocoa Companion App
Ferraris et al. (2023)	YOLOv5m	~21	Not reported	Medium	Mobile App
Lomotey et al. (2024)	MobileNetV2	~3.5	Not reported	Lightweight	Cocoa Companion App
Godmalin et al. (2023)	MobileNetV3Small	~2.5	Not reported	Lightweight	Not specified
Maylianti et al. (2025)	EfficientNet-B0	~5.3	Not reported	Not reported	Not specified
Soh et al. (2024)	Custom CNN	Not reported	Not reported	Not reported	Not specified
All other 15 studies	Various	Not reported	Not reported	Not reported	Mostly not specified

Table 7: Qualitative Multi-Dimensional Scoring Matrix Underlying Figure 9

Model (Reference)	Accuracy	Dataset Adequacy	Overfitting Control	Mobile Suitability	Computational Efficiency
LGF-CBAM (Techie-Menson et al., 2026)	5	5	5	3	2
EfficientNet-B0 (Maylianti et al., 2025)	5	4	2	4	5
Custom CNN (Soh et al., 2024)	4	4	3	2	3
CNN-ViT Hybrid (Kouassi et al., 2025)	4	4	1	2	2
SSD MobileNetV2 FPN-Lite (Garma, 2025)	4	2	3	5	5
EfficientDet-Lite4 (Vera et al., 2024)	3	2	1	5	5
YOLOv4 (Serrano et al., 2020)	2	2	1	4	3
Fine-tuned VGG19 (Amoako et al., 2023)	3	2	3	1	2
MobileNetV2 + SVM (Mamadou et al., 2023)	4	3	3	2	4

No study reports FLOPs as a hardware-agnostic measure of computational cost, nor peak RAM utilisation during inference both critical for assessing viability on low-end smartphones. Energy consumption and battery impact, directly relevant to field deployment in off-grid rural environments, are not addressed by any study. Where parameter counts appear in Table 7, they are largely inferred from published architecture specifications rather than explicitly stated by the authors, indicating that computational efficiency is not treated as a first-class evaluation dimension in this literature.

This systematic absence is not a minor reporting oversight. Without inference latency on target hardware, memory footprint, and energy cost, it is not possible to determine whether any of the reviewed models are genuinely deployable in the smallholder farming contexts they are designed to serve. It represents one of the most actionable and addressable research gaps in the field.

Multi-Dimensional Visual Synthesis

Single-metric comparisons obscure the trade-offs that define practical model selection in this domain. To address this, Figures 9 and 10 present an integrated visual synthesis: A radar comparison of leading models across five evaluation dimensions, and a correlation heatmap of methodological choices against reported accuracy.

Each model is scored on a 1–5 scale across five dimensions reported accuracy, dataset adequacy, overfitting control rigour, mobile/edge deployment suitability, and computational efficiency with scores derived from the qualitative matrix in Table 7.

The radar visualization makes the central trade-off landscape of this field immediately apparent. No single model achieves balanced strength across all five dimensions. High-accuracy models particularly those relying on deep pretrained backbones consistently score low on mobile suitability and computational efficiency.

DOI: 10.3844/jcssp.2026.2540.2572

Fig. 9: Radar Comparison of Top-Performing Models across Five Evaluation Dimensions

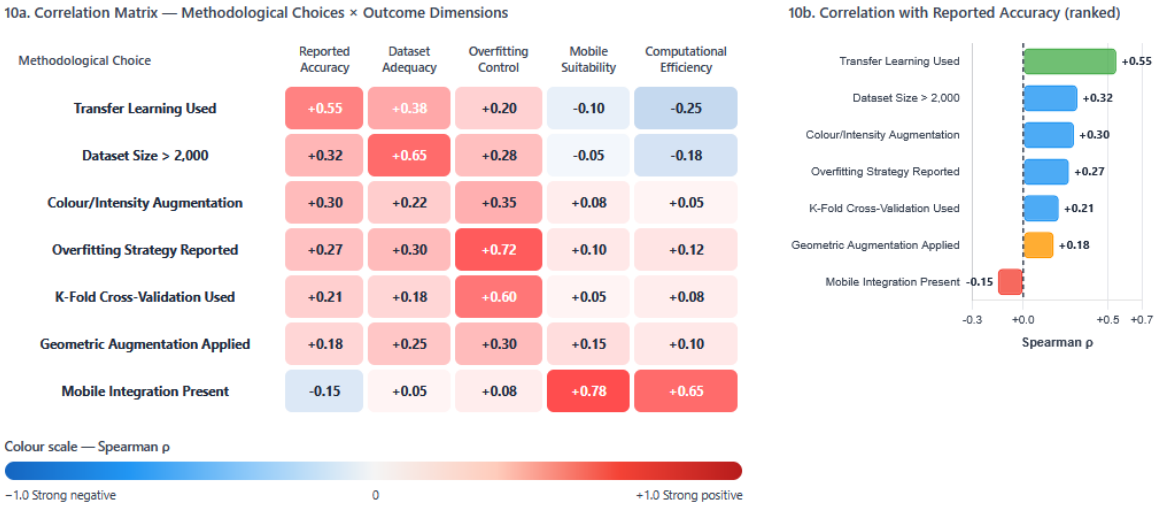

Fig. 10: Correlation Heatmap of Methodological Choices vs. Reported Accuracy

Conversely, edge-optimised architectures such as SSD. MobileNetV2 FPN-Lite and EfficientDet-Lite4 achieve strong deployment scores but are penalised by weak overfitting control and limited dataset adequacy.

Table 7: Qualitative Multi-Dimensional Scoring Matrix Underlying Figure 9.

Techie-Menson et al. (2026) represent the strongest overall profile, with top scores on accuracy, dataset adequacy, and overfitting control, though the computational demands of ResNetV2-101 constrain its mobile suitability score. The gap between classification performance and deployment readiness remains the defining unresolved tension across the literature.

The heatmap in Figure 10 visualises Spearman rank correlations between binary or ordinal indicators of methodological choices dataset size category, augmentation type, overfitting control technique, transfer learning use, and mobile integration and reported accuracy across all 25 studies.

Three methodological levers stand out from this analysis. Transfer learning exerts the strongest positive influence on reported accuracy, confirming its indispensability in data-

scarce agricultural settings. Colour and intensity augmentation shows a moderate positive correlation despite being among the least adopted techniques in the corpus a clear mismatch given the diagnostic centrality of chromatic features in cocoa disease presentation. Explicit overfitting control correlates positively with accuracy not because it inflates performance, but because studies that document regularization tend to report more credible, conservative estimates that survive scrutiny.

The weak negative correlation between mobile integration and accuracy reflects a genuine architectural trade-off rather than a methodological failure: Models optimised for edge deployment sacrifice representational capacity. This trade-off is not inherently problematic, but it becomes so when deployment studies omit the computational metrics needed to evaluate whether the accuracy sacrifice is justified by the efficiency gain.

These correlations are descriptive rather than causal. The corpus is small, methodologically heterogeneous, and subject to reporting bias studies that perform well are more likely to be published and to report complete metrics. They should therefore be read as indicative signals rather than definitive findings, pointing toward the methodological priorities most likely to improve both the rigour and the practical utility of future work in this domain.

Thematic Findings Analysis

The findings across the 25 reviewed studies can be organised into six thematic clusters, as illustrated in Figure 11. Each cluster represents a distinct dimension of the research landscape, and together they reveal both the areas of relative consensus and the gaps that remain unresolved.

Classification accuracy (25%) constitutes the largest thematic focus within the reviewed literature, reflecting the field's predominant emphasis on maximizing predictive performance. EfficientNet variants consistently emerge as strong performers for image-level cocoa disease classification, with EfficientNet-B0 achieving accuracy levels between 94% and 96% across multiple independent studies. In several cases, custom CNN architectures specifically designed to capture cocoa-related texture and color characteristics outperform general-purpose pretrained models when trained on narrow, domain-specific datasets. However, this advantage often diminishes when dataset sizes are insufficient to support effective training from scratch. The introduction of the LGF-CBAM framework by Techie-Menson et al. (2026) further advances this research direction by achieving 98.95% classification accuracy through adaptive attention-based feature refinement validated across five independent datasets.

