ISSN: 2581-1274

Benchmarking Deep Learning Models for Multi-Class Fruit Quality and Formalin Contamination Detection in Horticultural Fruits

Ahmad Fauzi [a,1,*], Achmad Lutfi Fuadi [a,2], Agus Heri Yunial [a,3]

[a] *Department of Informatics Engineering, Universitas Pamulang, Jl. Raya Puspiptek, South Tangerang 15310, Indonesia*
[1] *dosen02621@unpam.ac.id *;* [2] *dosen02524@unpam.ac.id;* [3] *dosen02525@unpam.ac.id*
**Corresponding author*

ARTICLE INFO

Article history:
Published
September 14, 2026

Keywords:
Deep learning
Fruit quality classification
Formalin detection
MobileNetV2
Model efficiency

ABSTRACT

The adulteration of fresh horticultural produce with formaldehyde (formalin) and the rapid physiological decay of fruit remain persistent food-safety and economic problems within agricultural supply chains. Deep learning offers a non-destructive route to automated fruit-quality screening; however, the literature has largely pursued peak accuracy using progressively heavier network backbones, leaving the computational cost of that accuracy, and therefore the feasibility of on-site deployment, insufficiently examined. This study presents a controlled, identical-protocol comparison of three visual-recognition paradigms of markedly different scale, namely MobileNetV2 (a depthwise-separable lightweight convolutional neural network), ResNet-50 (a deep residual network), and ViT-B/16 (a patch-based Vision Transformer), on the three-class task of distinguishing fresh, rotten, and formalin-mixed fruit. A total of 10,154 images from the public FruitVision dataset were processed under a single fixed preprocessing and training protocol, so that architecture was the only experimental variable, and performance was assessed through an efficiency lens that normalises accuracy by parameter count and by inference cost (GFLOPs). The lightweight MobileNetV2 attained the highest test accuracy (99.69%) while using about one-seventh of the parameters and one-thirteenth of the floating-point operations of ResNet-50 and roughly one twenty-fifth of the parameters of ViT-B/16, which recorded the lowest accuracy (99.28%) at the highest cost. Because the models were trained once, the small accuracy gap is not interpreted as a definitive advantage; the robust distinction is in computational cost. The compact model is positioned as a preliminary, non-destructive formalin-screening aid on resource-constrained devices, intended to flag suspicious samples for laboratory confirmation rather than to replace chemical testing. Per-class reliability, including one residual false negative, is reported without overstatement

I. Introduction

Horticultural produce occupies a central position in global food security and in the agrarian economy, both for its nutritional value and for the livelihoods it supports. Its commercial value, however, is undermined by two recurring hazards: the perishable nature of fresh fruit, which drives rapid post-harvest quality loss, and the illicit use of formaldehyde (commonly termed formalin) as an unauthorised preservative to prolong apparent freshness. Formaldehyde is a recognised toxicant, and its presence in food additives and preservatives has become a focus of toxicological and regulatory concern [1]. While legitimate preservation technologies such as edible barrier coatings are advancing [2], the persistence of chemical adulteration means that rapid, accessible screening at the point of sale remains a public-health priority.

Conventional confirmatory methods for chemical adulteration and quality grading, including wet-chemical assays and visible/near-infrared (Vis/NIR) spectroscopy, are accurate but generally

expensive, instrument-bound, and impractical for real-time monitoring within decentralised supply chains [3]. Computer vision powered by deep learning offers a scalable, non-destructive alternative that learns discriminative features directly from ordinary RGB imagery. Prior work has demonstrated strong performance for binary or coarse freshness grading of fruit [4], and the recent release of the public FruitVision benchmark, which explicitly incorporates a formalin-mixed class alongside fresh and rotten conditions, has enabled the more demanding multi-class formulation that couples quality assessment with chemical-contamination screening [5], [6].

