Article

Impact of Dataset Quality on Deep Learning Models for Dragon Fruit and Leaf Health Classification

Shahnawaz Ayoub [1,*], Imran Baig [2], Mudasir Ashraf [3], and Mahmoud Okasha [4,5]

[1] Glocal School of Science and Technology, Glocal University, Saharanpur, Uttar Pradesh 247121, India
[2] Cardiff School of Technologies, Cardiff Metropolitan University, Llandaff Campus, Western Avenue, Cardiff CF5 2YB, The UK
[3] School of engineering and IT, Manipal academy of Higher Education, Dubai, 345050, United Arab Emirates
[4] Agricultural Engineering Research Institute, Agricultural Research Center, Dokki, Giza 12611, Egypt
[5] Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Udine I-33100, Italy

* Correspondence: shahnawazayoub@outlook.com (S.A.)

Abstract

Accurate assessment of fruit and leaf health is essential for early disease detection, quality grading, and automated management in commercial dragon fruit production. Variability in illumination, symptom intensity, and morphological features often limits the reliability of conventional machine learning models trained on raw datasets. This study evaluates the effect of dataset quality on deep learning performance using a publicly available dragon fruit and leaf dataset containing 4,518 images across four classes: Healthy Fruit, Healthy Leaves, Infected Fruits, and Infected Leaves. Three dataset versions were constructed (i) the original dataset, (ii) an augmented dataset expanding each image threefold, and (iii) a cleaned augmented dataset created by removing mislabeled, ambiguous, or low-quality samples. Four deep architectures (MobileNetV3, InceptionV3, ResNet101, and VGG16) were trained under identical settings to assess classification performance. Across all models, the cleaned augmented dataset produced the most stable training behavior and highest accuracy. InceptionV3 achieved the strongest overall performance with an F1-score above 0.95 and validation accuracy approaching 0.97, while MobileNetV3 delivered competitive results (accuracy 0.9613) with minimal computational cost. Confusion matrices confirmed major reductions in fruit–fruit and leaf–leaf misclassification after dataset cleaning. The findings highlight that targeted data refinement, combined with augmentation, is critical for building reliable deep learning models for real-world agricultural applications.

Keywords: dragon fruit classification; deep learning; dataset augmentation; image cleaning; MobileNetV3; InceptionV3; plant disease detection

Citation: Ayoub S., Baig I., Ashraf M., Okasha M. Impact of Dataset Quality on Deep Learning Models for Dragon Fruit and Leaf Health Classification. *Impact in Agriculture*. **2025**, *1*, 1. https://doi.org/10.65500/agriculture-2025-001

Received: 16 August 2025 | Revised: 17 September 2025 | Accepted: 25 September 2025 | Published: 13 October 2025

1. Introduction

Dragon fruit has become an increasingly important tropical crop due to its nutritional value, expanding consumer demand, and suitability for semi-arid and subtropical regions. Although grown across Asia and Latin America, global production is dominated by Vietnam, which harvests more than one million metric tonnes

annually and serves as the leading exporter to China, the United States, the European Union, and several Asian markets [1]. Its rising trade value reflects strong international demand for antioxidant-rich fruits and has encouraged adoption by commercial growers. The plant's low water requirements, compatibility with degraded soils, and potential for multiple harvest cycles have further supported its rapid expansion in emerging agricultural economies [2].

Dragon fruit belongs to the *genera Hylocereus* and *Selenicereus*, with white-, red-, pink-, and yellow-fleshed varieties cultivated commercially. Despite its adaptability, production presents considerable agronomic challenges. The crop is sensitive to fluctuations in temperature, humidity, and soil conditions, while successful cultivation depends on appropriate trellis design, pollination management, and protection from environmental stress. Disease incidence, inconsistent fruit quality, and unreliable maturity assessment continue to limit productivity and export readiness. These issues directly affect market grade, post-harvest lifespan, and compliance with international quality standards.

A diverse set of biotic and abiotic threats affects both yield and fruit marketability. Bacterial infections such as soft rot and wilt spread rapidly in humid climates and can destroy entire plant sections [3]. Fungal pathogens including anthracnose and stem canker frequently affect branches and fruits, with symptoms that are difficult to detect in early stages. Additional problems arise from insect pests, sunburn, and mechanical injury, all of which alter external appearance and lower commercial grade. Because growers typically rely on visual inspection, variation in fruit morphology, lighting, and disease progression complicates judgement and introduces subjectivity into routine decision-making.

Accurate maturity assessment remains a major challenge. Manual grading depends on external color, spine condition, and subtle texture cues that vary across cultivars and environmental contexts. Inaccurate decisions lead to premature or delayed harvesting and non-uniformity in packing lines, undermining export consistency. Meeting international standards requires objective, reproducible indicators of fruit quality, size, firmness, and internal sweetness, creating a strong need for automated assessment technologies.

Deep learning provides an effective alternative to manual inspection and has shown strong performance in plant recognition [4], fruit detection [5], fruit quality detection [6] and maturity classification [7]. Convolutional neural networks have demonstrated high accuracy across diverse horticultural species, including dragon fruit [8]. Vision transformers and hybrid attention mechanisms have further improved performance on complex agricultural imagery by modelling long-range spatial relationships and subtle texture variations [9–11]. These advances have enabled new applications in automated sorting, robotic harvesting, and ecological monitoring.

However, the effectiveness of deep models depends heavily on the quality and diversity of training data. Dragon fruit exhibits substantial visual variability arising from cultivar differences, environmental lighting, occlusions, and ripening stages. Leaves display subtle differences between healthy and mildly infected states, while fruits show irregular shapes and evolving color patterns. Many existing datasets are small, laboratory-controlled, or limited to single tasks such as ripeness or disease classification, limiting generalization to real field conditions. Few datasets integrate both fruit and leaf organs or include multiple disease types within a unified structure, leaving significant gaps for developing robust classification models.

The recently published dragon fruit and leaf dataset addresses several of these limitations by providing 4518 annotated images representing a wide range of fruit and leaf conditions, including bacterial and fungal diseases, wilting, insect damage, sunburn, and healthy samples. Its diversity in environmental settings supports the development of models resilient to illumination changes, shadows, occlusions, and natural variation. Such datasets are increasingly essential as precision agriculture frameworks adopt automated diagnostics, robotic systems, and digital quality assessment pipelines to enhance grading efficiency and early disease detection [12].

The present study uses this dataset to evaluate how deep learning models handle four health classes across fruit and leaf samples under real field variation. The work examines the effect of data quality on accuracy, stability, and common error patterns. It focuses on practical challenges such as uneven illumination, complex backgrounds, and early symptom manifestation. The study provides a systematic analysis of how model performance evolves when transitioning from raw data to augmented data and subsequently to cleaned data. This analysis supports the development of simple and reliable tools for crop monitoring and routine inspection in production environments. A three-step data processing pipeline is proposed to quantify the impact of data quality on training and testing performance. Four models are trained under

identical experimental settings to ensure fair comparison. The results demonstrate that removing incorrect or ambiguous samples improves model stability and reduces misclassification, offering practical guidance for deploying lightweight and accurate systems in field conditions.

The main contributions are as follows.

- A three step dataset process is created to test how data quality shapes training and prediction results.
- Four deep learning models are trained under the same settings to provide a fair and direct benchmark.
- Cleaning of wrong and unclear samples is shown to improve accuracy and reduce common class errors.
- Practical insight is given for simple and lightweight systems that work in basic field and grading tasks.

2. Related Work

2.1 Deep learning in agriculture and fruit analytics

Deep learning has become a central method in agricultural automation for tasks such as plant disease recognition, fruit grading and yield estimation. A review of sixty one studies reported high performance for convolutional models including VGG, AlexNet, GoogleNet, ResNet and MobileNet, with accuracy levels often exceeding 98% across multiple crops including dragon fruit and mango [13]. These findings indicate that deep learning can replace manual inspection in many production settings and can enhance decision making in modern farming systems.

Several studies have attempted to reduce annotation cost and improve domain transfer for fruit detection. Domain adaptation has been explored through generative adversarial networks, cycle translation models and anchor free detection frameworks that transfer labels across species and fruit shapes [14,15]. Other research investigated kernel optimization for drone based detection, where adjustments in the shallow convolution layers of YOLOv5 improved performance for small and occluded fruit targets under aerial conditions [16]. These studies confirm that practical deployments require models that remain robust under variation in illumination, background noise and sample morphology.