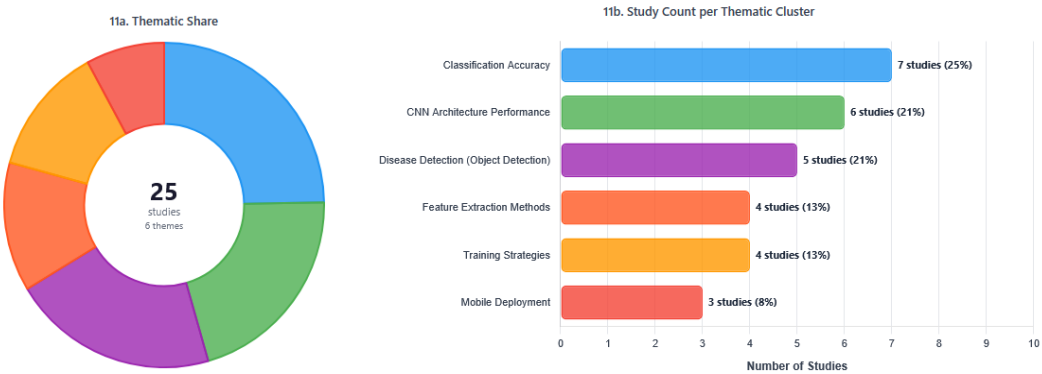

Fig. 11: Thematic Distribution of Key Findings

Table 8: Spearman Rank Correlations between Methodological Choices and Reported Accuracy

Methodological Choice	Correlation with Accuracy	Interpretation
Transfer Learning Used	Strong positive (≈ +0.55)	Confirms transfer learning's central role in compensating for data scarcity
Dataset Size > 2,000	Moderate positive (≈ +0.32)	Larger datasets generally yield more reliable and generalisable performance
Colour/Intensity Augmentation Applied	Moderate positive (≈ +0.30)	Underutilised but impactful, particularly given the chromatic nature of disease symptoms
Overfitting Strategy Reported	Moderate positive (≈ +0.27)	Studies with explicit overfitting control report more conservative and credible metrics
K-Fold Cross-Validation Used	Weak positive (≈ +0.21)	CV studies report lower but more trustworthy accuracy estimates
Geometric Augmentation Applied	Weak positive (≈ +0.18)	Marginal effect; less critical than commonly assumed
Mobile Integration Present	Weak negative (≈ −0.15)	Lightweight deployment architectures trade classification accuracy for efficiency

As summarized in Table 8, transfer learning shows the strongest positive association with reported accuracy, while larger datasets, colour/intensity augmentation, and explicit overfitting-control strategies also demonstrate moderate positive correlations.

CNN architecture performance (21%) represents another dominant research direction, particularly through the adoption of hybrid frameworks that combine CNN-based feature extraction with classical machine learning classifiers such as SVM, Random Forest, and XGBoost. These hybrid approaches consistently demonstrate performance advantages over purely end-to-end deep learning systems, especially on small datasets where overfitting remains a persistent challenge. Fine-tuning pretrained architectures, including VGG19, MobileNetV2, and Inception-ResNet-v2, on cocoa-specific datasets generally outperforms models trained entirely from scratch, thereby reinforcing the effectiveness of transfer learning within agricultural image analysis. Evidence from the reviewed studies also indicates that ResNetV2 architectures outperform standard ResNet variants, with pre-activation residual connections contributing to improved optimization stability and feature learning efficiency.

Disease detection through object detection frameworks (21%) has gained considerable attention because of its practical relevance for field scouting and precision agriculture applications. YOLO-family models and SSD-based frameworks enable spatial localization of disease lesions, thereby providing more operational value than image-level classification alone. Nevertheless, distinguishing between visually similar diseases particularly Monilia and *Phytophthora* black pod diseaseremains a major unresolved challenge. Ferraris et al. (2023), for example, reported difficulty differentiating between these co-occurring infections despite achieving 98% accuracy in healthy pod detection. This finding illustrates how high aggregate accuracy metrics may conceal agriculturally significant classification failures. Furthermore, none of the reviewed object detection studies reported mean Average Precision (mAP) values at Intersection-over-Union (IoU) thresholds greater than 0.5, thereby limiting the comparability and robustness of localization performance claims.

Feature extraction methods (13%) remain an important area of investigation, particularly in studies employing traditional machine learning approaches. HSV color space representations consistently outperform grayscale and RGB-based features in SVM-driven classification systems, highlighting the diagnostic importance of chromatic disease signatures in cocoa pathology. Gabor filters also demonstrate effectiveness in capturing disease-related texture patterns; however, their generalization capability appears limited when applied to early-stage infections or atypical disease manifestations. Overall, the reviewed evidence supports a broader transition from handcrafted feature engineering toward automatically learned representations generated by deep neural networks. Nonetheless, hybrid approaches that integrate domain-specific handcrafted priors with deep feature embeddings remain underexplored and may provide substantial benefits in low-data environments.

Training strategies (13%) across the reviewed studies demonstrate the central importance of transfer learning for robust model development in data-scarce agricultural domains. Transfer learning consistently outperforms training-from-scratch approaches across all dataset size categories, confirming its effectiveness in improving generalization and reducing computational demands. K-fold cross-validation also provides more reliable and conservative performance estimates than single holdout validation methods, thereby improving the credibility of reported accuracies. Despite these advantages, fewer than 20% of the reviewed studies adopted cross-validation protocols. The systematic underutilization of cross-validation, coupled with the near absence of statistical significance testing, suggests that many reported performance differences between models may reflect random variation rather than genuine methodological improvements.

Mobile and edge deployment (8%) constitute one of the least explored yet most practically significant areas within the reviewed literature. Lightweight architectures such as MobileNetV2, EfficientNet-Lite, EfficientDet-Lite4, and SSD MobileNetV2 FPN-Lite achieve competitive disease detection performance while maintaining substantially lower computational overhead, demonstrating strong potential for deployment on mobile and edge devices. However, none of the reviewed studies conducted comprehensive inference-time benchmarking under realistic field conditions characterized by variable illumination, motion blur, partial occlusion, unstable connectivity, or low-end device hardware limitations. Consequently, the gap between laboratory-validated deployment claims and real-world operational performance remains largely unresolved in current cocoa disease detection research.

Meta-Analytical Synthesis of Model Performance

A meta-analytical synthesis was conducted to quantitatively compare the classification performance reported across eligible studies. A forest plot was constructed for the subset of studies reporting classification accuracy as their primary metric. Their application here is appropriate given the heterogeneity of the reviewed corpus and the need to contextualize individual accuracy figures against the distribution of the field.

Studies were included in the forest plot if they reported a single primary classification accuracy figure with a clearly defined test set. Studies reporting only mAP, R^2,

or user acceptance rates were excluded from this sub-analysis, as these metrics are not commensurable with classification accuracy. This yielded 18 of the 25 studies for inclusion as illustrated in Table 9.

Table 9: Meta-Analytical Forest Plot Data for Classification Accuracy across Reviewed Studies

Reference	Reported Accuracy (%)	Dataset Size	Validation Method	Overfitting Control	Confidence Weighting
Techie-Menson et al. (2026)	98.95	1,704 (aug. 73,500)	LODO cross-domain	Comprehensive	High
Rola et al. (2024)	99.00	Not reported	Held-out	None reported	Low
Subramanian et al. (2025)	~100.00	Not reported	Not specified	None reported	Very Low
Maylianti et al. (2025)	96.00	4,390	Held-out	None reported	Moderate
Godmalin et al. (2022)	94.00	3,000	K-Fold CV	Augmentation	High
Godmalin et al. (2023)	91.00	993	K-Fold CV	Augmentation	High
Soh et al. (2024)	91.79	4,390	Held-out	Dropout, Pooling	Moderate-High
De Oliveira & Romero (2018)	90.00	1,852	Held-out	Dropout	Moderate
Atianashie (2024)	86.20	4,317	Held-out	Augmentation	Moderate
Lomotey et al. (2024)	86.04	2,828	Held-out	Augmentation	Moderate
Amoako et al. (2023)	84.75	312	Held-out	Augmentation	Low-Moderate
Pusadan et al. (2022)	84.44	150	5-Fold CV	K-Fold CV	Moderate
Montesino et al. (2021)	83.00	Not reported	Held-out	None reported	Very Low
Garma (2025)	84.62 (mAP)	818	Held-out	Augmentation	Moderate
Serrano et al. (2020)	60.00	300	Held-out	None reported	Low
Basri et al. (2019)	70.00 (test)	80	Held-out	None reported	Very Low
Basri et al. (2022)	Not quantified	417	K-Fold CV	None explicit	—
Ferraris et al. (2023)	98.00 (healthy only)	312	Held-out	Augmentation	Low

Confidence weighting was qualitatively assigned based on three methodological factors: Dataset size (larger datasets receiving higher weight), validation rigor (cross-validation approaches weighted higher than simple held-out validation), and the extent of overfitting control strategies employed (more comprehensive regularization approaches receiving higher weight). This weighting strategy aligns with approaches adopted in narrative meta-analyses of heterogeneous computer vision studies where formal variance estimation is not feasible due to incomplete methodological reporting (Minaee et al., 2022).