A dominant tendency in this literature is the pursuit of ever-higher accuracy through progressively larger and deeper backbones. Lightweight convolutional networks such as MobileNetV2 have been applied successfully in constrained settings [7]; deep residual networks such as ResNet-50 remain a strong default for complex visual tasks [19]; and, more recently, Vision Transformers have been introduced into agricultural recognition, including smartphone-oriented deployments [9] and self-attention fusion designs [10]. Comparative evaluations of competing classifiers under a common protocol, such as ConvMixer versus ResNet for medical image classification [18] and compact custom CNNs trained under minimal budgets [11], have proved valuable for isolating the practical trade-offs between paradigms. Yet the accuracy figures reported on quality-grading tasks are frequently near the ceiling, so that differences between models become marginal, while their computational and memory footprints differ by more than an order of magnitude. For field inspection on mobile or embedded hardware this footprint is decisive, yet the cost of the accuracy that heavier backbones deliver on the combined quality-and-formalin task has not been systematically quantified.

Three specific gaps follow from this observation. First, the accuracy return of additional architectural capacity for the combined fruit-quality-and-formalin classification task is unquantified: studies typically report the accuracy of a single backbone without relating it to the parameters and floating-point operations consumed, so it is unknown whether a heavier model is warranted. Second, no controlled, identical-protocol comparison places a depthwise-separable lightweight CNN, a deep residual CNN, and a patch-based Vision Transformer on the same three-class task; in particular, the suitability of Transformers, whose weak inductive bias is data-hungry, for the medium-scale (about 10^4 image) datasets typical of agricultural research is unverified. Third, efficiency in this domain is seldom expressed as accuracy normalised by computational cost, so architecture selection for embedded inspection remains ad hoc rather than evidence-based.

This study addresses these gaps through a controlled benchmark in which architecture is the sole experimental variable. Its contributions are threefold. (i) It provides a controlled, identical-protocol comparison of three visual-recognition paradigms, MobileNetV2, ResNet-50, and ViT-B/16, for combined fruit-quality-and-formalin classification, evaluated not by accuracy alone but through an explicit efficiency lens that reports accuracy per million parameters and accuracy per GFLOP. (ii) It shows empirically that, among the three architectures evaluated, the lightweight CNN is Pareto-optimal under a clearly defined accuracy-cost criterion (Section 2.5), matching or exceeding the heavier models at a fraction of the computational cost, whereas the Vision Transformer is Pareto-dominated on this medium-scale dataset. (iii) It reports an honest per-class reliability analysis, including a residual false negative in the safety-relevant formalin class, together with a computational assessment of embedded feasibility. The contribution is therefore empirical and methodological, a controlled evaluation under a common protocol rather than the introduction of a new model or algorithm; the study clarifies which existing paradigm is justified for practical, resource-aware deployment.

II. Method

The research followed a controlled, quantitative experimental design whose purpose was to isolate the effect of network architecture on multi-class fruit-quality and formalin classification. As summarised in Figure 1, a single dataset was passed through one fixed preprocessing pipeline, three architectures were trained under an identical protocol, and the resulting models were evaluated with both predictive and computational-efficiency metrics. Because every stage other than the architecture was held constant, observed differences can be attributed to the architectural paradigm rather than to preprocessing or optimisation choices.

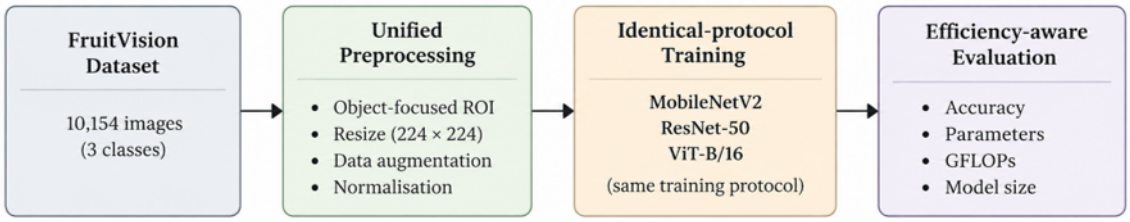

Fig. 1 Controlled benchmarking workflow in which architecture is the only experimental variable.

A. Dataset and Partitioning

The publicly available FruitVision dataset was used, comprising 10,154 RGB images of five fruit species (apple, banana, grape, mango, and orange) [5]. For this study the species dimension was not modelled; instead the images were organised into the three condition classes that define the task, fresh, rotten, and formalin-mixed. As documented by the dataset providers [5], the fresh and rotten conditions were photographed directly, whereas the formalin-mixed class was created by immersing fruit in a formalin solution and imaging it after a stabilisation period. Images were captured with several consumer cameras of differing resolution (approximately 12-64 MP) at ambient temperatures of about 24-34 degrees Celsius, and the five species were photographed across separate sessions, introducing device, illumination, and specimen variability that is relevant to external validity. The resulting class distribution is reported in Table 1. To prevent optimistic bias from data leakage, a group-aware stratified split was applied so that all images originating from the same physical fruit were confined to a single subset, using a 70% training, 20% validation, and 10% test partition; the identical split index was reused for every architecture.