2.2 Dragon fruit quality and maturity grading from RGB images

Earlier work on dragon fruit focused on external appearance and export grading. A convolutional neural network trained on images collected along packing lines achieved more than 96% accuracy for quality categories when combined with weight data from load cells [17]. Machine learning classifiers such as CNN, artificial neural networks and support vector machines have also been applied to features related to size, shape, colour and visible defects, producing effective grading and sorting results for commercial environments [18]. Species classification has been examined in the Thai context where a lightweight deep model separated seven dragon fruit species in laboratory and outdoor scenes with high accuracy [19].

Recent studies have applied deeper and more diverse architectures. Vision Transformer, VGG16, ResNet50, EfficientNet, Xception and InceptionV3 have been compared for quality grading and maturity recognition, with transformer based models achieving strong performance for quality and VGG16 remaining competitive for maturity detection [20]. Hybrid feature extraction approaches have combined DenseNet50 and ResNet50 with principal component analysis and support vector machine ensembles, reaching classification accuracy close to 98% for ripeness prediction [21]. Additional work fused features from multiple pretrained CNNs before classification with random forest, gradient boosting and AdaBoost to obtain high accuracy for maturity and quality grading [22].

Interpretability has also been emphasized. ResNet and Vision Transformer models have been paired with Grad CAM and attention maps to classify shelf life stages and explain morphological cues in the decision process [23]. Transfer learning with DenseNet201 has been used to classify ripeness stages, with guided visualization confirming that the model attends to relevant fruit regions [24]. Embedded systems have also been examined. A real time identification framework using EfficientNet and YOLOv8 on Raspberry Pi hardware achieved reliable performance for maturity, size and defect detection in field settings [25]. Collectively, these studies confirm that high accuracy can be obtained for fruit level grading tasks under controlled or semi controlled conditions.

2.3 Dataset contributions and graph based models

Dataset availability remains a constraint in dragon fruit research. A high resolution image dataset was introduced to support maturity and quality grading, with samples collected in multiple orchards under expert supervision [26]. This dataset has been proposed as a foundation for robotic harvesting and packaging research.

A more recent line of work examined richer image representations. A graph based classification framework combined superpixel segmentation, convolutional features and graph convolution networks to model structural relationships within images [27]. Experiments on tomato

disease, dragon fruit, tomato ripeness and a dragon fruit and leaf dataset showed that the graph based method outperformed standard CNNs and Vision Transformer models, especially on the more difficult datasets, suggesting that relational feature modelling enhances robustness under high intra class variation.

2.4 Leaf diseases and plant health modelling

Although fruit level analysis dominates current work, some studies have addressed leaf based disease identification. A semantic connection based learning method allowed incremental addition of new disease classes without degrading performance on earlier categories and achieved 92% accuracy after updates [28]. Leaf and fruit data have also appeared as part of broader benchmarks, although emphasis remained on classification accuracy rather than detailed plant health assessment [29]. Overall, existing leaf studies cover limited disease categories and do not combine leaf and fruit symptoms within a unified dataset.

2.5 Detection, counting and robotic harvesting in orchards

A considerable body of research has investigated detection and localization of dragon fruit in natural orchards. Lightweight models such as YOLOv4 LITE have been proposed to reduce computation through MobileNetv3 backbones and multi scale prediction, enabling accurate detection under occlusion while remaining suitable for embedded devices [30]. Improved YOLOX models with attention modules have also been introduced to handle variation in lighting and background complexity while maintaining fast inference [31].

Robotic harvesting has attracted similar attention. Multi pose detection using an optimal YOLOv7 tiny model has been used to identify fruits under strong, weak and artificial lighting, supporting real time operation on mobile devices [32]. A related approach combined YOLOv7 with PSPNet segmentation and ellipse based endpoint detection to extract head and root positions for robotic grasping [33]. Counting and growth monitoring have been addressed through video stream analysis, where YOLOv5 detection and improved ByteTrack tracking produced high accuracy for classification and counting of flowers, green fruits and mature fruits [27]. These studies highlight substantial progress in robotic applications but typically limit analysis to fruit level targets.

2.6 Navigation, remote sensing and yield estimation

Navigation and large scale monitoring have been explored through semantic segmentation and remote sensing. An improved DeepLabV3 plus model with attention components provided accurate road segmentation for navigation in dragon fruit orchards under varying illumination [34]. Unmanned aerial vehicle imagery has been used to identify dragon fruit plants and estimate yields in complex mountainous habitats. Vegetation indices, canopy height models and U Net based segmentation produced high accuracy across multiple ecological scenes [35]. A related study constructed a small sample dataset of dragon fruit trees from close range UAV images and demonstrated that U Net models trained on augmented samples achieved high recognition precision and improved robustness [36].

2.7 Multisensor and multimodal approaches

Multimodal sensing has also been investigated. Electrical impedance spectroscopy and texture features have been fused within a convolutional framework to estimate internal quality of yellow pitahaya, yielding accurate predictions of degrees Brix [37]. Dragon fruit peel has been used as a pH sensitive indicator film for fish freshness detection, where a smartphone based multimodal network combined color features with chemical indicators to reach almost perfect classification accuracy [38]. These approaches illustrate the promise of multimodal sensing but do not address full plant health modelling.

2.8 Synthesis and research gap

Across all studies, strong progress has been made in quality grading, maturity recognition, species classification, disease identification, detection for robotic harvesting, navigation and yield estimation [15–27,30–39] . Most research focuses on fruit level tasks or on a limited number of leaf diseases. Only a few datasets combine fruits and leaves, and these studies emphasize accuracy rather than broad representation of plant health conditions [26,29]. Training often occurs in controlled environments, which restricts generalization to diverse field conditions. Transformer based models and graph based representations show potential but have not been benchmarked comprehensively on unified fruit and leaf datasets. Incremental learning and domain adaptation methods remain unexplored in settings that involve multiple disease categories, physiological stresses and healthy samples across both organs of the plant [14,15,28].

These limitations indicate a need for integrated datasets and models that capture a wider spectrum of health conditions in fruits and leaves together. Such datasets support more complete modelling of plant condition and strengthen applications in disease

management, ecological assessment and automated monitoring.

3. Material and Methods

3.1 Dataset Description

The study used a public dragon fruit and leaf dataset collected under natural field conditions [12]. Images show healthy samples, bacterial infection, fungal infection, insect damage, wilting, and sunburn. Light, angle, and background vary across samples, which supports training models for real environments. Figure 1 shows example images.

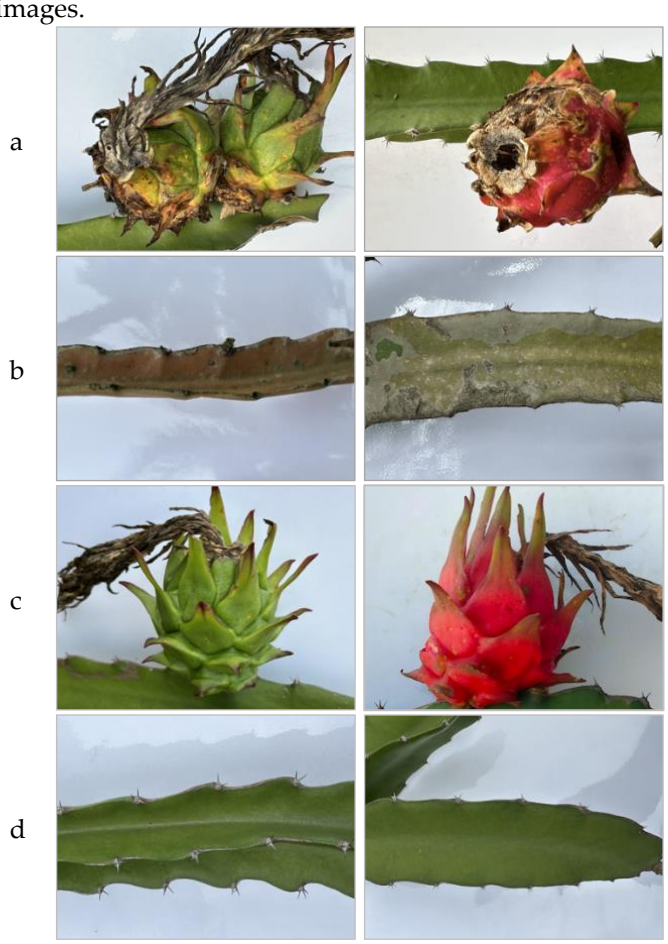


Figure 1. Samples of Dragon Dataset (a) Healthy Fruit (b) Healthy Leaves (c) Healthy Fruit (d) Infected Leaves

The original dataset contains 4518 images and four classes. Healthy Leaves 2242, Healthy Fruits 333, Infected Leaves 1615, Infected Fruits 307. These were split into 3146 for training, 674 for validation, and 677 for testing as listed in Table 1. This set provided the baseline.