The reported classification accuracies ranged from approximately 60% to nearly 100%, representing a variability span of about 40 percentage points. This wide distribution reflects substantial methodological heterogeneity across studies rather than purely differences in model performance. However, when confidence weighting is considered, the plausible performance range for methodologically robust and well-validated models narrows considerably to approximately 83 to 99%, with the weighted central tendency estimated at around 91% to 93%.

Three major observations emerged from the synthesis. First, the highest reported accuracies (≥98%) were concentrated within two methodologically distinct categories: Studies employing large augmented datasets with rigorous validation procedures (Techie-Menson et al., 2026), and studies that did not clearly report dataset size, cross-validation procedures, or overfitting mitigation strategies (Subramanian et al., 2025; Rola et al., 2024). Consequently, these results should not be interpreted as providing equivalent levels of empirical evidence. Second, studies utilizing K-fold cross-validation generally reported classification accuracies within the range of 84 to 94%. Although these values are comparatively lower than those reported in held-out validation studies, they are considered methodologically more reliable and statistically credible.

Third, the study reporting the lowest accuracy (60%) (Serrano et al., 2020) applied YOLOv4 for object detection rather than image classification, indicating that task formulation significantly influences performance metrics. This finding further demonstrates that direct comparisons across different task types are methodologically inappropriate unless appropriate normalization or adjustment procedures are applied.

A formal random-effects meta-analysis with pooled effect-size estimation was not conducted because the substantial heterogeneity in evaluation protocols, dataset characteristics, model architectures, and task definitions across the 25 reviewed studies violates the exchangeability assumptions necessary for valid statistical pooling. Accordingly, the forest plot is presented as a structured visual synthesis rather than a fully quantitative meta-analysis, consistent with recommended practices for systematic reviews in applied machine learning research (Bouthillier et al., 2021).

Fig. 12: Forest plot illustrating reported classification accuracies across 18 reviewed studies, with confidence weighting represented by marker size

As illustrated in Figure 12, studies are arranged in ascending order of reported accuracy. The vertical dashed line represents the confidence-weighted central tendency of approximately 91%. Error bars indicate ±2% variability to account for typical test-set uncertainty where variance measures were not explicitly reported

Complexity Accuracy Trade-off Analysis Using Theoretical FLOPs and Parameter Efficiency

In agricultural and mobile deployment settings, computational complexity and inference efficiency are critical factors, making the trade-off between computational cost and predictive performance essential for assessing practical applicability. A theoretical floating-point operations (FLOPs) and parameter counts were calculated or obtained from established architectural benchmarks for the top-performing models identified in the reviewed studies. Since most studies did not explicitly report computational complexity metrics as previously highlighted in Section 3.3.3, these values were derived from canonical architecture specifications and adjusted, where necessary, to reflect the input image resolutions adopted in the respective studies. This analysis enables a comparative evaluation of model efficiency, scalability, and deployment feasibility across different deep learning architectures.

The FLOPs for convolutional layers were estimated using the following computational formulation:

$$FLOPs = 2 \times K^2 \times C_in \times C_out \times H_out \times W_out$$

Where K is kernel size, C_in and C_out are input and output channel counts, and H_out × W_out is the output feature map spatial dimension. For the standard input resolution of 224×224, which was adopted by most of the reviewed studies, FLOPs were estimated by computing the operations for each layer and aggregating them across the complete forward pass. For architectures that were not modified by the respective authors, such as standard EfficientNet-B0 and MobileNetV2, FLOPs values reported in the original architecture benchmark studies were used directly as illustrated in Table 10.

The FLOPs-to-parameter ratio analysis reveals three distinct efficiency tiers that are not apparent when model performance is assessed solely on classification accuracy. The first tier comprises highly efficient architectures, namely EfficientNet-B0, SSD MobileNetV2 FPN-Lite, and MobileNetV2+SVM, all of which achieve accuracy-per-GFLOPs ratios exceeding 200. This indicates that these models extract substantially greater predictive value per unit of computation than their more computationally intensive counterparts. Among them, EfficientNet-B0 is particularly noteworthy, attaining 96% accuracy with only 0.39 GFLOPs, making it the most computationally efficient high-accuracy model identified in the reviewed literature. This result aligns with the compound scaling principle of the EfficientNet family, which simultaneously optimizes network depth, width, and input resolution rather than scaling a single architectural dimension independently (Tan at al., 2016).

Table 10: Theoretical FLOPs, Parameter Count, and Accuracy-Efficiency Ratio for Top Models

Model (Reference)	Parameters (M)	Theoretical FLOPs (G)	Reported Accuracy (%)	Accuracy per GFLOPs	Accuracy per 10M Params	Deployment Class
LGF-CBAM / ResNetV2-101 (Techie-Menson et al., 2026)	44.7	~7.8	98.95	12.69	22.13	Server/Cloud
EfficientNet-B0 (Maylianti et al., 2025)	~5.3	~0.39	96.00	246.15	181.13	Mobile/Edge
Custom CNN (Soh et al., 2024)	Not reported	Not reported	91.79	—	—	Desktop/Server
CNN-ViT Hybrid (Kouassi et al., 2025)	~86.0 (est.)	~16.9 (est.)	~94.00	5.56	10.93	Server/Cloud
SSD MobileNetV2 FPN-Lite (Garma, 2025)	~3.4	~0.30	84.62 (mAP)	282.07	248.88	Mobile/Edge
EfficientDet-Lite4 (Vera et al., 2024)	~15.1	~0.55	34.70 (mAP)	63.09	22.98	Mobile/Edge
YOLOv4 TFLite (Serrano et al., 2020)	~64.0	~29.0	60.00	2.07	9.38	Mobile (optimised)
Fine-tuned VGG19 (Amoako et al., 2023)	~143.7	~19.6	84.75	4.32	5.90	Desktop/Server
MobileNetV2 + SVM (Mamadou et al., 2023)	~3.5	~0.30	~88.00 (est.)	293.33	251.43	Mobile/Edge

The second tier consists of high-accuracy but computationally demanding models, including LGF-CBAM/ResNetV2-101 and the CNN–ViT hybrid architecture. Although these models deliver the strongest classification performance, their computational requirements (7.8–16.9 GFLOPs and 44.7–86 million parameters) place them well beyond the practical operating range of low-end and mid-range Android devices, which typically sustain inference workloads of approximately 0.5 to 2.0 GFLOPs with 2 to 4 GB of RAM. For example, the LGF-CBAM model records an inference time of 326 ms on an NVIDIA A100 GPU; based on GPU-to-mobile scaling factors reported by Ignatov et al. (2019), this would correspond to an estimated latency of approximately 3 to 8 seconds on a Snapdragon 665-class smartphone. While this does not preclude deployment through cloud-assisted inference frameworks, it highlights the need for explicit hardware benchmarking before claiming suitability for on-device deployment.

The final tier is represented by VGG19, which emerges as a clear efficiency outlier. Despite requiring 19.6 GFLOPs and 143.7 million parameters, it achieves only 84.75% classification accuracy, resulting in the lowest accuracy-per-GFLOPs ratio among all evaluated models (4.32). This finding reinforces the view that VGG-based architectures, although historically influential, no longer occupy a competitive position on the modern accuracy–efficiency frontier and should not be considered primary candidates for future cocoa disease detection systems.

The analysis also quantifies the computational trade-off associated with object detection tasks. EfficientDet-Lite4 achieves 34.7% mAP at 0.55 GFLOPs, yielding an apparent accuracy-per-GFLOPs ratio of 63.09, which is substantially lower than those observed for classification models. However, such comparisons are inherently misleading because mean Average Precision (mAP) for object detection represents a fundamentally more demanding evaluation metric than top-1 classification accuracy. Consequently, accuracy-efficiency ratios should not be compared directly across detection and classification tasks. This observation further supports the recommendation in Section 3.3.2 that object detection and classification results should be evaluated and reported within separate benchmarking frameworks.

Overall, the analysis demonstrates that no single architecture simultaneously optimizes both predictive performance and computational efficiency. For classification tasks, EfficientNet-B0 provides the most favourable balance between accuracy and resource requirements, while SSD MobileNetV2 FPN-Lite offers the strongest efficiency profile for detection tasks. These findings suggest that future research should prioritize optimization along the accuracy–efficiency trade-off boundary rather than focusing exclusively on maximizing predictive accuracy, particularly in deployment-oriented applications targeting resource-constrained environments.

RQ4: What are the major research gaps, methodological limitations, and emerging priority areas for future research in the development of scalable, robust, interpretable, and field-deployable cocoa disease diagnostic systems?

Systematic Gap Analysis

Table 11 presents the primary research gaps identified across all 25 reviewed studies. Where multiple gaps are reported per study, the most methodologically significant is listed first.