Table 1. Class distribution of the FruitVision images used in this study

Condition class	Training	Validation	Test	Total
Fresh	2,660	760	380	3,800
Rotten	2,225	635	318	3,178
Formalin-mixed	2,223	635	318	3,176
Total	7,108	2,030	1,016	10,154

B. Unified Preprocessing

A single preprocessing pipeline was applied uniformly to all architectures and therefore constitutes a controlled variable rather than an object of study. Each image was first passed through a one-time, object-focused region-of-interest step that isolates the fruit from background clutter using instance segmentation, an approach shown to improve real-time recognition in agricultural imagery [12], [13]. The stage employed a pretrained YOLOv8-seg instance-segmentation model (Ultralytics); for each image the segmented candidate with the largest area whose confidence exceeded 0.30 was selected as the primary fruit object, expanded with 10% padding, and letterbox-resized to 224 x 224 pixels, then normalised using ImageNet statistics. Images for which no candidate exceeded the confidence threshold were treated as extraction failures and excluded. Data augmentation, namely horizontal flipping, rotation within plus or minus 30 degrees, and mild photometric adjustment, was applied to the training subset only, while the validation and test subsets received resizing and normalisation alone, so that reported metrics reflect generalisation to unseen data.

This exclusion criterion removed 40 of the 1,016 test images (3.9%). The removals were not evenly distributed across classes: 15 fresh (3.9%), 18 rotten (5.7%), and 7 formalin-mixed (2.2%), leaving an effective test set of 976 images (fresh 365, rotten 300, formalin-mixed 311). Because the segmentation step is agnostic to the classifier and to the ground-truth label, and because the safety-relevant formalin-mixed class had the lowest exclusion rate, this curation is unlikely to bias the evaluation in favour of formalin detection; the higher exclusion in the rotten class is consistent with its more pronounced morphological deformation. As the pipeline is fixed and identical for all architectures, the study does not compare preprocessing strategies; its findings concern architecture under a realistic, held-constant input representation.

C. Candidate Architectures and Complexity

Three architectures spanning a wide complexity range were selected to represent distinct design philosophies. MobileNetV2 is an efficiency-oriented convolutional network whose inverted-residual, depthwise-separable convolutions minimise parameters and computation. ResNet-50 is a deep residual network that provides high representational capacity through skip connections. ViT-B/16 is a patch-based Vision Transformer that processes an image as a sequence of 16 x 16 patches through multi-head self-attention and therefore lacks the spatial locality bias intrinsic to convolution. All three models were implemented in PyTorch using torchvision's ImageNet-pretrained weights; the final fully connected layer of each was replaced by a three-unit classification head, and all layers were fine-tuned end-to-end (no layers were frozen). Inputs were 224 x 224 x 3 RGB tensors. Their parameter counts, inference costs, and on-disk sizes are contrasted in Table 2 and span roughly two orders of magnitude in cost.

Table 2. Design characteristics and computational complexity of the three architectures

Architecture	Design paradigm	Parameters (M)	GFLOPs	Model size (MB)
MobileNetV2	Lightweight (depthwise-separable)	3.50	0.65	14
ResNet-50	Deep residual CNN	25.56	8.27	98
ViT-B/16	Vision Transformer (self-attention)	86.38	33.70	330

D. Training Protocol

To guarantee comparative validity, all models were trained under one identical configuration: the Adam optimiser (beta1 = 0.9, beta2 = 0.999) with an initial learning rate of 1 x 10^{-4} (0.0001), a batch size of 32, a maximum of 50 epochs, categorical cross-entropy loss, and a ReduceLROnPlateau scheduler (factor 0.1, patience 3). All experiments were run in a Kaggle Notebook environment on a single NVIDIA Tesla T4 GPU (16 GB) using PyTorch. For each architecture the checkpoint with the lowest validation loss was retained and evaluated once on the held-out test set. Each model was trained a single time under this configuration; consequently, the small accuracy differences between models are interpreted with corresponding caution (Section 3.1). No architecture-specific tuning, class weighting, or additional regularisation was introduced, so that the comparison reflects the paradigms under equal conditions rather than the outcome of unequal optimisation effort. The complete training configuration and code can be made available by the authors on request.

