Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data Leakage, Acquisition Bias, and Cross-Dataset Generalization in Guava Disease Image Classification

Abdul Ahad

ahad.abdul01@northsouth.edu

North South University https://orcid.org/0009-0005-1253-5314

Mirza Farhan Shahriar

North South University

Md. Naimur Rahman

North South University

Nourash Azmine Chowdhury

North South University

Research Article

Keywords: Data leakage, Near-duplicate detection, Group-disjoint evaluation, Domain shift, Computer vision, Plant disease recognition

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11860184/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Purpose

We examined how sharing images from the same photograph or acquisition session across training and test sets affects reported accuracy in guava disease recognition.

Methods

We audited four public datasets. For the Guava Fruit Disease Dataset (GFDD24), we inferred groups of related images and checked common photographic origin using filenames where available. A paired experiment held the evaluation images and training-set size fixed while changing whether training included images from the held-out groups. We also evaluated session holdouts and transfer between datasets with shared labels.

Results

All 382 official GFDD24 test images belonged to inferred groups represented in training. Filename records covered 139 of those test images and confirmed a training image from the same augmentation family for 132. On the same 354 evaluation images, replacing 219 training images with images from held-out groups increased mean MobileNetV2 accuracy from 94.54% to 97.83% across three seeds (paired gain 3.30 percentage points, 95% group-bootstrap interval 0.68–7.36). On a separate three-class leaf task, central-region balanced accuracy fell from 84.8% under grouped cross-validation to 45.9% on held-out sessions. Cross-dataset accuracy with frozen features ranged from 72.4% to 86.8%.

Conclusion

Sharing related images can inflate apparent accuracy. Reporting the source photograph, acquisition session, and split assignment for each image would make benchmarks easier to audit and their results easier to interpret.

Introduction

Guava (Psidium guajava L.) is an economically important fruit crop in South Asia. Anthracnose, caused by Colletotrichum species, and infestation by tephritid fruit flies are among the main causes of unmarketable fruit, and early recognition at the orchard or packing stage limits losses. Convolutional neural networks (CNNs) have become the standard tool for image-based recognition of plant disease [1, 2], but they need labelled images, and public image data for guava fruit remain scarce.

The Guava Fruit Disease Dataset (GFDD24) [3] partly fills this gap. It contains photographs of guava fruit from orchards in Rajshahi and Pabna, Bangladesh, labelled as Anthracnose, Fruit fly, or Healthy and verified by a plant pathologist. Its creators preprocessed the 473 photographs (contrast-limited adaptive histogram equalization and unsharp masking), expanded them to 3,784 images by augmentation, and

released the result already divided into training, validation, and test folders. Studies on GFDD24 report accuracies from 95.6% to about 99.99% [4, 5, 6, 7].

These figures are hard to interpret because of how the dataset was built. Each of the 473 photographs appears as eight augmented images, and the augmentation happened before the split. A test image may therefore have one or more versions of the same photograph in the training set, and a classifier can score well by recognizing the photograph rather than the disease. Data leakage of this kind is a recognized source of inflated results across machine-learning-based science [8], and augmenting before splitting has been shown to inflate plant-disease detection metrics severely [9]. Plant-disease benchmarks are also prone to shortcuts unrelated to the symptoms: a model trained on a few background pixels of PlantVillage images classified far above chance [10], and models trained on PlantVillage lost most of their accuracy on images from other sources [1]. Published GFDD24 studies either merged the folders and re-split them at the image level [5] or augmented the data further before a random image-level split [6]. Neither procedure keeps the versions of a photograph together.

Two further issues are rarely examined. Many public plant-disease datasets are collected in a few sessions, so that a class can coincide with a site, a date, or the weather, and a model can learn the setting instead of the symptoms. Explanations, in turn, are usually shown as a few qualitative heatmaps [6, 7], although saliency maps can look plausible while being insensitive to the model [11], and their faithfulness can be tested directly [12].

This paper makes the following contributions:

1. A leakage audit of GFDD24. We use candidate retrieval and geometric verification to identify groups of related images and quantify their overlap across the official partitions.
2. A group-disjoint benchmark. We compare four backbones across three training seeds on one fixed split, report image- and group-level accuracy, and bootstrap test groups to reflect dependence among related images. A paired experiment evaluates how replacing training images with members of held-out groups changes accuracy on identical test images.
3. A controlled experiment using published original images. On a second dataset that releases its original photographs [13], we trace augmented images to their sources by keypoint matching, identify repeated views in its “original” folder, and measure accuracy on identical held-out photographs with and without their augmented copies in training, with frozen and fine-tuned networks.
4. **Acquisition confounding and transfer across datasets.** Using two further public datasets, we show that collection session predicts the class in one of them and that holding sessions out removes much of the apparent accuracy, and we measure how accuracy changes when a model is tested on another dataset with the same classes.
5. A five-step checklist translates the audit findings into decisions about benchmark design and reporting.