The augmented dataset increased diversity through rotation, shift, and zoom. Each original image was expanded into three images, giving a total of 11024. Healthy Leaves 4270, Healthy Fruits 999, Infected Leaves 4834, Infected Fruits 921. The split was 7715 training, 1653 validation, and 1656 testing as shown in Table 2. This version improved variation in shape and appearance.

Table 1. Details of the original dataset

Class Name	Training	Validation	Test	Total
Healthy Fruit	233	50	50	333
Healthy Leaves	1569	336	337	2242
Infected Fruits	235	46	47	328
Infected Leaves	1130	242	243	1615
Total	3167	674	677	4518

Table 2. Details of the augmented dataset

Class Name	Training	Validation	Test	Total
Healthy Fruit	699	150	150	999
Healthy Leaves	2989	640	641	4270
Infected Fruits	644	138	139	921
Infected Leaves	3383	725	726	4834
Total	7715	1653	1656	11024

The original dataset was first divided into training, validation, and test subsets as shown in Table 1. Data augmentation was then applied separately within each subset based on this initial split, ensuring that all augmented images derived from an original sample remained in the same subset and that no augmented variants appeared across different subsets. A cleaned dataset was subsequently prepared from the augmented images using a structured visual screening process. Each image was manually inspected by a single experienced researcher with prior experience in agricultural image datasets and plant disease classification. Incorrect samples were removed, including leaves appearing in fruit folders, fruits appearing in leaf folders, healthy images placed in infected folders, and infected images without visible symptoms. Ambiguous samples with unclear visual cues caused by shadowing, motion blur, occlusion, or illumination artifacts were also excluded. Images labeled as infected but lacking visible disease indicators, such as lesions or discoloration, were removed. No relabeling was performed during this process. The final cleaned dataset (Table 3) contained 10,144 images, comprising 3,603 healthy leaves, 948 healthy fruits, 4,732 infected leaves, and 861 infected fruits. The dataset was split into 7,099 training images, 1,521 validation images, and 1,524 testing images.

Table 3. Details of the augmented and cleaned dataset

Class Name	Training	Validation	Test	Total
Healthy Fruit	663	142	143	948
Healthy Leaves	2522	540	541	3603
Infected Fruits	602	129	130	861
Infected Leaves	3312	710	710	4732
Total	7099	1521	1524	10144

Using three datasets allowed stepwise analysis of data quality. The original set tested baseline behavior. The augmented set improved generalisation. The cleaned set raised label purity and produced the most stable results.

3.2 Data Preprocessing

All images were prepared with a simple and consistent pipeline. Each image was resized to the input size required by the tested models. Pixel values were scaled to a fixed range to keep training stable across classes and datasets. The same steps were applied to fruits and leaves.

The original dataset was already divided into training, validation, and test sets. These splits were kept unchanged across all experiments to allow fair comparison. The augmented dataset followed the same split ratio in its source files. The cleaned dataset followed the same structure, with wrong samples removed before splitting. The cleaning process followed a conservative exclusion strategy, where only clearly inconsistent or visually unreliable samples were removed, ensuring that the retained dataset favored label precision over dataset size. Images previously identified as visually inconsistent or incorrectly labeled during the cleaning stage were excluded prior to dataset splitting. No further manual correction or relabeling was applied.

All datasets were stored in separate folders. Each folder held one class. This layout allowed direct loading through standard data loaders. No extra balancing or resampling was applied. The goal was to evaluate the effect of raw data quality and size, not synthetic reweighting.

The same preprocessing steps were used for MobileNetV3, InceptionV3, VGG16, and ResNet101. This ensured that changes in performance came from data quality and model design, not from preprocessing differences.

3.3 Model Architectures

Four deep models were used to evaluate performance under different data conditions. These were MobileNetV3, InceptionV3, VGG16, and ResNet101. Each model offers a distinct feature extraction style. This allowed a clear view of how network depth and structure affect fruit and leaf classification.

MobileNetV3 uses lightweight blocks that reduce computation while keeping stable feature quality. InceptionV3 uses parallel filters to capture patterns at multiple scales. VGG16 applies stacked 3 by 3 filters that form a simple and steady feature pipeline. ResNet101 uses residual blocks to support deeper learning without loss of gradient flow.

All models were loaded with pretrained weights. The final classifier was replaced with a four class output layer. Images were resized to the standard input size of each model. No other architectural changes were made. The goal was to compare these models under identical training settings and across the three datasets.

3.4 Training Setup

All models were trained under identical settings to ensure fair comparison. The learning rate was fixed at 1×10^{-5} after preliminary experiments with values ranging from 1×10^{-1} to 1×10^{-6}, where higher learning rates led to unstable validation loss and lower values resulted in slower convergence without performance gains. Batch sizes of 8, 16, and 32 were evaluated; a batch size of 16 was selected as it provided stable validation performance while reducing training time and memory usage. The number of epochs was set to 100 to allow observation of full training and validation behavior.

Early stopping was intentionally not applied to the original dataset in order to explicitly observe the onset of overfitting under limited data conditions. In contrast, early stopping was applied to the augmented and cleaned datasets to prevent unnecessary overfitting once sufficient data diversity was available and to reflect practical training scenarios, where training is terminated when the validation loss begins to rise. All models used the same optimizer and cross-entropy loss function. The same preprocessing steps and data loaders were used for each run to ensure fair comparison. Validation was performed at the end of each epoch, and test evaluation was conducted once per dataset after training was completed.

Models were trained on Windows 10 Pro with an Intel i5 processor at 2.9 GHz, 16 GB RAM, and an Nvidia GeForce GTX 1660 GPU. Python version 3.8 was used with OpenCV version 4.7 and Keras version 2.8. This setup provided stable timing and reproducible runs.

Evaluation used accuracy, precision, recall, and f1 score. These metrics were computed for all four classes. Macro averages gave equal weight to each class. Weighted averages reflected the class distribution. Confusion

matrices were prepared to show correct and wrong predictions at class level. These outputs supported a clear view of model performance under the original, augmented, and cleaned datasets.

4. Results and Discussion

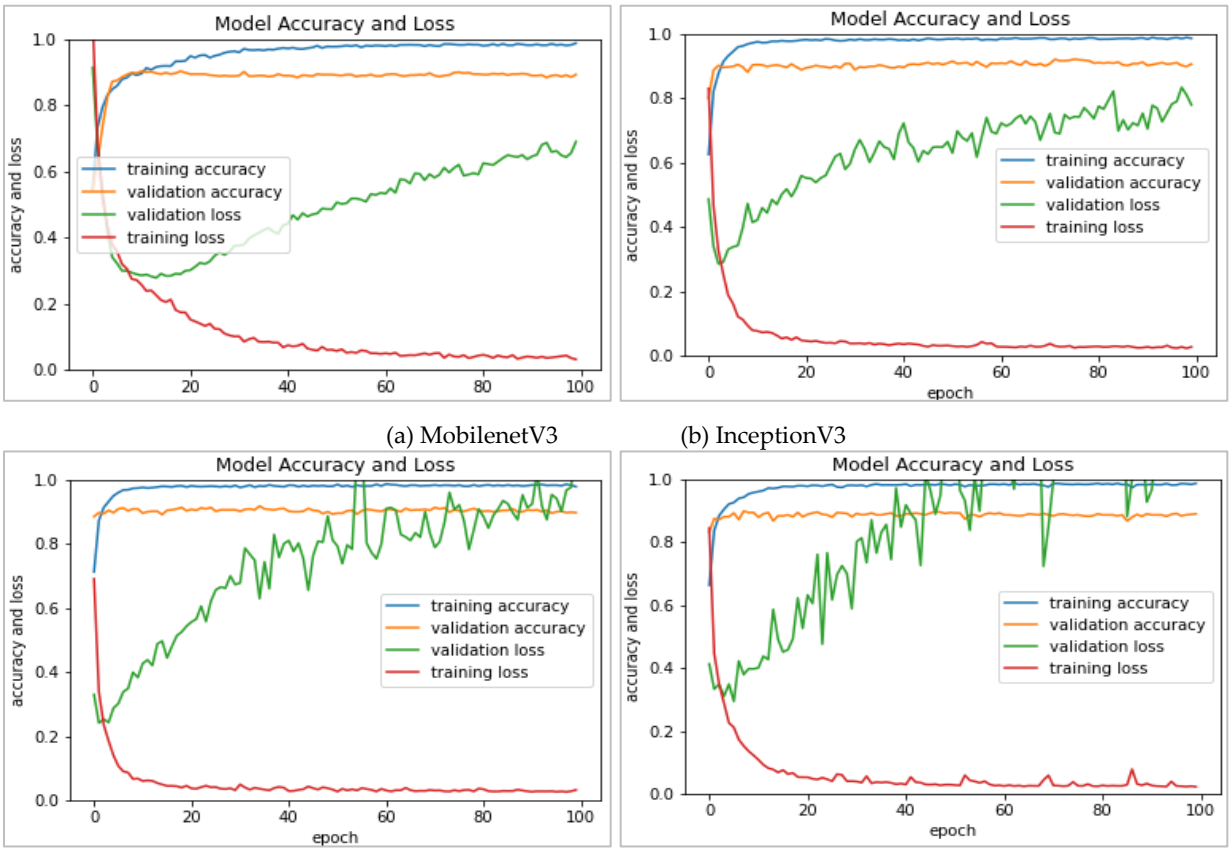
This section presents the results from all four deep learning models on the three dataset versions. The analysis covers training and validation behavior, test performance, and error patterns. The models were evaluated on the original dataset, the augmented dataset, and the cleaned augmented dataset. Each experiment used the same training settings to ensure a fair comparison. The results are grouped into three parts. The first part reports model behavior on the original dataset. The next two parts describe the effect of augmentation and cleaning on accuracy, stability, and class wise performance. The discussion highlights how data quality and class distribution shape model strengths and limitations.