E. Evaluation Metrics

Predictive performance was quantified on the disjoint test set using accuracy, and macro-averaged precision, recall, and F1-score, defined in the usual way from the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Because predictive accuracy alone cannot express deployability, two efficiency ratios were additionally computed to relate accuracy to resource consumption. For a model with test accuracy A (in per cent), P million parameters, and G GFLOPs, the parameter efficiency and the computational efficiency were defined as:

$$E_{\text{param}} = \frac{A}{P} \quad (5)$$

$$E_{\text{flop}} = \frac{A}{G} \quad (6)$$

A model was regarded as Pareto-optimal, among the three architectures evaluated, when no competing model achieved higher accuracy at lower or equal computational cost. Computational cost is quantified by parameter count, GFLOPs, and model size, which serve as hardware-independent proxies for deployability; actual on-device latency was not measured in this study and is identified as future work. This efficiency-aware framing, rather than the maximisation of accuracy in isolation, forms the basis of the analysis reported below.

III. Results and Discussion

A. Overall Classification Performance

Table 3 reports the performance of the three architectures on the independent test set. All models classified the three conditions to a high standard, with test accuracies confined to the narrow band between 99.28% and 99.69%. The lightweight MobileNetV2 recorded the highest accuracy (99.69%), marginally ahead of ResNet-50 (99.59%), while the Vision Transformer ViT-B/16 was lowest (99.28%). The macro-averaged precision, recall, and F1-score tracked accuracy closely for every model, indicating balanced behaviour across the three classes rather than a metric inflated by class imbalance. The central observation is therefore not that any one architecture is dramatically more accurate, but that accuracy is near-saturated for all three: a difference of 0.41 percentage points separates the best and worst models, equivalent to only a handful of test images. Because each model was trained once, this 0.41-percentage-point gap lies within the run-to-run variability expected of single-run training and is not interpreted as evidence of a meaningful accuracy advantage; the robust, reproducible distinction between the models lies instead in their computational cost, which differs deterministically by more than an order of magnitude. Accuracy alone thus provides weak grounds for model selection, which motivates the efficiency analysis that follows.

Table 3. Test-set classification performance of the three architectures

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MobileNetV2	99.69	99.70	99.71	99.69
ResNet-50	99.59	99.59	99.59	99.59
ViT-B/16	99.28	99.28	99.28	99.28

B. Accuracy-Efficiency Trade-off

When the near-identical accuracies are related to computational cost, the models separate sharply. Table 4 augments the accuracy figures with the two efficiency ratios of Equations (5) and (6), and Figure 2 places the models in the accuracy-versus-cost plane. Among the three architectures evaluated, MobileNetV2 is Pareto-optimal: it attains the highest accuracy at the lowest cost, so no other model improves accuracy without increasing computation. Its parameter efficiency of 28.5 accuracy points per million parameters is roughly 7.3 times that of ResNet-50 and 24.8 times that of ViT-B/16, and its computational efficiency of 153.4 accuracy points per GFLOP is about 12.7 and 52.0 times greater, respectively. Figure 3 visualises these ratios directly. It should be emphasised that these ratios rank computational efficiency, not measured runtime; the deployment feasibility they imply is a computational estimate to be confirmed by on-device benchmarking.

Table 4. Accuracy normalised by computational cost, with Pareto status

Architecture	Acc (%)	Params (M)	GFLOPs	Size (MB)	Acc./ M-param	Acc./ GFLOPs	Status
MobileNetV2	99.69	3.50	0.65	14	28.48	153.37	Pareto-optimal
ResNet-50	99.59	25.56	8.27	98	3.90	12.04	Dominated
ViT-B/16	99.28	86.38	33.70	330	1.15	2.95	Dominated

[a] *Acc./M-param = test accuracy (%) divided by parameters (millions); Acc./GFLOP = test accuracy (%) divided by GFLOPs; higher is better. "Pareto-optimal" denotes a model not dominated by any other in the accuracy-cost plane (no competitor offers higher accuracy at equal or lower GFLOPs) among the three architectures evaluated.*