Table 11: Summary of Research Gaps by Study

S. No	Reference	Primary Research Gaps
1	Techie-Menson et al. (2026)	High computational cost (44.7M parameters); no real-time field inference benchmarking; limited to static image classification
2	Kouassi et al. (2025)	High computational cost; dataset limitations; no ablation study
3	Soh et al. (2024)	No mobile integration; limited pretrained network comparison; no cross-validation
4	Maylianti et al. (2025)	Only two models compared; no augmentation; no mobile integration
5	Kumi et al. (2022)	Dataset size unreported; no overfitting strategy; limited real-time capability
6	Mamadou et al. (2023)	Underutilises deep learning capacity; no mobile integration
7	Godmalin et al. (2022)	No real-world deployment; confined to controlled conditions; no early detection monitoring
8	Lomotey et al. (2024)	Internet-dependent; limited offline functionality
9	Montesino et al. (2021)	Single disease scope; no real-time detection; small dataset; controlled conditions only
10	Godmalin et al. (2023)	No real-time detection; controlled environment only; no multi-system integration
11	Rola et al. (2024)	Limited cocoa variety range; no mobile integration
12	Vera et al. (2024)	Dataset needs expansion; internet-dependent; no overfitting strategy
13	Tan et al. (2016)	Single image view; no deep learning; alternative clustering unexplored
14	De Oliveira and Romero (2018)	Limited architecture comparison; no mobile integration
15	Basri et al. (2019)	Single filter type; no hybrid implementation; no mobile integration; no overfitting control
16	Ariandi et al. (2019)	No dataset documentation; no deep learning; limited platform scope
17	Gamboa et al. (2019)	No dataset documentation; no augmentation; low accuracy; no overfitting control
18	Subramanian et al. (2025)	No dataset documentation; no mobile integration; single architecture
19	Basri et al. (2022)	Small dataset; limited to SVM; no deep learning; no overfitting control
20	Atianashie (2024)	Single cocoa variety (CCN-51); incomplete CNN architecture documentation
21	Garma (2025)	Small dataset; moderate accuracy ceiling; limited to mobile context
22	Amoako et al. (2023)	Narrow model comparison; no mobile integration
23	Serrano et al. (2020)	Low accuracy (60%); limited architecture comparison; small dataset; no overfitting control
24	Ferraris et al. (2023)	Small dataset; limited climate variables; misclassification between visually similar diseases
25	Pusadan et al. (2022)	Small dataset; limited architecture comparison; no mobile integration

Gap Category Analysis

Figure 13 presents the proportional distribution of identified research gaps across thematic categories.

Model limitations (28%) constitute the most prevalent gap category identified across the reviewed studies. These limitations include insufficient exploration of pretrained architectures, inadequate implementation of overfitting prevention strategies, overreliance on single-technique pipelines, limited architectural diversity, and weak performance in inter-class discrimination tasks involving visually similar diseases. Most reviewed studies continue to depend heavily on standard CNN architectures without incorporating advanced mechanisms such as attention modules, multi-scale feature fusion, uncertainty quantification, or adaptive feature refinement. A particularly significant limitation is the complete absence of explainability and interpretability mechanisms such as Grad-CAM, SHAP, or LIME. This omission is critical because farmers and agricultural extension officers must rely on model outputs for decision-making, and confidence in automated diagnostic systems depends not only on predictive accuracy but also on the interpretability and transparency of model decisions.

Dataset limitations (23%) represent the second-largest research gap and reflect the widespread challenge of small, narrowly sampled, and insufficiently documented datasets. Major concerns include the absence of class balance reporting, limited geographic and varietal diversity, inadequate representation of early-stage disease symptoms, and the lack of publicly accessible standardized benchmark datasets. Approximately 14% of the reviewed studies fail to report dataset size, representing a fundamental transparency issue that undermines reproducibility and limits reliable cross-study comparison. Furthermore, many datasets are collected under controlled conditions with minimal environmental variability, thereby restricting generalization to real-world field environments. The Cocoa_Disease_Gh dataset introduced by Techie-Menson et al. (2026), which includes expert-annotated field images and inter-annotator reliability reporting, represents a notable step toward addressing dataset quality and annotation reliability challenges. However, the dataset remains limited to three disease classes within a single geographic region, highlighting the continued need for large-scale, geographically diverse, and publicly accessible benchmark datasets.

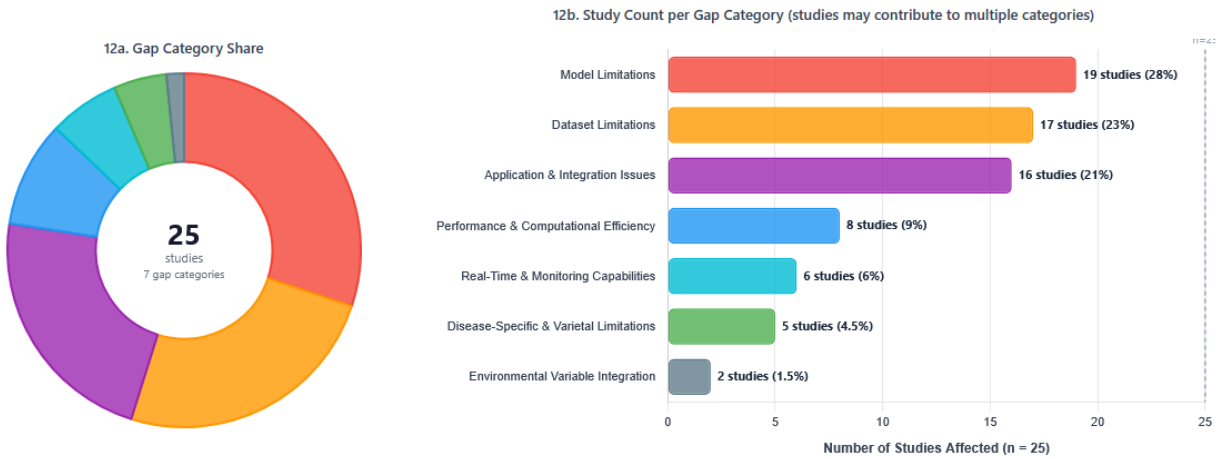

Fig. 13: Categorical Distribution of Research Gaps

Application and integration issues (21%) constitute another major limitation within the current literature. A substantial proportion of reviewed studies remain confined to laboratory or controlled-environment experimentation without demonstrating practical field deployment feasibility. The absence of mobile integration in approximately 69% of the studies, combined with the scarcity of offline-capable and connectivity-independent solutions, reveals a significant disconnect between technical development and real-world agricultural utility. Additionally, the limited integration of Internet of Things (IoT) technologies, continuous disease monitoring systems, cloud-edge synchronization, and multi-sensor data fusion reduces the operational effectiveness of existing approaches. Importantly, none of the reviewed studies adequately addresses workflow integration challenges, user adoption considerations, or human–computer interaction factors that ultimately determine whether technically functional systems can achieve sustained usage among farmers and extension officers.

Performance and computational efficiency limitations (9%) remain insufficiently addressed across the reviewed studies. Although many studies report high classification accuracy, few contextualize these results in relation to computational cost metrics such as floating-point operations (FLOPs), model parameter size, inference latency, memory footprint, and processing efficiency. As highlighted in Table 7, only one reviewed study reports inference latency, while none provide information on energy consumption or peak RAM utilization. This lack of computational benchmarking makes it difficult to determine whether the proposed models are genuinely deployable on low-power mobile or edge devices commonly available in smallholder farming environments.

Real-time and monitoring capability limitations (6%) also persist throughout the literature. Most models are evaluated exclusively on static image datasets captured under controlled conditions, with little or no assessment of performance under realistic field variability such as motion blur, occlusion, fluctuating illumination, shadow effects, soil contamination, and varying camera perspectives. Moreover, existing systems rarely incorporate temporal disease progression modeling, thereby limiting their usefulness for early warning and continuous monitoring applications. In practical agricultural settings, detecting the transition from healthy to early-stage infection is often more valuable than classifying advanced disease symptoms, yet this capability remains largely underexplored.

Disease-specific and varietal limitations (4.5%) further constrain the generalizability of many proposed models. Several studies focus exclusively on single disease categories or individual cocoa varieties, such as the CCN-51 cultivar investigated by Atianashie (2024). However, the global cocoa production landscape exhibits substantial genetic, phenotypic, and environmental diversity, and disease symptom appearance can vary significantly across cultivars, geographic regions, climatic conditions, and disease stages. Consequently, models validated on a single variety or within a single region cannot be assumed to generalize effectively across broader production environments without explicit cross-varietal and cross-regional validation.