4.1 Performance Evaluation on the Original Dataset

Figure 2(a–d) presents the training and validation curves for MobileNetV3, InceptionV3, ResNet101, and VGG16 on the original dataset. All four models converged rapidly within the first 10–15 epochs, but their validation behavior differed noticeably. MobileNetV3 showed the most stable pattern, with smooth validation accuracy and a steadily declining validation loss. This indicates strong generalization and minimal sensitivity to noise in fruit-level samples. InceptionV3 converged to a similar level but exhibited moderate oscillation in validation loss after epoch 20, suggesting a higher sensitivity to illumination variation and background clutter. ResNet101 achieved the highest training and validation accuracy among the models, with a sharp reduction in loss and consistent validation accuracy around 0.91. Although its validation loss fluctuated, the amplitude remained lower than VGG16, indicating a better balance between capacity and generalization. In contrast, VGG16 showed the highest instability: validation loss oscillated heavily throughout training, even as accuracy held near 0.89. This confirms that VGG16 overfit more than the other architectures, largely due to its large parameter count relative to the size of the dataset.

Figure 3(a–d) shows the confusion matrices for the four models. Across all architectures, the most frequent misclassification occurred between Healthy Fruit and Infected Fruits. This aligns with the visual similarity of early-stage infection symptoms, where discoloration and surface texture changes are subtle. Healthy Leaves were consistently identified with high confidence, driven by a larger sample size and clearer structural features. Infected Leaves showed moderate confusion across models, particularly under strong shadow or variable color transitions, which creates overlap with Healthy Leaves in feature space. ResNet101 produced the cleanest confusion matrix, followed by MobileNetV3. VGG16 and InceptionV3 displayed more dispersion along the off-diagonal entries, consistent with the instability observed in their validation loss curves.

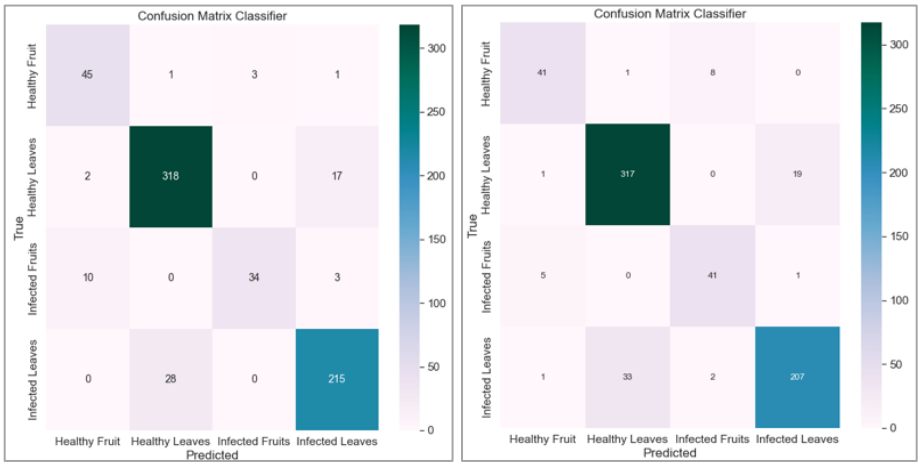
Table 4 summarizes the class-wise precision, recall, and F1 scores for all four models. ResNet101 achieved the strongest overall performance, reflected in its balanced precision–recall profile across all classes and the highest accuracy of 0.9129. MobileNetV3 also performed well, with high precision for Infected Fruits but lower recall, indicating that it tends to miss darker or low-contrast infected samples. InceptionV3 and VGG16 produced similar overall accuracy (0.8951), but VGG16 showed weaker stability and larger variability across classes, especially for Infected Fruits. The results in Table 4 are consistent with the training behaviors shown in Figure 2 and the error distributions in Figure 3. Overall, the performance trends indicate that model behavior on the original dataset is constrained by class imbalance and visual overlap among fruit-level samples, while leaf-level classes remain easier to separate.



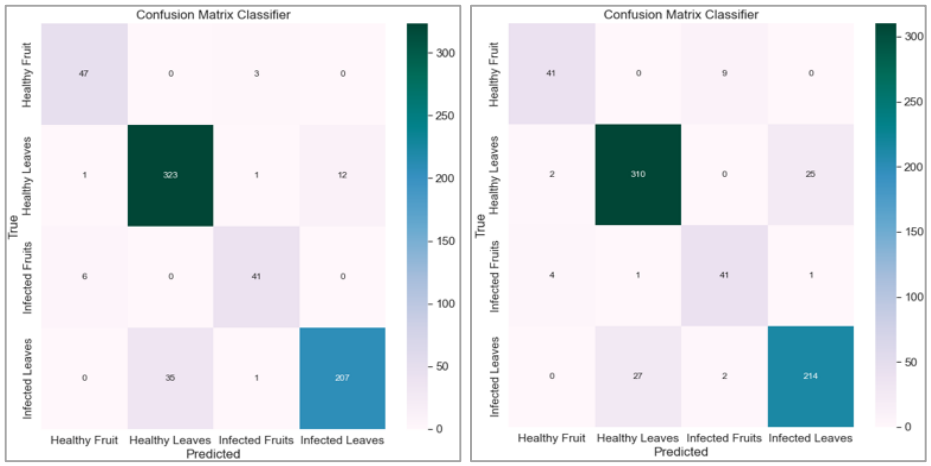
(a) MobilenetV3 (b) InceptionV3

(c) ResNet101 (d) VGG16

Figure 2. Training and validation curves for original dataset



(a) MobilenetV3 (b) InceptionV3



(c) ResNet101 (d) VGG16

Figure 3. Confusion matrices for original dataset

Table 4. Class wise metrics for original dataset

Model	Metric	Healthy Fruit	Healthy Leaves	Infected Fruits	Infected Leaves
MobileNetV3	Precision	0.789474	0.916427	0.918919	0.911017
	Recall	0.900000	0.943620	0.723404	0.884774
	F1 Score	0.841121	0.929825	0.809524	0.897704
InceptionV3	Precision	0.854167	0.903134	0.803922	0.911894
	Recall	0.820000	0.940653	0.872340	0.851852
	F1 Score	0.836735	0.921512	0.836735	0.880851
ResNet101	Precision	0.870370	0.902235	0.891304	0.945205
	Recall	0.940000	0.958457	0.872340	0.851852
	F1 Score	0.903846	0.929496	0.881720	0.896104
VGG16	Precision	0.872340	0.917160	0.788462	0.891667
	Recall	0.820000	0.919881	0.872340	0.880658
	F1 Score	0.845361	0.918519	0.828283	0.886128