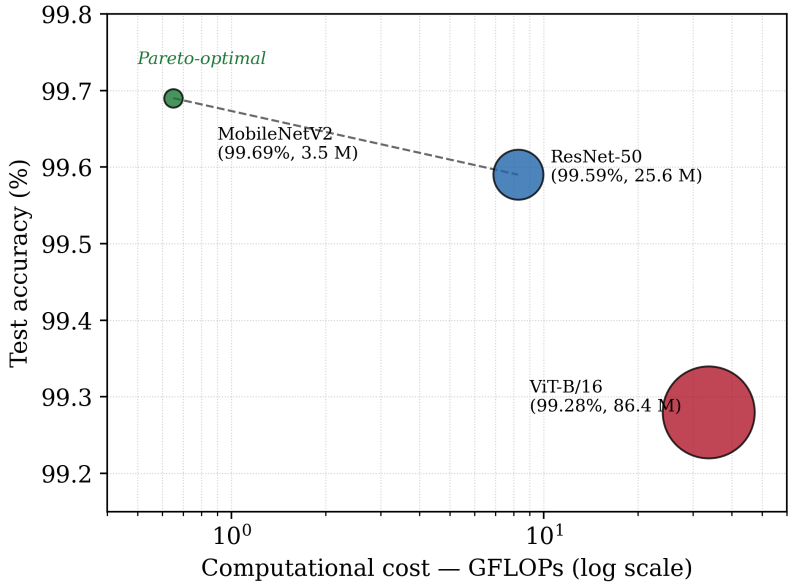

Fig. 2. Test accuracy versus computational cost (GFLOPs, log scale); marker area is proportional to parameter count

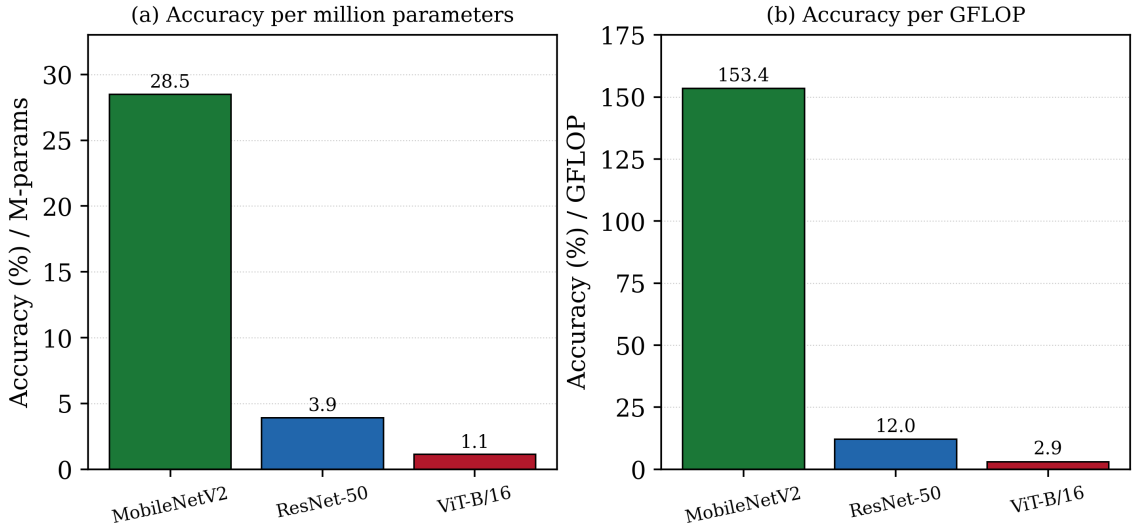

Fig. 3. Efficiency of the three architectures expressed as (a) accuracy per million parameters and (b) accuracy per GFLOP

The Vision Transformer is Pareto-dominated: it consumes the most parameters and computation yet returns the lowest accuracy. This outcome is consistent with the theoretical expectation that self-attention, lacking the local-connectivity and translation-equivariance priors of convolution, requires substantially larger training corpora to learn spatial structure effectively, so that on a medium-scale dataset of roughly ten thousand images its weak inductive bias becomes a liability rather than an asset. Qualitatively, the same effect was visible during optimisation: the convolutional models reached a stable validation plateau within the first several epochs, whereas the Transformer exhibited greater early-epoch volatility and required more epochs to stabilise. The practical implication is not that Transformers are unsuitable for agricultural vision in general; they have been applied effectively where data are abundant or where fusion mechanisms are introduced [9], [10]. Rather, for the data regime typical of this problem, their additional computational cost is not repaid in accuracy.