2 Related Work

2.1 Guava disease recognition

Early guava studies used private datasets. Mostafa et al. [14] compared five CNNs on augmented guava images, with ResNet-101 reaching 97.74%, and Almadhor et al. [15] classified high-resolution DSLR images of guava disease with hand-crafted features and classical classifiers. Since GFDD24 became available, several studies have used it (Table 1). Kilci and Koklu [4] combined SqueezeNet features with boosting classifiers (95.6% with gradient boosting); their journal extension [16] replaced the boosting classifiers with SVM and ANN on InceptionV3 features and reported 99.83% and 99.58% under 10-fold cross-validation over all 3,784 images, a protocol that permits augmented siblings to occur in different folds. Rijon and Arafath [5] cascaded EfficientNet-B0 features with boosting ensembles and reported about 99.99%. Ahmed et al. [6] fine-tuned ResNet-101 on an expanded set of 4,632 images split randomly 80/10/10, reached 98.48%, and showed Grad-CAM maps. Ghosh et al. [7] reported 98.15% with InceptionV3 and 94.46% with ResNet50, using CutMix and MixUp augmentation and SHAP and LIME for interpretation; the split is described only as training, validation, and testing sets with 20% held out for validation, so the unit of partitioning is unclear. Shahade et al. [17] are the closest prior work to ours: they split the 473 original photographs 70/20/10 before augmenting, expanded only the 331 training photographs to 2,647 images, and state that no augmented copy of a validation or test image was used in training. Their reported accuracy is nevertheless difficult to interpret, because the text describes a 47-image original test set while their results tables, confusion matrix, and error analysis all report 382 test images; 382 is exactly the size of GFDD24's official augmented test folder, every image of which belonged to an inferred group represented in official training (Section 4.1). We therefore treat [17] as the first study to describe a source-original split for GFDD24, and as an illustration of how difficult such a split is to carry out and report reliably when a dataset is distributed only in augmented form. In a broader study of guava leaf and fruit diseases, Biswas et al. [18] used GAN-generated samples and five-fold cross-validation and reported 98.20% for a ResNet50–DenseNet121 fusion. Other guava image datasets from Bangladesh provide original, non-augmented photographs of fruit and leaves [19, 20], which makes them usable as independent test sets. Of the GFDD24 studies we found, only Shahade et al. [17] report a split that keeps the augmented versions of each photograph together, and their reported test-set size is internally inconsistent (Table 1); the remaining studies used image-level splits or provided insufficient detail to establish the partitioning unit.

Table 1
Published results on GFDD24

Study	Model	Split as described	Reported accuracy	Explanations
Kilci and Koklu [4, 16]	SqueezeNet or InceptionV3 features + boosting, SVM, ANN	not stated in [4]; [16] uses 10-fold cross-validation over all 3,784 images, by image	95.6% [4]; 99.83% [16]	—
Rijon and Arafath [5]	EfficientNet-B0 features + boosting ensembles	train and validation folders merged and re-split 80/20 by image; official test folder	≈ 99.99%	—
Ahmed et al. [6]	ResNet-101	further augmented to 4,632 images, then split 80/10/10 by image	98.48%	Grad-CAM
Ghosh et al. [7]	InceptionV3, ResNet50	not stated beyond “split into training, validation and testing sets”; 20% validation, split unit unclear	98.15%	SHAP, LIME
Shahade et al. [17]	EfficientNetB3, ResNet50, DenseNet121, ViT-B16	473 originals split 70/20/10, then augmentation of the training split only; results tables report 382 test images	98.5%	—
This work	ResNet50, EfficientNetB0, MobileNetV2, ConvNeXt-Tiny	by inferred group, 70/15/15	96.10 ± 0.44% (EfficientNetB0)	Grad-CAM, LIME, deletion test

2.2 Data leakage and shortcut learning

Kapoor and Narayanan [8] catalogued leakage in machine-learning-based science and identified non-independence between training and test data as a recurring cause. For plant imagery, Gerz and Jelali [9] showed with reproducible case studies that applying augmentation before the train–test split inflated detection metrics by tens of percentage points. Noyan [10] showed that PlantVillage labels can be predicted from background pixels alone. Near-duplicate images are commonly found with perceptual hashes [21], which tolerate resizing and mild photometric changes but not rotation or mirroring.

2.3 Explanations and their evaluation

Grad-CAM [22] weights convolutional features by the gradient of a class score. LIME [23] fits a local surrogate on perturbed superpixels [24]. Fahmid et al. [25] used probabilistic class activation mapping for weakly supervised localization in chest radiographs. Here the question is whether the regions highlighted on a guava image affect its predicted score. Motivated by saliency sanity checks [11] and deletion tests [12], we compare the probability drop under image-specific and control masks.