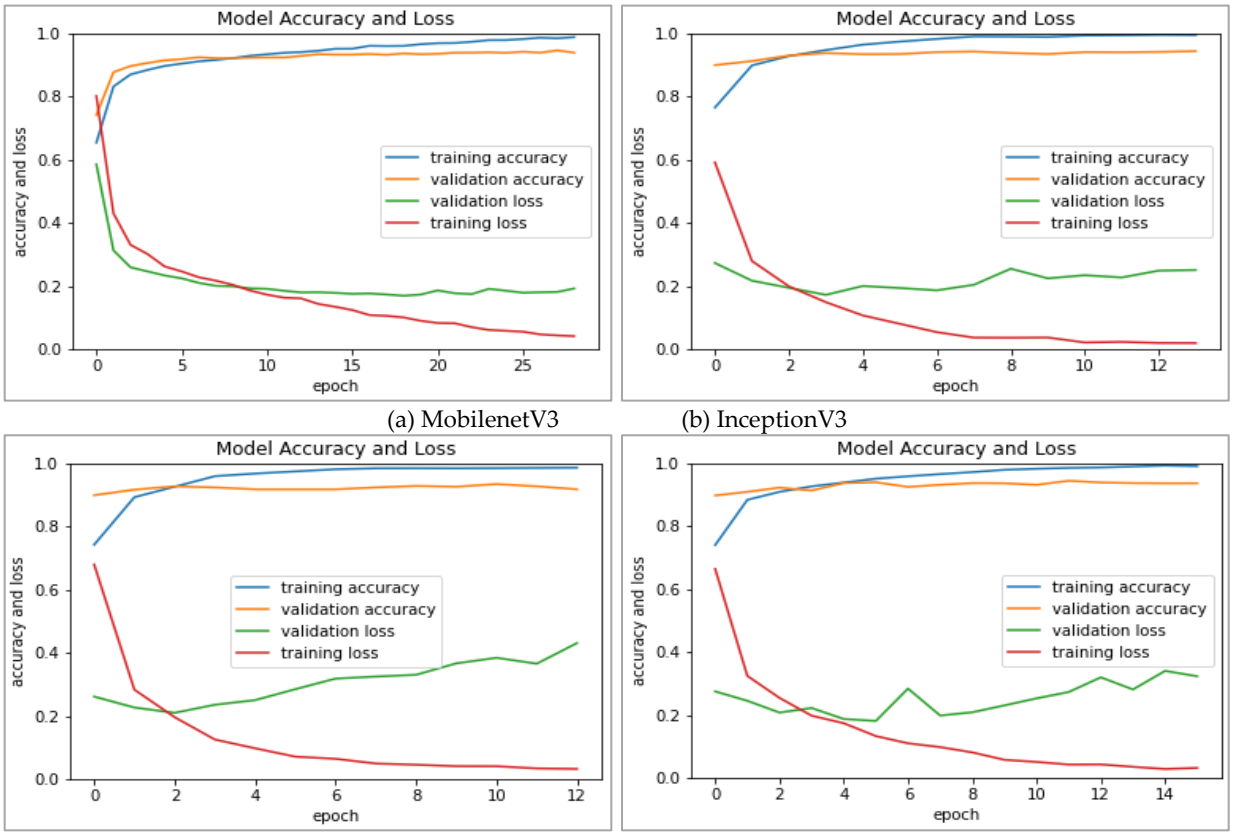
4.2 Performance Evaluation on the Augmented Dataset

Figure 4(a–d) shows the training and validation curves for MobileNetV3, InceptionV3, ResNet101, and VGG16 on the augmented dataset. All models reached high training accuracy within the first five epochs. The validation accuracy curves stayed stable with minor variation, which shows that early stop prevented long-epoch overfitting. Validation loss decreased steadily for all models, and none showed the oscillation seen in the original dataset. MobileNetV3, InceptionV3, and VGG16 reached smooth convergence. ResNet101 showed a slower decline in validation loss but kept a stable validation accuracy above 0.90 across the curve.

Figure 5(a–d) presents the confusion matrices for the four models. The main errors occurred among fruit level classes. Healthy Fruit and Infected Fruits showed the largest reduction in misclassification compared to the original dataset. Healthy Leaves reached strong separation in all models because the augmented samples increased visual diversity and reduced bias from lighting variation. Infected Leaves still produced mild confusion in VGG16 and InceptionV3, driven by variation in lesion shape and background clutter. MobileNetV3 and ResNet101 produced more balanced predictions on leaf classes due to stronger recall.

Table 5 reports the class wise precision, recall, and F1 score for all four models. InceptionV3 achieved the highest overall accuracy with an F1 score of 0.950887 for Infected Leaves and 0.943606 for Healthy Leaves. MobileNetV3 reached stable results with balanced performance across all classes and improved recall for Infected Fruits. ResNet101 delivered strong precision for Infected Fruits with

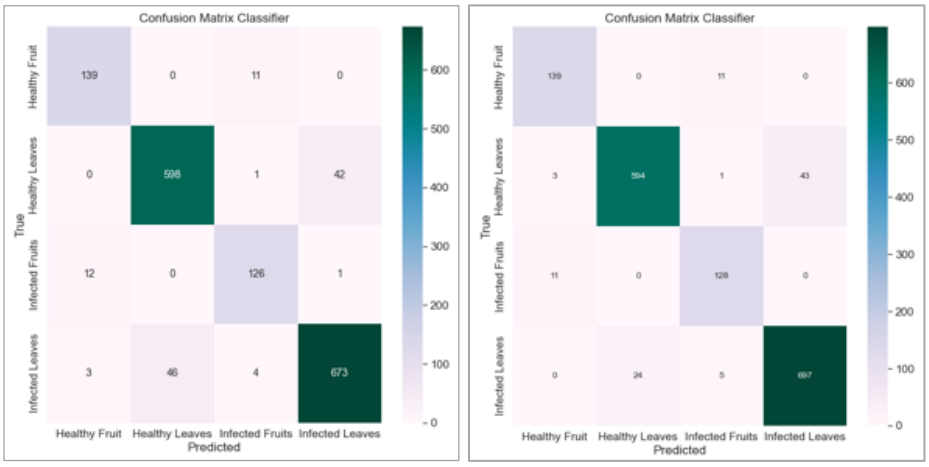
competitive recall for Healthy Leaves. VGG16 reached the highest precision for Healthy Fruit but produced lower stability in fruit level recall. The improvements across all models confirm that the augmented dataset reduced imbalance and increased robustness to background and illumination changes, which explains the tighter convergence in Figure 4 and the reduced misclassification seen in Figure 5.



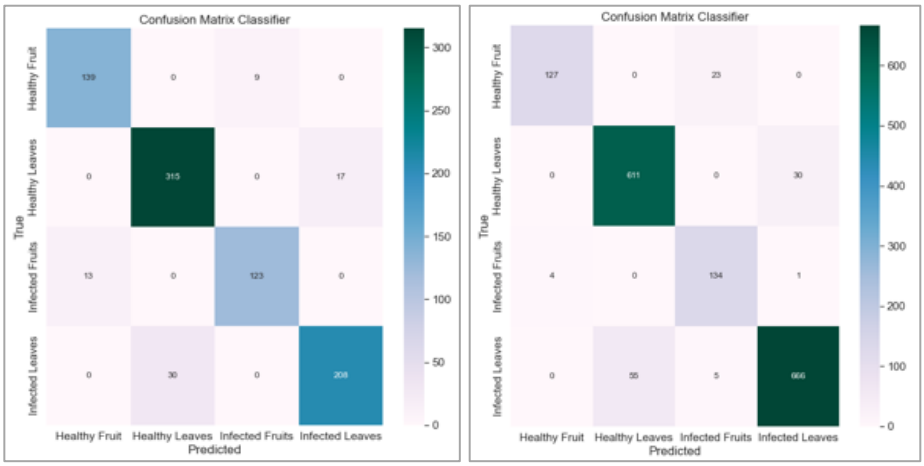
(a) MobilenetV3 (b) InceptionV3

(c) ResNet101 (d) VGG16

Figure 4. Training and validation curves for augmented dataset



(a) MobilenetV3 (b) InceptionV3



(c) ResNet101 (d) VGG16

Figure 5. Confusion matrices for augmented dataset

Table 5. Class wise metrics for augmented dataset

Model	Metric	Healthy Fruit	Healthy Leaves	Infected Fruits	Infected Leaves
MobileNetV3	Precision	0.902597	0.928571	0.887324	0.939944
	Recall	0.926667	0.932917	0.906475	0.926997
	F1Score	0.914474	0.930739	0.896797	0.933426
InceptionV3	Precision	0.908497	0.961165	0.882759	0.941892
	Recall	0.926667	0.926677	0.920863	0.960055
	F1Score	0.917492	0.943606	0.901408	0.950887
ResNet101	Precision	0.914474	0.913043	0.931818	0.924444
	Recall	0.939189	0.948795	0.904412	0.873950
	F1Score	0.926667	0.930576	0.917910	0.898488
VGG16	Precision	0.969466	0.917417	0.827160	0.955524
	Recall	0.846667	0.953198	0.964029	0.917355
	F1Score	0.903915	0.934966	0.890365	0.936051