C. Per-Class Reliability for Safety-Critical Screening

Because a missed formalin-contaminated fruit carries a direct consumer-health consequence, the class-level behaviour of the Pareto-optimal MobileNetV2 was examined in detail. Table 5 reports its per-class precision, recall, and F1-score, and Figure 4 shows the corresponding confusion matrix on the test set. The rotten class was separated without error. The formalin class was detected with a recall of 99.68%, corresponding to a single false negative among 311 contaminated samples, together with two false positives in which fresh fruit was flagged as formalin-mixed. The residual error is thus concentrated almost entirely at the fresh-formalin boundary, which is consistent with the subtle, fine-grained micro-texture cues that distinguish chemically treated skin from natural fresh skin, as opposed to the gross morphological changes that make rotten fruit easy to separate.

Table 5. Per-class performance of the Pareto-optimal MobileNetV2 on the test set

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Fresh	99.73	99.45	99.59	365
Rotten	100.00	100.00	100.00	300
Formalin-mixed	99.36	99.68	99.52	311
Macro average	99.70	99.71	99.70	976

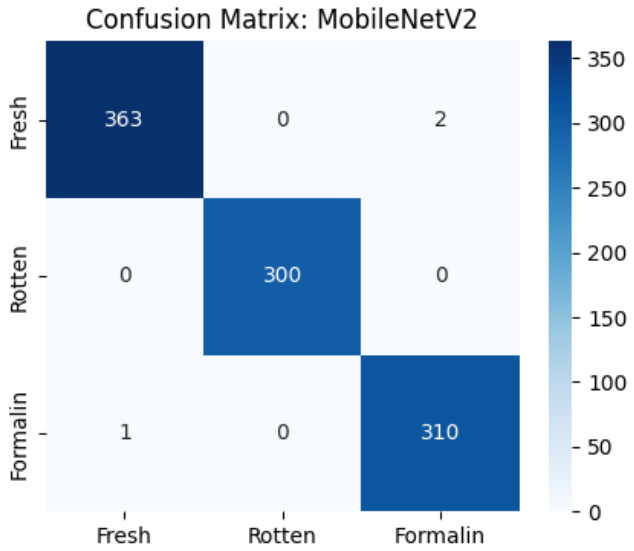

Fig. 4. Confusion matrix of MobileNetV2 on the test set (rows: true class; columns: predicted class)

This result is reported without overstatement. Although the false-negative count is very low, it is not zero: one contaminated fruit in the test set was predicted as fresh. In a consumer-safety context this single error is material, and it indicates that an automated screen of this kind should be positioned as a rapid pre-screening aid rather than a replacement for confirmatory chemical testing. Where the cost of a missed contaminant is high, the operating threshold for the formalin class can be lowered to trade a small increase in false positives for a further reduction in false negatives; the near-balanced precision and recall observed here suggest that such calibration is feasible without a large penalty.

D. Comparison with Prior Work

Table 6 positions the present findings against representative deep-learning studies of fruit quality and related image-classification tasks, reporting, where available, the dataset, number of classes, model, accuracy, F1-score, computational cost, and deployment setting; values not reported in the original studies are marked NR. Earlier fruit work generally addressed only fresh-versus-rotten discrimination with comparatively heavy backbones whose deployment cost was not quantified [4], [5], [6]. Comparative studies in adjacent domains, such as ConvMixer versus ResNet on the Kvasir dataset [18], a ResNet-50-based diagnostic model [19], and compact custom CNNs trained under minimal budgets [11], reinforce the value of evaluating competing paradigms under a common protocol and of reporting efficiency alongside accuracy [14], [15]. The distinctive element of this study is neither a new architecture nor a higher headline accuracy; it is the explicit demonstration that,

for the combined quality-and-formalin task, a compact model matches heavier alternatives while adding the formalin class and reporting its computational cost.