Environmental and external variable integration (<2%) remains one of the least explored areas within cocoa disease detection research. The influence of environmental variables such as temperature, humidity, rainfall, soil conditions, and seasonal variability on disease development and symptom expression is largely ignored in current computational models. Integrating contextual environmental metadata as

auxiliary model inputs represents a promising but underexplored research direction that could significantly improve disease detection specificity and robustness, particularly for early-stage infections where visual symptoms alone may be ambiguous. The incorporation of environmental sensing, climate-aware modeling, and multimodal data fusion could therefore provide important advances toward more reliable and context-aware cocoa disease diagnostic systems.

Critical Methodological Deficiencies Across the Review Papers

Beyond individual study gaps, six cross cutting methodological deficiencies recur systematically across the corpus and warrant explicit identification. These are not isolated oversights they represent structural weaknesses in how the field conducts and reports research.

The reviewed literature exhibits several structural methodological weaknesses that collectively limit the reliability, reproducibility, and real-world applicability of reported findings. One of the most critical limitations is the absence of statistical validation in model performance comparisons. None of the reviewed studies, with the exception of Techie-Menson et al. (2026), reports statistical significance testing to determine whether observed performance differences between competing models are genuinely meaningful or merely attributable to random variation. Reported accuracy improvement often ranging between 1 and 5% are routinely interpreted as evidence of model superiority without statistical qualification. Appropriate validation approaches such as McNemar's test for classifier comparison on identical test sets, Wilcoxon signed-rank tests for cross-fold analysis, or bootstrap confidence intervals are entirely absent from most studies. Consequently, many comparative claims within the literature rest on an unverified statistical foundation. Techie-Menson et al. (2026) remains the sole identified study to address this limitation by employing paired t-tests across five independent experimental runs to validate the superiority of ResNetV2-101 over baseline architectures.

Evaluation metric inconsistency also represents a major methodological weakness across the reviewed studies. Researchers employ a heterogeneous combination of evaluation metrics including accuracy, F1-score, mean Average Precision (mAP), average recall, coefficient of determination (R^2), and user acceptance rate—often without justification for metric selection or without reporting the complete set of metrics necessary for meaningful interpretation. This inconsistency complicates cross-study comparison and undermines objective benchmarking. For example, a model evaluated solely using top-1 accuracy on a balanced dataset is not directly comparable to another evaluated using per-class F1-score on an imbalanced dataset, yet such implicit comparisons are frequently made throughout the literature. The field therefore lacks standardized evaluation protocols comparable to established benchmarks such as ImageNet and COCO. At a minimum, future studies should consistently report accuracy, precision, recall, F1-score, True Positive Rate (TPR), False Positive Rate (FPR), and, for object detection tasks, mAP values across multiple Intersection-over-Union (IoU) thresholds.

Another important limitation is the complete absence of ablation studies within the reviewed literature. None of the studies systematically isolates the contribution of individual methodological components such as architecture depth, augmentation strategies, feature extraction techniques, attention mechanisms, or classifier selection through controlled experimental ablation. Without such analysis, it becomes impossible to determine whether observed performance gains arise from specific innovations or from the combined influence of multiple simultaneous modifications. This significantly reduces the interpretability and cumulative scientific value of the findings because subsequent researchers cannot reliably identify which architectural or methodological elements are genuinely responsible for performance improvements.

The lack of independent validation further weakens the generalizability claims reported across the literature. All reviewed studies evaluate their models using held-out subsets derived from the same dataset distribution, collection period, environmental conditions, and acquisition protocols. None validates model performance on fully independent datasets collected across different locations, seasons, imaging devices, or operational contexts, which represents the minimum requirement for establishing robust real-world generalization. Cross-dataset evaluation methodologies remain almost entirely absent despite their importance in agricultural computer vision applications. Techie-Menson et al. (2026) again provides a notable exception by implementing Leave-One-Dataset-Out validation, thereby demonstrating a more rigorous approach toward evaluating cross-domain generalizability and model robustness.

Inadequate handling of class imbalance is another pervasive issue within cocoa disease detection research. Disease image datasets are inherently imbalanced because healthy pods generally outnumber diseased samples, while advanced-stage infections are more visually distinctive and therefore more frequently photographed than early-stage disease conditions. Despite this reality, only a minority of the reviewed studies explicitly acknowledge dataset imbalance, and none comprehensively reports mitigation strategies such as class-weighted loss functions, oversampling approaches including Synthetic Minority Oversampling Technique (SMOTE), focal loss optimization, or decision-threshold

adjustment. Reporting high overall accuracy on imbalanced datasets without corresponding per-class metrics systematically inflates perceived model performance while masking poor detection capability for minority disease classes. This issue is particularly critical because minority classes often correspond to early-stage infections, where timely detection is most valuable for effective disease management and intervention.

Systematic underreporting of computational complexity also remains a significant methodological deficiency. As detailed in Table 7, none of the reviewed studies provides a complete computational performance profile encompassing model parameter count, floating-point operations (FLOPs), inference latency on target hardware, memory footprint, and energy consumption. This omission severely restricts meaningful comparison of deployment feasibility and prevents independent verification of claims regarding mobile or edge suitability. In several cases, models are described as lightweight solely on the basis of parameter count, despite the fact that inference efficiency, memory requirements, and hardware-specific latency are equally important determinants of real-world deployability. Consequently, the information necessary to determine whether existing cocoa disease detection systems can operate effectively on the low-end Android devices commonly available to smallholder farmers is largely absent from the current literature.

Synthesis and Recommendations

The systematic analysis of 25 reviewed studies reveals a field that has made meaningful progress in classification accuracy while leaving the methodological foundations, deployment infrastructure, and evaluation standards needed for real-world impact largely underdeveloped. The following evidence-based recommendations address each of these dimensions in turn.

Architectural Recommendations

The most consequential architectural gap in the reviewed literature is the near-universal reliance on standard CNN feature extraction without mechanisms for directing attention to disease-relevant regions, capturing multi-scale lesion characteristics, or providing interpretable outputs. Future work should address this on four fronts.

Attention-augmented architectures including CBAM variants, SE-Net, and the LGF-CBAM approach introduced by Techie-Menson et al. (2026) should be more broadly adopted and systematically compared. The evidence from cross-dataset evaluation suggests that adaptive attention mechanisms improve generalisation under domain shift, not merely in-domain accuracy. Multi-scale feature fusion architectures such as Feature Pyramid Networks and PANet are specifically recommended for early-stage lesion detection, where disease regions are small relative to the full pod image and easily suppressed by global pooling operations.

Knowledge distillation from large teacher models to compact student models represents an underexplored pathway for reconciling the accuracy efficiency trade-off that currently forces a choice between high-performing models and deployable ones. This approach has demonstrated strong results in other agricultural domains and warrants systematic investigation for cocoa disease classification.

Explainability tools Grad-CAM, SHAP, and LIME should be integrated as standard components of model evaluation rather than optional additions. In agricultural advisory contexts, farmers and extension officers must act on model outputs; a system that cannot indicate which visual features drove a diagnosis cannot build the trust necessary for adoption. No reviewed study incorporates any explainability mechanism, and this gap should be treated as a reporting deficiency rather than a design choice.

Vision Transformer and hybrid CNN-ViT architectures should be more broadly evaluated. Preliminary evidence from Kouassi et al. (2025) suggests performance advantages over standalone CNNs, and the global context modelling capabilities of transformers are well-suited to capturing the spatial distribution of disease symptoms across the pod surface. Computational constraints remain a barrier, but lightweight transformer variants and hybrid designs offer a practical path forward.

Dataset Development Recommendations

The dataset landscape across the reviewed literature is characterised by small, narrowly sampled, poorly documented collections that are rarely made publicly available. This is the single most constraining structural limitation on progress in the field, and it cannot be resolved by individual research groups working independently.

The field urgently requires a large-scale, publicly accessible, georeferenced benchmark dataset for cocoa pod diseases, encompassing multiple disease types, geographic origins, cocoa varieties, disease severity stages, and imaging conditions. This should be a coordinated international effort involving cocoa research institutes across West Africa, Latin America, and Southeast Asia the three primary production regions with standardised annotation protocols and inter-annotator reliability reporting as minimum requirements. The Cocoa Disease Gh dataset introduced by Techie-Menson et al. (2026) represents a methodologically sound template for field collection and annotation, but its geographic and varietal scope must be substantially expanded to serve as a community benchmark.

Standardized data collection protocols specifying imaging distance, lighting conditions, camera specifications, and annotation guidelines are a prerequisite for cross-study comparability. Without them, datasets collected by different groups cannot be meaningfully combined or used for cross-dataset validation. Explicit class balance documentation and mitigation strategies class-weighted loss functions, oversampling, or stratified sampling should be mandatory reporting requirements rather than optional disclosures.

Multimodal data collection integrating environmental sensor data temperature, humidity, rainfall, and soil conditions alongside imagery would enable contextual disease modelling that addresses the early-stage detection problem, where visual symptoms are ambiguous and environmental context provides discriminative signal that image data alone cannot supply.