4.3. Performance Evaluation on the Cleaned Augmented Dataset

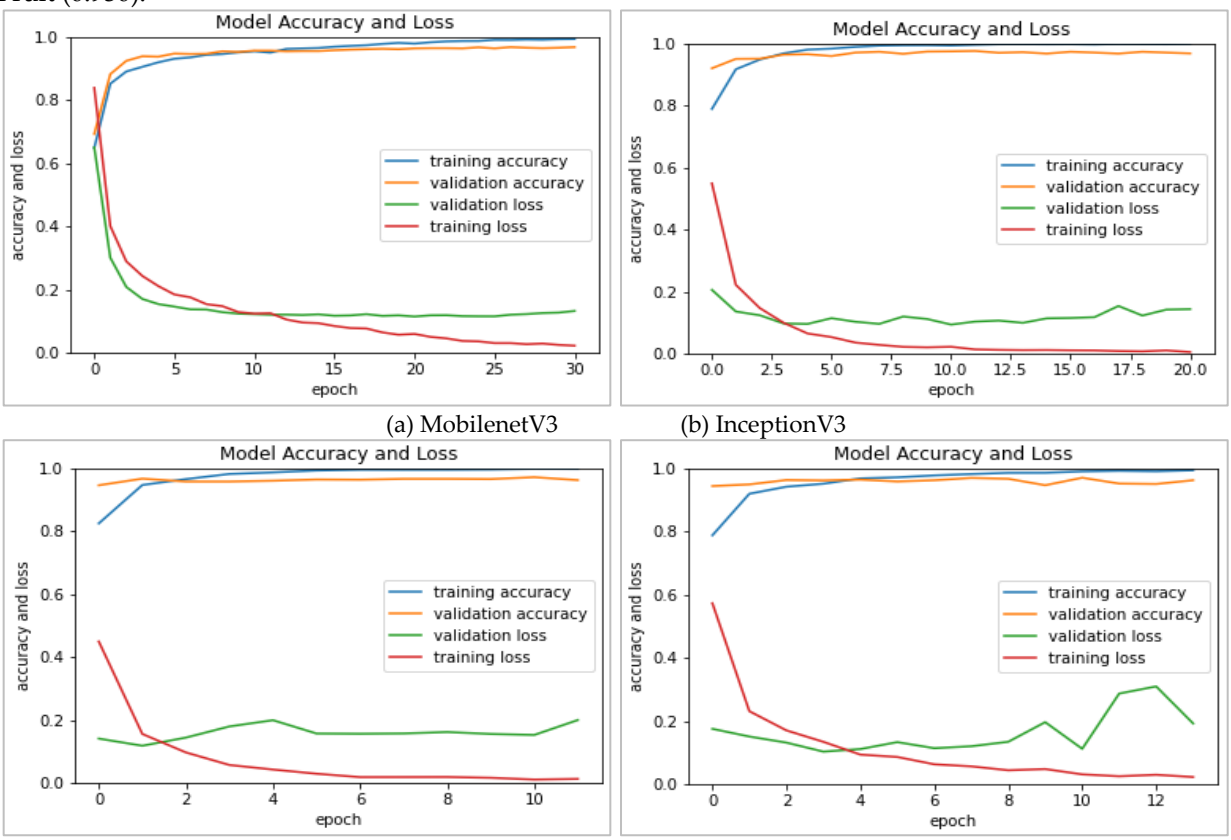
The cleaned augmented dataset represents the final and most stable version of the entire pipeline. After removing noisy, under-represented, and visually ambiguous samples, all four models showed improved convergence, reduced validation loss oscillation, and stronger generalization compared to both the original and simply-augmented datasets. Figure 6(a–d) illustrate the training and validation curves for MobileNetV3, InceptionV3, ResNet101, and VGG16 on the cleaned dataset. Across all four models, the validation accuracy saturates between 0.94 and 0.97, while the validation loss stabilizes at extremely low values without major divergence episodes, confirming that cleaning the augmented images had a measurable positive impact on stability and convergence.

The corresponding confusion matrices in Figure 7(a–d) provide additional evidence of the improved generalization. All models show very strong separation between the four classes—Healthy Fruit, Healthy Leaves, Infected Fruits, and Infected Leaves—with only minor confusion between adjacent health states (e.g., Healthy vs Infected Leaves). Notably, misclassifications dropped sharply compared with earlier datasets, particularly in Infected Fruits, which previously showed the largest confusion when noise was present in the augmented set.

From a class-wise perspective, InceptionV3 achieved the strongest overall stability and balance, with the highest macro-F1 (0.9578) and weighted-F1 (0.9665). MobileNetV3 produced the highest recall for Infected Leaves (0.9746) and

delivered the second-highest overall accuracy (0.9613) while maintaining a significantly smaller model size, making it attractive for real-time field deployment. VGG16 achieved strong precision, especially on Healthy Fruit (0.978) and Infected Leaves (0.978), but displayed slightly larger recall variance between classes. ResNet101, although deep and expressive, showed marginally lower performance due to over-regularization at later epochs, reflected in its lower recall for Infected Fruits (0.892) and Healthy Fruit (0.930).

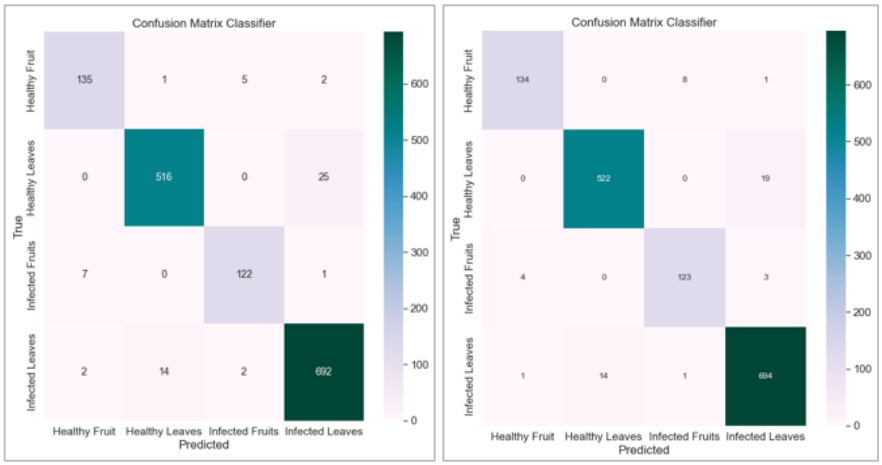
Table 6 summarizes all class-wise metrics for the cleaned dataset. The cleaned dataset enables each model to achieve above 94% overall classification accuracy, demonstrating the combined benefit of augmentation plus targeted cleaning. InceptionV3 remains the top performer, but MobileNetV3's competitive accuracy with significantly lower computational cost makes it the best practical model for agricultural deployment.



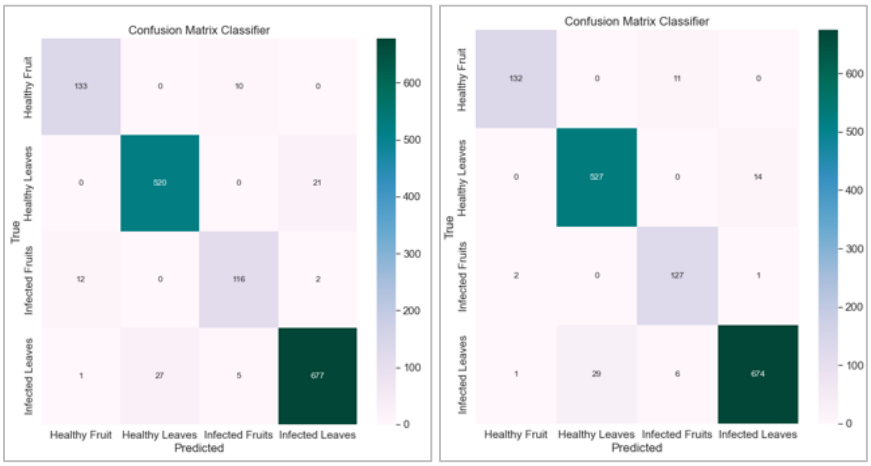
(a) MobilenetV3 (b) InceptionV3

(c) ResNet101 (d) VGG16

Figure 6. Training and validation curves for cleaned augmented dataset



(a) MobilenetV3 (b) InceptionV3



(c) ResNet101 (d) VGG16

Figure 7. Confusion matrices for cleaned augmented dataset

Table 6. Class wise metrics for cleaned augmented dataset.