Table 6. Positioning relative to representative fruit-quality deep-learning studies

Study / approach	Detection focus	Reported accuracy	Cost reported
CNN transfer learning [4]	Fresh / rotten (2 classes)	~98%	No
Heavier CNN grading [8]	Quality grading	High	No
This study (MobileNetV2)	Fresh / rotten / formalin	99.69%	Yes (0.65 GFLOPs, 14 MB)

E. Practical Implications and Limitations

The deployment implication is computational rather than empirical. At 3.50 million parameters, 0.65 GFLOPs, and a 14 MB footprint, MobileNetV2 lies well within the resource envelope of contemporary smartphones and single-board devices, and such footprints are compatible with real-time inspection systems already deployed on low-cost hardware such as the Raspberry Pi [8]. However, no inference-latency or throughput measurement on a physical edge device was performed here, so on-site feasibility is inferred from computational proxies and remains to be confirmed by hardware benchmarking.

Several further limitations temper the conclusions and are stated plainly. First, the evaluation rests on a single dataset whose formalin class was produced by controlled induction; it may not represent the full range of adulteration concentrations, fruit-maturity stages, camera types, and environmental conditions found in real markets, so domain shift and dataset bias are genuine risks. Second, the uniformly high accuracy (99.28-99.69%) may partly reflect the separability of a curated dataset; although the instance-segmentation step removes background and thereby reduces the risk of background-based shortcuts, only cross-dataset or external validation can establish that the models rely on genuine fruit-surface cues, and such validation was not performed. Third, the comparison covers three representative paradigms rather than the full space of modern architectures, and all results derive from single training runs, so the reported differences should be read as indicative rather than statistically established. Finally, the reliance on visual cues that can be masked by illumination means the method is best understood as a first-line screening aid that complements, and does not replace, laboratory chemical confirmation [16], [17].

IV. Conclusion

This study asked not which architecture is most accurate for multi-class fruit-quality and formalin classification, but which is most justified once computational cost is taken into account. Under an identical preprocessing and training protocol on 10,154 FruitVision images, three paradigms of widely differing scale produced near-saturated and closely spaced accuracies (99.28-99.69%), yet differed in cost by up to two orders of magnitude. Among the three architectures evaluated, the lightweight MobileNetV2 was Pareto-optimal, achieving the highest accuracy (99.69%) at roughly one-seventh the parameters and one-thirteenth the computation of ResNet-50, while the Vision Transformer was Pareto-dominated, delivering the lowest accuracy at the highest cost on this medium-scale dataset. For the formalin class the compact model achieved 99.68% recall with a single residual false negative, which argues for its use as a preliminary screening aid alongside confirmatory chemical testing rather than as a standalone decision-maker. The broader implication is that, under the experimental conditions and single dataset used in this study, the lightweight model is computationally preferable and a promising candidate for embedded, on-site screening, subject to confirmation on real edge hardware and independent datasets. Because the accuracy differences are small and derive from single runs, they are not claimed as definitive. Future work will measure on-device latency, incorporate repeated runs with statistical testing, pursue cross-dataset validation, extend the evaluation to a wider range of fruit species, contaminants, and acquisition conditions, and examine post-training quantisation to compress the model further.

References

[1] S. V Binorkar and M. S. Bhoyar, “Toxicological Insights into Food Additives and FP-Impact, Trends, and Resolves,” *Annals of Ayurvedic Medicine*, vol. 14, no. 2, pp. 164–174, 2025, doi: 10.5455/AAM.

[2] Y. Cui, Y. Cheng, Z. Xu, B. Li, W. Tian, and J. Zhang, "Cellulose-Based Transparent Edible Antibacterial Oxygen-Barrier Coating for Long-Term Fruit Preservation," *Advanced Science*, vol. 11, no. 48, Dec. 2024, doi: 10.1002/advs.202409560.

[3] C. Wang, X. Li, Z. Zhang, X. Luo, J. Cai, and A. Wang, "Nondestructive Quality Detection of Characteristic Fruits Based on Vis/NIR Spectroscopy: Principles, Systems, and Applications," Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/agriculture15202167.

[4] U. Amin, M. I. Shahzad, A. Shahzad, M. Shahzad, U. Khan, and Z. Mahmood, "Automatic Fruits Freshness Classification Using CNN and Transfer Learning," *Applied Sciences (Switzerland)*, vol. 13, no. 14, Jul. 2023, doi: 10.3390/app13148087.