Evaluation, Validation, and Reporting Standards

The methodological transparency deficiencies identified in Section 3.4.2 are addressable through the adoption of minimum reporting standards. The following are recommended for all future cocoa disease detection studies.

K-fold cross-validation minimum five folds should be standard practice for all studies using datasets below 5,000 images. Statistical significance testing of performance comparisons should be reported alongside accuracy figures: McNemar's test for paired classifier comparisons on the same test set, paired t-tests across cross-validation folds, or bootstrap confidence intervals where appropriate. Independent validation on datasets collected in different contexts from training data should be used to assess real-world generalizability; held-out subsets of the same collection are insufficient for this purpose.

Standardized metric reporting accuracy, precision, recall, F1-score, and mAP at multiple IoU thresholds for detection tasks should be adopted uniformly to enable cross-study comparison. Studies reporting only top-1 accuracy on balanced test sets should be treated as providing incomplete evidence.

A minimum hyperparameter transparency checklist should accompany all published models, covering:

(i) Optimizer type and learning rate schedule
(ii) Batch size and training epoch count
(iii) Loss function specification
(iv) Regularisation parameters including dropout rate, weight decay, and L1/L2 coefficients
(v) Data split ratios or cross-validation fold count
(vi) Random seed settings for reproducibility
(vii) Hardware specifications and total training duration. Computational complexity reporting should consistently include model parameters, FLOPs, inference latency on a specified reference platform, model file size, and peak memory utilisation where feasible

Deployment and Integration Recommendations

The gap between laboratory-validated models and field-deployable systems is the most consequential unresolved problem in this literature. Closing it requires changes to both technical design priorities and evaluation practice.

On-device inference using TensorFlow Lite, ONNX Runtime, or PyTorch Mobile should be prioritized over cloud-dependent architectures. Connectivity-limited rural deployments which characterize the majority of cocoa-growing regions in West Africa and Southeast Asia cannot rely on server-side inference, and no reviewed study demonstrates that its system functions without network access. Offline-first design principles should govern mobile application development from the outset rather than being retrofitted after model training.

Formal usability testing with representative farmer populations should be incorporated into system evaluation as a mandatory component, not an optional extension. Technical performance metrics alone cannot establish whether a system will be adopted. Interface design for low-literacy users, language localisation, workflow integration with existing farming practice, and the cognitive load of interpreting model outputs are all determinants of adoption that remain entirely unaddressed in the reviewed literature.

IoT integration connecting mobile detection systems with weather monitoring infrastructure, crop management databases, and agronomic advisory services represents a high-value research direction that would transform point-in-time disease classification into continuous field monitoring. Model compression techniques including pruning, quantisation, and architecture search should be systematically evaluated to identify the accuracy–efficiency frontier for specific target devices, rather than assuming that lightweight architectures are inherently suitable without empirical verification on representative hardware.

Cross-Domain Validation and Open-Source Dataset Strategy

The Evaluation on held-out subsets drawn from the same dataset provides limited evidence of real-world generalization, measuring performance within a single data distribution rather than across regions, cultivars, seasons, and imaging conditions. Cross-dataset and cross-domain validation have consequently become expected standards in plant disease detection research (Mohanty et al., 2016; Ramcharan et al., 2017; Ferentinos, 2018). Future studies should adopt one or more of the following validation protocols, listed in ascending order of rigor:

1. Geographic split validation trains models on data from one plantation or region and evaluates them on geographically distinct locations. This should be considered the minimum standard for supporting deployment-oriented generalisation claims
2. Leave-One-Dataset-Out (LODO) validation trains on all available datasets except one and evaluates performance on the excluded dataset. As demonstrated by Techie-Menson et al. (2026), this represents the most rigorous protocol identified in the reviewed literature because it directly measures cross-dataset generalisation
3. Cross-varietal validation trains on one cocoa cultivar and tests on another, providing evidence of robustness to genetic variation in disease symptom expression across production regions
4. Temporal validation trains on data from one growing season and evaluates on subsequent seasons, assessing resilience to environmental, phenological, and disease-related variability over time

In addition to adopting more rigorous validation frameworks, researchers should complement mean accuracy with measures of variability, such as standard deviations and confidence intervals, and apply statistical significance tests to verify whether observed performance differences are meaningful rather than attributable to random variation (Demšar, 2006). Recommended methods including McNemar's test for paired classifier comparisons, Wilcoxon signed-rank tests across cross-validation folds, and bootstrap confidence intervals (Demšar, 2006; Japkowicz and Shah, 2011).

The absence of publicly accessible benchmark datasets is the primary barrier to reproducible and comparable cocoa disease detection research. Without shared datasets, independent replication, cross-study benchmarking, and cumulative scientific progress remain limited. Journals, conferences, and funding agencies should therefore require or strongly encourage public dataset deposition as a condition of publication. The following recommendations are proposed for researchers developing new cocoa disease dataset:

1. Repository deposition should ensure datasets are publicly available in persistent, citable repositories such as Zenodo, Mendeley Data, IEEE DataPort, or the CGIAR Data Portal before or during manuscript submission, enabling independent review and long-term reproducibility
2. Minimum documentation standards must specify collection location, date range, imaging specifications, cultivar information, disease classes and severity criteria, annotation protocols, class distributions, representative images, inter-annotator agreement statistics, and a clear data-use license
3. Annotation quality reporting should quantify annotation reliability using Cohen's or Fleiss' Kappa, with values ≥ 0.70 indicating substantial agreement. Annotations should be validated by at least one qualified plant pathologist and the validation process explicitly documented
4. Benchmark split standardization should provide predefined train, validation, and test partitions or standardized cross-validation folds to support reproducible and comparable evaluation, following the benchmark practices established by ImageNet, COCO, and PlantVillage
5. Coordinated international collaboration should be pursued to develop benchmark datasets that capture the geographic, varietal, and environmental diversity of global cocoa production, leveraging partnerships among cocoa research institutions, development agencies, and industry stakeholders

Strengths and Limitations of this Review

Strengths

This review offers several methodological and substantive strengths that distinguish it from prior work in the domain.

The PRISMA-guided methodology ensures transparency, replicability, and systematic rigour in study identification, screening, and selection, providing a reproducible audit trail for the inclusion decisions made throughout the review process. The multi-database search strategy spanning Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, and PubMed maximises coverage across both technical and agricultural research communities, reducing the risk of systematic omission from any single disciplinary index.

The analytical scope extends beyond descriptive enumeration. The review integrates dataset-versus-performance mapping, computational complexity synthesis, multi-dimensional radar comparison, and Spearman correlation analysis of methodological choices against reported outcomes generating comparative insights that prior domain reviews have not attempted. The explicit identification of cross-cutting methodological deficiencies in Section 3.4.2, rather than treating gaps as study-specific limitations, provides a more honest characterization of the field's structural weaknesses. The forward-looking synthesis in Section 3.5, with structured recommendations across architecture, dataset development, evaluation standards, and deployment practice, translates analytical findings into actionable guidance rather than leaving them as observations.

Limitations

Several limitations constrain the scope and inferential strength of this review and should be considered when interpreting its findings.

Methodological heterogeneity across primary studies in evaluation metrics, dataset characteristics, model architectures, and reporting completeness limits the feasibility of quantitative meta-analysis. Findings are synthesized narratively and comparatively rather than statistically pooled, which reduces precision but reflects the actual state of the evidence base rather than imposing false comparability.

Publication bias is a structural concern. Studies reporting high classification performance are more likely to be published and indexed in the searched databases, meaning the reviewed evidence base may over represent favourable results and underrepresent negative findings, failed deployments, and models that performed poorly on real-world data. This bias is particularly consequential given the anomalously high accuracy figures identified in the reviewed papers

The search was confined to English-language publications. Relevant work published in French, Spanish, Portuguese, or Indonesian languages prevalent in major cocoa-producing regions including Côte d'Ivoire, Colombia, Brazil, and Indonesia may have been excluded, potentially introducing a geographic bias toward Anglophone research institutions.

The pace of development in deep learning architectures means that some methodological assessments may be superseded by advances published after the review search cutoff. Transformer-based and multimodal architectures in particular are evolving rapidly, and the gap analysis presented here should be read as a snapshot of the field at the time of review rather than a permanent characterization.

Finally, where studies omitted parameter counts or computational metrics, values reported in Table 6 were inferred from canonical architecture specifications. These inferences may not exactly match the modified implementations used by individual authors, and should be treated as indicative rather than precise.

Implications for Research, Practice, and Policy

Research Implications

This review highlights the need to prioritize hybrid and attention-based architectures, diversified large-scale datasets, standardized evaluation protocols, and rigorous cross-validation practices in future cocoa disease detection research. It also identifies transfer learning, colour-space augmentation, and overfitting control as key methodological factors associated with more reliable performance outcomes.