Model	Metric	Healthy Fruit	Healthy Leaves	Infected Fruits	Infected Leaves
MobileNetV3	Precision	0.9375	0.97175	0.94574	0.96111
	Recall	0.94406	0.95379	0.93846	0.97465
	F1Score	0.94077	0.96269	0.94209	0.96783
InceptionV3	Precision	0.96403	0.97388	0.93182	0.96792
	Recall	0.93706	0.96488	0.94615	0.97746
	F1Score	0.95036	0.96936	0.93893	0.97267
ResNet101	Precision	0.91096	0.95064	0.88550	0.96714
	Recall	0.93007	0.96118	0.89231	0.95352
	F1Score	0.92042	0.95588	0.88889	0.96028
VGG16	Precision	0.97778	0.94784	0.88194	0.97823
	Recall	0.92308	0.97412	0.97692	0.94930
	F1Score	0.94964	0.96080	0.92701	0.96355

4.4. Comparative Discussion

A cross-dataset comparison reveals clear and consistent trends regarding model behavior, dataset quality, and the effect of augmentation plus targeted cleaning. The original dataset produced the weakest stability across all models. Training accuracy climbed rapidly, but validation accuracy plateaued early and remained notably lower. Validation loss also showed wide oscillations, especially for MobileNetV3 and VGG16, confirming that the raw dataset suffered from class imbalance, illumination variability, and insufficient sample diversity. The confusion matrices for the original dataset showed concentrated errors in two regions: (i) misclassification between Healthy Fruit and Infected Fruits, and (ii) confusion within leaf classes when shadows or strong light were present. These patterns indicate that the original distribution lacked enough representation for borderline or early-stage symptoms.

The augmented dataset brought an immediate improvement. All models achieved higher validation accuracy, and the gap between training and validation performance narrowed. Augmentation strengthened the model's ability to generalize by introducing geometric and photometric variations that were missing earlier. However, validation loss curves still showed moderate fluctuations, most prominently in ResNet101 and VGG16, which implies that some augmented samples introduced new noise instead of meaningful diversity. The confusion matrices for the augmented dataset showed reduced fruit–fruit confusion, but leaf–leaf confusion persisted, suggesting that augmentation alone could not compensate for visually

inconsistent or noisy samples originating from the raw dataset.

The cleaned augmented dataset delivered the most stable and highest-performing results across all models. Both training and validation accuracies exhibited smooth convergence, and validation loss became consistently low without the oscillations seen earlier. This indicates that removing low-quality augmented images and visually ambiguous samples allowed the models to focus on consistent patterns rather than noise. The confusion matrices confirmed substantial reductions in misclassifications across all classes, particularly for Infected Fruits, which previously showed the largest error range. The class-wise improvements were most visible in F1 scores, where all models crossed the 0.93 threshold, with InceptionV3 reaching the highest macro-F1 and MobileNetV3 demonstrating strong recall while retaining computational efficiency.

Comparing models across datasets also reveals consistent behavior. InceptionV3 remained the most balanced and stable across all settings, reflecting its robust feature extraction capability. MobileNetV3 showed the largest relative improvement when moving from original to cleaned augmented data, demonstrating its sensitivity to dataset quality and its suitability for resource-constrained environments once the data is refined. ResNet101 benefitted from augmentation but showed more fluctuation in early datasets due to its depth and higher regularization demands. VGG16 showed strong precision on cleaned data but was more prone to overfitting in the original dataset due to limited feature diversity.

Overall, the comparative analysis confirms two key insights. First, dataset quality has a stronger influence on performance than model depth or architecture. Second, augmentation alone is insufficient unless followed by targeted cleaning to remove visually inconsistent or low-information samples. The cleaned augmented dataset consistently enabled all four models to achieve high validation accuracy, stable convergence, and balanced class-wise performance, demonstrating that the combined approach produces the most reliable and generalizable classifier for fruit and leaf health assessment.

5. Conclusion and Future Work

This study evaluated four deep convolutional architectures (MobileNetV3, InceptionV3, ResNet101, and VGG16) on a multi-class fruit and leaf health dataset under three conditions: the original dataset, an augmented dataset, and a cleaned augmented dataset. The analysis shows that model performance is shaped primarily by dataset quality rather than architectural depth. On the original dataset, all models reached high training accuracy but exhibited clear gaps in validation accuracy and unstable validation loss, indicating the presence of visually inconsistent samples and class imbalance. Augmentation reduced this gap and improved generalization by introducing necessary variations; however, it also introduced noise, which limited stability in the validation curves.

The cleaned augmented dataset produced the most consistent gains. All four models achieved stable convergence, reduced loss fluctuations, and strong class-wise performance. InceptionV3 delivered the highest overall scores, while MobileNetV3 demonstrated notable improvement, making it effective for deployment on lightweight agricultural devices. Confusion matrices confirmed substantial reductions in misclassification across fruit and leaf categories, particularly for early-stage infections, which had previously posed classification challenges. These findings highlight that targeted cleaning combined with augmentation is essential to achieve reliable performance in real-world agricultural scenarios.

For future work, three directions are recommended. First, expanding the dataset to include more varieties, growth stages, and environmental conditions will further reduce model bias and improve generalization. Second, incorporating transformer-based models or diffusion-driven feature enhancement may offer additional gains, particularly for fine-grained disease discrimination. Third, integrating explainability tools such as Grad-CAM or SHAP can help agronomists interpret model decisions and improve trust when deploying the system in field environments. The results of this study demonstrate that well-curated and balanced datasets, combined with appropriately tuned architectures, can deliver robust and scalable solutions for automated crop health monitoring.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dragon fruit & leaf Dataset from Bangladesh for Classification and Ecological Research [12] .

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

References:

1. Luu, T.T.H.; Le, T.L.; Huynh, N.; Quintela-Alonso, P. Dragon Fruit: A Review of Health Benefits and Nutrients and Its Sustainable Development under Climate Changes in Vietnam. *https://cjfs.agriculturejournals.cz/doi/10.17221/139/2020-CJFS.html* 2021, *39*, 71–94, doi:10.17221/139/2020-CJFS.
2. Trivellini, A.; Lucchesini, M.; Ferrante, A.; Massa, D.; Orlando, M.; Incrocci, L.; Mensuali-Sodi, A. Pitaya, an Attractive Alternative Crop for Mediterranean Region. *Agronomy 2020, Vol. 10, Page 1065* 2020, *10*, 1065, doi:10.3390/AGRONOMY10081065.
3. Raju, C.; Pazhanivelan, S.; Perianadar, I.V.; Kaliaperumal, R.; Sathyamoorthy, N.K.; Sendhilvel, V. Climate Change as an Existential Threat to Tropical Fruit Crop Production—A Review. *Agriculture 2024, Vol. 14, Page 2018* 2024, *14*, 2018, doi:10.3390/AGRICULTURE14112018.
4. Amri, E.; Gulzar, Y.; Yeafi, A.; Jendoubi, S.; Dhawi, F.; Mir, M.S. Advancing Automatic Plant Classification System in Saudi Arabia: Introducing a Novel Dataset and Ensemble Deep Learning Approach. *Model Earth Syst Environ* 2024, *10*, 2693–2709, doi:10.1007/S40808-023-01918-9/METRICS.
5. Gulzar, Y.; Ünal, Z. Optimizing Pear Leaf Disease Detection Through PL-DenseNet. *Applied Fruit Science* 2025, *67*, 1–13, doi:10.1007/s10341-025-01265-2.
6. Gulzar, Y.; Ünal, Z. Time-Sensitive Bruise Detection in Plums Using PlmNet with Transfer Learning. *Procedia Comput Sci* 2025, *257*, 127–132, doi:10.1016/J.PROCS.2025.03.019.
7. Gulzar, Y. Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique. *Sustainability* 2023, *15*, 1906.
8. Kulkarni, V.; Kosamkar, P.; Singh, C.; Ingle, P.; Modi, V. Detection and Classification of Diseases and Maturity of Dragon Fruits. *Lecture Notes in Networks and Systems* 2022, *321*, 365–374, doi:10.1007/978-981-16-5987-4_37.
9. Saranya, T.; Deisy, C.; Sridevi, S. Efficient Agricultural Pest Classification Using Vision Transformer with Hybrid Pooled Multihead Attention. *Comput Biol Med* 2024, *177*, 108584, doi:10.1016/J.COMPBIOMED.2024.108584.
10. Gulzar, Y. Papaya Leaf Disease Classification Using Pre-Trained Deep Learning Models: A Comparative Study. *Applied Fruit Science* 2025, *67*, 1–10, doi:10.1007/S10341-025-01533-1/METRICS.
11. Gulzar, Y. PapNet: An AI-Driven Approach for Early Detection and Classification of Papaya Leaf Diseases. *Applied Fruit Science 2025 67:4* 2025, *67*, 1–11, doi:10.1007/S10341-025-01466-9.
12. Sarkar, P.; Pranta, G.K.; Mojumdar, M.U. Dragon Fruit & Leaf Dataset from Bangladesh for Classification and Ecological Research. 2024, *1*, doi:10.17632/CFCHFDPFW5.1.
13. Khan, A.; Radzi, S.A.; Zaimi, M.Z.M.; Amsan, A.N.; Mohd Saad, W.H.; Abd Razak, N.A.; Hamid, N.A.; Samad, A.S.A. Revolutionizing Agriculture with Deep Learning Current Trends and Future Directions. *International Journal of Integrated Engineering* 2024, *16*, 192–211, doi:10.30880/ijie.2024.16.03.018.
14. Zhang, W.; Zheng, C.; Wang, C.; Guo, W. DomAda-FruitDet: Domain-Adaptive Anchor-Free Fruit Detection Model for Auto Labeling. *Plant Phenomics* 2024, *6*, doi:10.34133/plantphenomics.0135.
15. Zhang, W.; Chen, K.; Zheng, C.; Liu, Y.; Guo, W. EasyDAM_V2: Efficient Data Labeling Method for Multishape, Cross-Species Fruit Detection. *Plant Phenomics* 2022, *2022*, doi:10.34133/2022/9761674.
16. Rahmania, R.; Corputty, F.; Wibowo, S.A.; Saputra, D.E.; Arrahmah, A.I. Exploration of The Impact of Kernel Size for YOLOv5-Based Object Detection on Quadcopter. *International Journal on Informatics Visualization* 2022, *6*, 726–735, doi:10.30630/joiv.6.3.898.
17. Minh Trieu, N.; Thinh, N.T. Quality Classification of Dragon Fruits Based on External Performance Using a Convolutional Neural Network. *Applied Sciences (Switzerland)* 2021, *11*, doi:10.3390/app112210558.
18. Patil, P.U.; Lande, S.B.; Nagalkar, V.J.; Nikam, S.B.; Wakchaure, G.C. Grading and Sorting Technique of Dragon Fruits Using Machine Learning Algorithms. *J Agric Food Res* 2021, *4*, doi:10.1016/j.jafr.2021.100118.
19. Yusamran, N.; Hiransakolwong, N. DIPDEEP: Classification for Thai Dragon Fruit. *Engineering and Applied Science Research* 2022, *49*, 521–530.
20. Pallavi, N.; Vijayakarthik, P.; Sushma, B. Optimizing Dragon Fruit Quality and Maturity Classification Through Deep Learning Techniques. *SN Comput Sci* 2025, *6*, doi:10.1007/s42979-025-04195-8.
21. Abhishek, A.G.S.; Ravikumar, T.; Terlapu, P.V.; Tippana, C.; Pondreti, R. Intelligent Fruit Detection System Using Optimized Hybrid Deep Learning Models. *Journal of Machine and Computing* 2025, *5*, 1386–1395, doi:10.53759/7669/jmc202505109.
22. Nikam, S.B.; Lande, S.B.; Nagalkar, V.J.; Wakchaure, G.C.; Kumar, P.S. Predictive Classification Model for Quality Grading and Maturity Detection of Dragon Fruit Using Fused Deep CNN Feature and Ensemble Learning. *J Food Process Preserv* 2025, *2025*, doi:10.1155/jfpp/6938071.
23. da Silva-Ferreira, M.V.; Barbon Junior, S.; Turrisi da Costa, V.G.; Barbin, D.F.; Lucena-Barbosa, J.E. De Deep Computer Vision System and Explainable Artificial Intelligence Applied for Classification of Dragon Fruit (Hylocereus Spp.). *Sci Hortic* 2024, *338*, doi:10.1016/j.scienta.2024.113605.
24. Vo, H.T.; Thien, N.N.; Mui, K.C. A Deep Transfer Learning Approach for Accurate Dragon Fruit Ripeness Classification and Visual Explanation Using Grad-CAM. *International Journal of Advanced Computer Science and Applications* 2023, *14*, 1344–1352, doi:10.14569/IJACSA.2023.01411137.
25. Cometa, L.M.A.; Garcia, R.K.T.; Latina, M.A.E. Real-Time Visual Identification System to Assess Maturity, Size, and Defects in Dragon Fruits †. *Engineering Proceedings* 2025, *92*, doi:10.3390/engproc2025092039.
26. Khatun, T.; Nirob, M.A.S.; Bishshash, P.; Akter, M.; Uddin, M.S. A Comprehensive Dragon Fruit Image Dataset for Detecting the Maturity and Quality Grading of Dragon Fruit. *Data Brief* 2024, *52*, doi:10.1016/j.dib.2023.109936.
27. Li, X.; Wang, X.; Ong, P.; Yi, Z.; Ding, L.; Han, C. Fast Recognition and Counting Method of Dragon Fruit Flowers and Fruits Based on Video Stream. *Sensors* 2023, *23*, doi:10.3390/s23208444.
28. Ha, D.M.; Hung, T.; Kieu, N.X.; Vuong, N.G.; Thuy, Q.D.T. Semantic Connection-Based Learning for Dragon Fruit Disease Classification. *Journal of Information Hiding and Multimedia Signal Processing* 2024, *15*, 281–291.
29. Nguyen, T.P.T.; Nguyen, T.T.; Nguyen, H.Q.; Nguyen, T.D.; Nguyen, C.K.; Cu, N.G. An Enhanced Image Classification Model Based on Graph Classification and Superpixel-Derived CNN Features for Agricultural Datasets. *Computers, Materials and Continua* 2025, *85*, 4899–4920, doi:10.32604/cmc.2025.067707.
30. Wang, J.; Gao, K.; Jiang, H.; Zhou, H. Method for Detecting Dragon Fruit Based on Improved Lightweight Convolutional Neural Network; 基于改进的轻量化卷积神经网络火龙果检测方法. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of*

Agricultural Engineering 2020, *36*, 218–225, doi:10.11975/j.issn.1002-6819.2020.20.026.

31. Shang, F.; Zhou, X.; Liang, Y.; Xiao, M.; Chen, Q.; Luo, C. Detection Method for Dragon Fruit in Natural Environment Based on Improved YOLOX; 基于改进 YOLOX 的自然环境中火龙果检测方法. *Smart Agriculture* 2022, *4*, 120–131, doi:10.12133/j.smartag.SA202207001.
32. Wang, J.; Zhou, J.; Zhang, Y.; Hu, H. Multi-Pose Dragon Fruit Detection System for Picking Robots Based on the Optimal YOLOv7 Model; 基于优选 YOLOv7 模型的采摘机器人多姿态火龙果检测系统. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering* 2023, *39*, 276–283, doi:10.11975/j.issn.1002-6819.202208031.
33. Zhou, J.; Zhang, Y.; Wang, J. A Dragon Fruit Picking Detection Method Based on YOLOv7 and PSP-Ellipse. *Sensors* 2023, *23*, doi:10.3390/s23083803.
34. Zhu, L.; Deng, W.; Lai, Y.; Guo, X.; Zhang, S. Research on Improved Road Visual Navigation Recognition Method Based on DeepLabV3+ in Pitaya Orchard. *Agronomy* 2024, *14*, doi:10.3390/agronomy14061119.
35. Zhou, Z.; Peng, R.; Li, R.; Li, Y.; Huang, D.; Zhu, M. Remote Sensing Identification and Rapid Yield Estimation of Pitaya Plants in Different Karst Mountainous Complex Habitats. *Agriculture (Switzerland)* 2023, *13*, doi:10.3390/agriculture13091742.
36. Li, Q.; Yan, L.; Huang, D.; Zhou, Z.; Zhang, Y.; Xiao, D. Construction of a Small Sample Dataset and Identification of Pitaya Trees (Selenicereus) Based on UAV Image on Close-Range Acquisition. *J Appl Remote Sens* 2022, *16*, doi:10.1117/1.JRS.16.024502.
37. Yu, J.; Sun, Y.; Latinovic, N.; Kong, C.; Han, B.; Zhang, X. Nondestructive Internal Quality Detection Method for Yellow Pitaya Based on EIS and Tactile Multimodal Perception Data-Driven Approach. *Journal of Food Composition and Analysis* 2025, *144*, doi:10.1016/j.jfca.2025.107744.
38. Pan, Y.; Wang, Y.; Zhou, Y.; Zhou, J.; Chen, M.; Liu, D.; Li, F.; Liu, C.; Zeng, M.; Jiang, D.; et al. A Smartphone-Based Non-Destructive Multimodal Deep Learning Approach Using PH-Sensitive Pitaya Peel Films for Real-Time Fish Freshness Detection. *Foods* 2025, *14*, doi:10.3390/foods14101805.
39. Xu, T.; Song, L.; Lu, X.; Zhang, H. Dual-Index Detection Method of Pitaya Quality and Maturity Based on YOLO v7-RA; 基于 YOLO v7 RA 的火龙果品质与成熟度双指标检测方法. *Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery* 2024, *55*, 405–414, doi:10.6041/j.issn.1000-1298.2024.07.040.