[5] M. H. I. Bijoy, S. Z. Tasnim, S. A. Awsaf, and M. Z. Hasan, "FruitVision: A benchmark dataset for fresh, rotten, and formalin-mixed fruit detection," *Data Brief*, vol. 61, Aug. 2025, doi: 10.1016/j.dib.2025.111752.

[6] P. Dhiman *et al.*, "A Novel Deep Learning Model for Detection of Severity Level of the Disease in Citrus Fruits," *Electronics (Switzerland)*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/electronics11030495.

[7] Z. Cao and Z. Cao, "Design of a MobilNetV2-Based Retrieval System for Traditional Cultural Artworks," *Int. J. Gaming Comput. Mediat. Simul.*, vol. 16, no. 1, pp. 1–17, 2023, doi: 10.4018/IJGCMS.334700.

[8] W. Nugroho, R. Zahabiyah, M. J. F. Arifiant, and A. Afianto, "Automated Component Detection for Quality PCB Using YOLO Algorithm with IoT Real-Time Streaming on Raspberry Pi," *JURNAL INFOTEL*, vol. 17, no. 2, Jul. 2025, doi: 10.20895/infotel.v17i2.1313.

[9] U. Barman *et al.*, "ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture," *Agronomy*, vol. 14, no. 2, Feb. 2024, doi: 10.3390/agronomy14020327.

[10] H. M. Albarakati *et al.*, "A Novel Deep Learning Architecture for Agriculture Land Cover and Land Use Classification from Remote Sensing Images Based on Network-Level Fusion of Self-Attention Architecture," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 6338–6353, 2024, doi: 10.1109/JSTARS.2024.3369950.

[11] Z. Muthi'ah, Oktalia Triananda Lovita, Oktalia Triananda Lovita, and Mohd Iqbal Muttaqin, "Deep Learning Based Classification of TB and Normal Chest X-rays Using a Custom CNN with Minimal Epoch Training," *Jurnal Inotera*, vol. 10, no. 2, pp. 272–279, Jul. 2025, doi: 10.31572/inotera.vol10.iss2.2025.id503.

[12] M. Ben Ammar, "Enhancing real-time instance segmentation for plant disease detection with an improved YOLOv8-seg algorithm," Int. J. Inf. Technol. Secur., vol. 16, no. 2, pp. 27–38, 2024, doi: https://doi.org/10.59035/BCNL3199

[13] G. Al-refai, H. Elmoaqet, A. Al-Refai, A. Alzu'bi, T. Al-Hadhrami, and A. Alkhateeb, "Two-stage object detection in low-light environments using deep learning image enhancement," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.2799.

[14] M. H. Ashraf *et al.*, "HIRD-Net: an explainable CNN-based framework with attention mechanism for diabetic retinopathy diagnosis using CLAHE-D-DoG enhanced fundus images," Life, vol. 15, no. 9, art. 1411, 2025, doi: 10.3390/life15091411.

[15] Y. S. Cho and P. C. Hong, "Applying Machine Learning to Healthcare Operations Management: CNN-Based Model for Malaria Diagnosis," *Healthcare (Switzerland)*, vol. 11, no. 12, Jun. 2023, doi: 10.3390/healthcare11121779.

[16] M. Abedin, S. Islam, and N. Sultana, "Comprehensive data of 10 fruit leaf classes captured for agricultural AI applications," *Data Brief*, vol. 61, Aug. 2025, doi: 10.1016/j.dib.2025.111879.

[17] A. Bhatt and M. Joshi, "De dicate d dataset of Sapota (Manilkara zapota) fruit for machine vision applications," *Data Brief*, Jul. 2025, doi: 10.17632/jgtb95x6kf.2.

[18] Yuliana et al., "Comparison of ConvMixer method and ResNet method in classification and detection of gastrointestinal diseases using Kvasir dataset," Jurnal Inotera, vol. 10, no. 2, pp. 288–297, 2025, doi: 10.31572/inotera.Vol10.Iss2.2025.ID493

[19] V. Cambay et al., "Automated detection of gastrointestinal diseases using a ResNet-50-based explainable deep feature engineering model with endoscopy images," Sensors, vol. 24, no. 23, art. 7710, 2024, doi: 10.3390/s24237710.