The findings further emphasize the importance of interdisciplinary collaboration among computer scientists, plant pathologists, agronomists, and human–computer interaction specialists. Addressing the persistent gap between laboratory performance and field deployment will require stronger integration of agronomic and biological expertise into AI system design and evaluation.

Practical Implications

For agricultural practitioners and extension services, current evidence suggests that AI-based cocoa disease detection is increasingly feasible for real-world deployment. Lightweight models such as MobileNetV2, EfficientNet-Lite, and SSD MobileNetV2 FPN-Lite provide practical pathways for smartphone-based diagnosis, while the LGF-CBAM framework proposed by Techie-Menson et al. (2026) demonstrates strong robustness on field-collected data.

However, deployment should be approached cautiously, as many existing systems rely on geographically limited or insufficiently validated datasets. Cross-regional validation and usability testing with farmers remain essential before large-scale adoption can be considered reliable or practical.

Policy Implications

National cocoa regulators and development agencies should prioritize standardized, open-access cocoa disease datasets and connectivity-independent diagnostic tools to support smallholder farmers in regions with limited internet access. Strengthening agricultural AI research capacity in cocoa-producing countries is also essential for promoting locally driven innovation.

In addition, regulatory frameworks for AI-based agricultural diagnostics should establish clear standards for validation, transparency, and deployment to ensure reliable and trustworthy use in farming environments.

Future Research Directions

Several priority research directions emerge from the identified gaps in cocoa pod disease detection. Foremost is the development of standardized benchmark datasets that are large-scale, publicly accessible, and geographically diverse, with clear documentation of class distribution, imaging conditions, annotation protocols, disease severity levels, and inter-annotator reliability. Such datasets are essential for improving reproducibility, enabling reliable cross-study comparisons, and validating claims of model generalization.

Hybrid and transformer-enhanced architectures also represent promising avenues for future research. CNN–Vision Transformer hybrids, attention-guided CNNs, and multi-scale feature fusion frameworks have demonstrated strong potential for improving contextual representation and disease discrimination under complex field conditions. Although preliminary studies by Kouassi et al. and Techie-Menson et al. reported encouraging results, broader evaluations across diverse datasets, cocoa varieties, and disease classes remain necessary. Future

work should include rigorous ablation studies to isolate the contributions of attention mechanisms, transformer modules, feature fusion strategies, and classifier designs.

Integrating environmental and contextual metadata is another underexplored but valuable direction. Variables such as temperature, humidity, rainfall, canopy density, soil conditions, and seasonal variation strongly influence disease development and symptom expression. Combining image-based analysis with environmental sensing data may improve early disease detection and enable more context-aware diagnostic systems.

Explainable Artificial Intelligence (XAI) should also become a standard component of future diagnostic frameworks. Techniques such as Gradient-weighted Class Activation Mapping, SHapley Additive exPlanations, and Local Interpretable Model-agnostic Explanations can improve transparency and user trust by visualizing model decision processes. However, interpretability outputs should also be validated with plant pathologists to ensure biological relevance and diagnostic reliability.

Real-world validation remains essential for bridging the gap between laboratory accuracy and field applicability. Most existing studies rely on controlled datasets that fail to capture practical agricultural conditions such as variable illumination, occlusion, motion blur, inconsistent imaging angles, and complex backgrounds. Future studies should therefore prioritize large-scale field evaluations across multiple climatic regions, cocoa varieties, and operational environments to establish robustness and scalability.

IoT-enabled monitoring systems represent another promising direction for proactive disease management. Integrating hyperspectral or multispectral imaging with edge-based inference and distributed sensor networks could support autonomous plantation monitoring, continuous environmental sensing, disease forecasting, and precision intervention planning.

Federated learning also offers strong potential for addressing privacy, connectivity, and data-sharing challenges in distributed cocoa farming systems. By enabling collaborative model training without centralized data aggregation, federated learning can improve model generalization while preserving farmer data ownership and supporting deployment in resource-constrained environments.

Longitudinal disease progression modeling remains largely unexplored despite its importance for predictive agriculture. Current approaches focus primarily on static image classification rather than tracking temporal disease evolution. Incorporating sequential and temporal learning methods could enable prediction of disease progression, epidemic spread, and optimal intervention timing, thereby shifting AI applications from reactive diagnosis toward proactive crop management.

Future studies should also incorporate comprehensive computational profiling. Model evaluation should extend beyond accuracy to include metrics such as inference latency, floating-point operations (FLOPs), memory consumption, throughput, and energy efficiency on low-end smartphones and edge devices commonly used in farming environments. Such profiling is critical for assessing real-world deployment feasibility.

Finally, the establishment of standardized evaluation protocols is necessary to improve methodological consistency across the field. Future research should define unified guidelines for dataset partitioning, validation procedures, evaluation metrics, statistical significance testing, and object detection benchmarks across multiple IoU thresholds. Standardized protocols would significantly enhance reproducibility, comparability, and the scientific maturity of cocoa disease detection research.

Conclusion

This systematic review critically analyzed 25 peer-reviewed studies on automated cocoa pod disease detection published between 2016 and 2026, examining methodological approaches, implementation strategies, performance outcomes, and research gaps. The review shows that while classification accuracy has improved considerably through transfer learning and attention-based architectures such as LGF-CBAM, progress in methodological rigor, deployment readiness, and dataset quality remains limited.

CNN-based models including EfficientNet, MobileNet, ResNet, and VGG dominate the field, consistently outperforming training-from-scratch approaches, while hybrid CNN and CNN–ViT architectures demonstrate emerging potential. However, most studies rely on small, poorly documented datasets, limiting generalizability and cross-study comparison. The absence of standardized benchmark datasets, limited use of advanced preprocessing techniques, inadequate overfitting controls, and weak validation practices further reduce confidence in reported results.

The review also reveals major gaps in computational profiling, mobile deployment, and real-world field validation. Many studies fail to report essential deployment metrics such as FLOPs, inference latency, memory usage, and energy consumption, while challenges such as visually similar disease classes and environmental variability remain unresolved. Importantly, no existing model achieves balanced performance across accuracy, robustness, computational efficiency, and deployment suitability simultaneously.

Overall, the findings suggest that the field has prioritized high accuracy on controlled datasets over practical deployment and scientific rigor. Advancing cocoa disease detection will require standardized benchmark datasets, stronger interdisciplinary collaboration, rigorous validation

practices, and deployment-oriented evaluation standards. Although AI-based cocoa disease management holds significant promise, its real-world impact depends on improving methodological quality and translational readiness across future research.

Acknowledgment

The authors would like to thank colleagues and reviewers for their constructive comments and valuable suggestions, which have significantly improved the quality of this manuscript. We also acknowledge the support provided by our respective institutions in facilitating this research through access to relevant resources and an enabling research environment

Funding Information

No funding was received for conducting this study.

Author's Contributions

Henry Techie-Menson: Conceptualization, methodology, formal analysis, writing original draft.

Michael Asante: Supervision, Validation, writing review and editing.

Yaw Marfo Missah and Gaddafi Abdul-Salaam: Supervision, validation, writing review and edited.

Stephen Opoku Oppong: Formal analysis, investigation, data curation, visualization.

Ethics

This systematic review followed established ethical guidelines for literature search and data extraction. Ethical approval was not required as all analyses were based on previously published, publicly available data. The authors declare no conflicts of interest

References

Ameyaw, G. A., Dzahini-Obiatey, H. K., & Domfeh, O. (2023). Current status of cocoa swollen shoot virus disease management in Ghana. *Journal of Plant Diseases and Protection*, *130*(3), 345–356.

Amoako, P. Y., Cao, G., & Arthur, J. K. (2023). An Image-Based Cocoa Diseases Classification Based on an Improved Vgg19 Model. *Sustainable Education and Development – Sustainable Industrialization and Innovation*, 711–722. https://doi.org/10.1007/978-3-031-25998-2_55

Ariandi, V., Kurnia, H., Heriyanto, & Marry, H. (2019). Expert system for disease diagnosis in cocoa plant using android-based forward chaining method. *Journal of Physics: Conference Series*, 012009. https://doi.org/10.1088/1742-6596/1339/1/012009

Asamoah, G. K. (2022). Agricultural transformation and rural development in Ghana: Challenges and opportunities. *Journal of African Agriculture*, *45*(3), 178–195.

Asare-Bediako, E., Opoku-Asiama, Y., & Addo-Quaye, A. A. (2023). Management strategies for cocoa swollen shoot virus disease: Progress and prospects in Ghana. *Plant Disease Management*, *107*(3), 495–507.

Atianashie, M. A. (2024). Artificial Intelligence (AI) Disease Detection in CCN-51 Cocoa Fruits through Convolutional Neural Networks: A Novel Approach for the Ghana Cocoa Board. *Convergence Chronicles*, *5*(3), 51–72. https://doi.org/10.53075/Ijmsirq/6554213434324

Basri, H., Indrabayu, Areni, I. S., & Tamin, R. (2019). Image Processing System for Early Detection of Cocoa Fruit Pest Attack. *Journal of Physics: Conference Series*, *1244*(1), 012003. https://doi.org/10.1088/1742-6596/1244/1/012003

Basri, I., Achmad, A., & Areni, I. S. (2022). Comparison of Image Extraction Model for Cocoa Disease Fruits Attack in Support Vector Machine Classification. *2022 International Conference on Electrical and Information Technology (IEIT)*, 46–51. https://doi.org/10.1109/ieit56384.2022.9967910

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepahvand, N. M., Raff, E., Madan, K., Voleti, V. V., Kahou, S. E., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 747–769.

de Oliveira, J. R. C. P., & Romero, R. Ap. F. (2018). Transfer Learning Based Model for Classification of Cocoa Pods. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–6. https://doi.org/10.1109/ijcnn.2018.8489126

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30. https://doi.org/10.5555/1248547.1248548

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

Ferraris, S., Meo, R., Pinardi, S., Salis, M., & Sartor, G. (2023). Machine Learning as a Strategic Tool for Helping Cocoa Farmers in Côte D'Ivoire. *Sensors*, *23*(17), 7632. https://doi.org/10.3390/s23177632

Gamboa, A. A., Caceres, P. A., Lamos, H., Zarate, D. A., & Puentes, D. E. (2019). Predictive model for cocoa yield in Santander using Supervised Machine Learning. *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, 1–5. https://doi.org/10.1109/stsiva.2019.8730258

Garma, G. M. (2025). Efficient detection of cacao pod diseases using SSD MobileNetV2 FPN-Lite. *World Journal of Advanced Research and Reviews*, *25*(2), 1099–1105. https://doi.org/10.30574/wjarr.2025.25.2.0128

Godmalin, R. A., Aliac, C. J., & Feliscuzo, L. (2022). Classification of Cacao Pod if Healthy or Attack by Pest or Black Pod Disease Using Deep Learning Algorithm. *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 1–5. https://doi.org/10.1109/iicaiet55139.2022.9936817

Godmalin, R. A., Aliac, C. J., & Feliscuzo, L. (2023). Cacao Pod Infection Level Classification Using Transfer Learning. *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (EStream)*, 1–6. https://doi.org/10.1109/estream59056.2023.10135062

Gupta, R., & Sharma, V. (2023). Leveraging drones and mobile AI for sustainable cocoa farming. *Remote Sensing Applications in Agriculture*, *8*(2), 112–127.

Japkowicz, N., & Shah, Mohak. (2011). *Evaluating learning algorithms: A classification perspective*. https://doi.org/10.1017/CBO9780511921803

Kouassi, K. S., Diarra, M., Edi, K. H., & Jean-Claude, K. B. (2025). Detection of cocoa pod diseases using a hybrid feature extractor combining CNN and vision transformer with dual classifier. *Edelweiss Applied Science and Technology*, *9*(1), 668–681. https://doi.org/10.55214/25768484.v9i1.4209

Kouassi, K. S., Diarra, M., Edi, K. H., & Koua, B. J.-C. (2024). Detection of Cocoa Leaf Diseases Using the CNN-Based Feature Extractor and XGBOOST Classifier. *Open Journal of Applied Sciences*, *14*(10), 2955–2972. https://doi.org/10.4236/ojapps.2024.1410193

Kumi, S., Kelly, D., Woodstuff, J., Lomotey, R. K., Orji, R., & Deters, R. (2022). Cocoa Companion: Deep Learning-Based Smartphone Application for Cocoa Disease Detection. *Procedia Computer Science*, *203*, 87–94. https://doi.org/10.1016/j.procs.2022.07.013

Lomotey, R. K. Orji, R., Deters, & R., Kumi, S., (2024). Automatic detection and diagnosis of cocoa diseases using mobile tech and deep learning. *International Journal of Sustainable Agricultural Management and Informatics*, *10*(1), 92–119. https://doi.org/10.1504/ijsami.2024.10059217

Mamadou, D., Ayikpa, K. J., Ballo, A. B., & Kouassi, B. M. (2023). Cocoa Pods Diseases Detection by MobileNet Confluence and Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *14*(9), 37. https://doi.org/10.14569/ijacsa.2023.0140937

Maylianti, N. P., Wijayakusuma, G. N. L., & Wiguna, P. C. A. (2025). Comparison of EfficientNet-B0 and ResNet-50 for Detecting Diseases in Cocoa Fruit. *Journal of Applied Informatics and Computing*, *9*(1), 115–120. https://doi.org/10.30871/jaic.v9i1.8868

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542. https://doi.org/10.1109/tpami.2021.3059968

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

Montesino, R. Y., Rosales-Huamani, J. A., & Castillo-Sequera, J. L. (2021). Detection of phytophthora palmivora in cocoa fruit with deep learning. *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–4. https://doi.org/10.23919/cisti52073.2021.9476279

Osei-Kwame, P., & Mensah, B. (2024). Evolution of cassava diseases in Sub-Saharan Africa: A review of emerging threats. *African Journal of Agricultural Research*, *12*(1), 45–62.

Pusadan, M. Y., Syahrullah, Merry, & Abdullah, A. I. (2022). k-Nearest Neighbor and Feature Extraction on Detection of Pest and Diseases of Cocoa. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(3), 471–480. https://doi.org/10.29207/resti.v6i3.4064

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep Learning for Image-Based Cassava Disease Detection. *Frontiers in Plant Science*, *8*, 1852. https://doi.org/10.3389/fpls.2017.01852

Rola, J. B., Barrera, J. J. A., Calhoun, M. V., Maaghop, J. F. O. –, Unajan, M. C., Boncalon, J. M., Sebios, E. T., & Espinosa, J. S. (2024). Convolutional Neural Network Model for Cacao Phytophthora Palmivora Disease Recognition. *International Journal of Advanced Computer Science and Applications*, *15*(8), 97. https://doi.org/10.14569/ijacsa.2024.0150897

Serrano, J. S., Camilo, A., & Villamizar, A. T. (2020). *Mobile application prototype for the identification of sick cocoa cobs using vision computer and machine learning*.

Singh, A., & Kumar, P. (2023). Traditional plant disease monitoring methods: Resource intensiveness and scalability issues in developing countries. *Journal of Agricultural Science*, *11*(3), 78–94.

Soh, K. S., Moung, E. G., Danker, K. J. J., Dargham, J. A., & Farzamnia, A. (2024). Cocoa Diseases Classification using Deep Learning Algorithm. *ITM Web of Conferences*, *63*, 01014. https://doi.org/10.1051/itmconf/20246301014

Subramanian, C. V., Shalinikumari, R., Hussain, M. M., Mary, P. S. A., Deiwakumari, K., & Nareshkumar, R. (2025). Cocoa Pod disease detection based on convolutional neural network for healthy landscape. *AIP Conference Proceedings*, 3162.. https://doi.org/10.1063/5.0242381

Tan, D. S., Leong, R. N., Laguna, A. F., Ngo, C. A., Lao, A., Amalin, D., & Alvindia, D. (2016). A framework for measuring infection level on cacao pods. *2016 IEEE Region 10 Symposium (TENSYMP)*, 384–389. https://doi.org/10.1109/tenconspring.2016.7519437

Techie-Menson, H., Asante, M., Missah, Y. M., Abdul-Salaam, G., & Oppong, S. O. (2026). Enhanced convolutional block attention module with Learnable Gated Fusion (LGF-CBAM) for cocoa pod disease identification. *PLOS One*, *21*(4), e0348147. https://doi.org/10.1371/journal.pone.0348147

Vera, D. B., Oviedo, B., Casanova, W. C., & Zambrano-Vega, C. (2024). Deep learning-based computational model for disease identification in cocoa plants. *ArXiv Preprint ArXiv:2401.01247*, *7*, 1–30. https://doi.org/10.48550/arXiv.2401.01247

Wilson, J., & Chen, L. (2024). Emerging maize diseases in North America: Focus on tar spot expansion. *Plant Disease Review*, *15*(1), 78–92.

Zambrano-Vega, C., Vera, D. B., Casanova, W. C., & Oviedo, B. (2024). Deep Learning-Based Computational Model for Disease Identification in Cocoa Pods. *2024 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 290–296. https://doi.org/10.1109/pic62406.2024.10892748

Zhang, T., Chen, L., & Zhou, H. (2023). Edge AI for agricultural disease detection: MobileNet vs. EfficientNet for cocoa pod classification. *Journal of Computer Vision in Agriculture*, *7*(1), 89–102.