Materials and Methods

3.1 Dataset

GFDD24 [3] contains 512×512 RGB PNG images in three classes (Table 2). Every class count is a multiple of eight, consistent with eight augmented versions of each of the 473 photographs. The class labelled Fruit fly denotes fruit-fly infestation damage rather than a pathogen; we keep the dataset's labels. The dataset was distributed with training, validation, and test folders containing 2647, 755, and 382 images, respectively, as counted in the files. The repository description lists 775 validation images, whereas the counted 755 reconciles the published total of 3,784. The illustrated augmented versions in Online Resource 1 differ by rotation through arbitrary angles, with the corners filled by stretched edge pixels, together with zoom, small shifts, shears, and mirroring; the fruit itself is unchanged.

AI assistance. Claude (Anthropic) assisted with analysis code and manuscript drafting. ChatGPT (OpenAI) assisted with revision, figure preparation, and the filename analysis and thumbnail-based visual comparisons in Online Resource 1. Those visual judgments are identified separately from algorithmic matches.

Table 2
GFDD24 composition

Class	Images	Source photographs (images ÷ 8)
Anthracnose	1,544	193
Fruit fly	1,312	164
Healthy	928	116
Total	3,784	473

3.2 Inferring source-photograph groups

The augmentations used to build GFDD24 include rotations by arbitrary angles, with the corners filled by stretched edge pixels, as well as zoom, shifts, shears, and flips (Online Resource 1). We therefore grouped images by keypoint matching with geometric verification, the approach also used in case study 2 (Section 3.6).

Candidates. For each image, candidate partners of the same class were its 30 nearest neighbours by cosine similarity of ImageNet-pretrained [26] EfficientNetB0 [27] embeddings (1,280-dimensional pooled features at 224×224, averaged over four right-angle rotations, each with and without mirroring), together with every image whose 64-bit perceptual hash [21], computed at 128×128 for each of these eight versions, lay within $\tau = 10$ bits.

Verification. Each candidate pair was compared using SIFT keypoints [28] (at most 500, detected in the central disc of the image at 320 pixels so that the filled corners are ignored, and also on the mirrored

image to cover flips), Lowe's ratio test (0.75), and an affine transform fitted with RANSAC [29] (reprojection threshold 5 pixels); the score is the number of inlier matches. We treated 1,500 random cross-class pairs as presumed nonmatches. Their maximum score was nine inliers, below the prespecified floor of 15. Among thresholds of 15 through 74 inliers, the value whose number of groups came closest to the known 473 photographs was kept, with ties resolved in favour of the larger threshold; the selected threshold, 74 inliers, was the upper end of the tested range. The target count guided threshold selection. Agreement with that count cannot independently validate source identity or exclude fragmented families. Verified pairs were linked, and linked images formed a group. A completion pass then compared images of incomplete groups (fewer than eight images) whose sizes add up to at most eight, as a heuristic, and added the pairs that passed verification.

We compared the recovered group counts with the expected counts for Anthracnose (193), Fruit fly (164), and Healthy (116). We examined the group-size distribution, visually inspected six randomly selected groups of eight images each, and audited the split (Section 3.3). For reference we also report a conventional perceptual hash without rotation invariance (Hamming-distance threshold of eight bits), the rotation/flip-invariant hash alone, and agglomerative clustering (average linkage, cosine distance) of the embeddings into the known number of photographs per class. Separately, we parsed the class and the filename prefix before "_unsharp_clahe_augmented_" as a reference family, retaining the final digit as the augmentation index. The visual matching algorithm used image content. We compared its direct links with filename-family membership on 144 candidate pairs sampled equally across eight inlier-score bands, and enumerated the families in every component larger than eight. Each of the 144 pairs also received an AI-assisted visual judgment in randomized order with identifiers and inlier scores concealed. Online Resource 1 provides the analysis notebook excerpt, validation script, matched-experiment results, and annotated image-group audit.

3.3 Data splits

Group-disjoint split. Within each class, groups were shuffled with seed 42 and assigned one at a time, largest first, to whichever partition was furthest below its target as a fraction of its target number of images (70% training, 15% validation, 15% test), so that every image of a group shares a partition. Each inferred group occurred in exactly one partition. Table 3 gives image and group counts for both splits. The audit compares, for the official and new splits, the cosine similarity of each test image to its nearest training image in the invariant embedding space and the number of test images whose invariant hash distance to some training image is at most τ.

Official split. The folders as distributed were used unchanged for the leakage baseline.

Matched contamination experiment. Within each of the 55 test groups, images were shuffled with a fixed random seed. For a group of size n, the first $\min(4, \max(0, n - 2))$ images were reserved for transfer into training. This rule selected 219 of the 573 test images and left 354 for evaluation. The clean arm retained all 2,636 training images. The contaminated arm retained 2,417 randomly selected clean training images and added the 219 reserved images. Both arms used the same validation set, 354

evaluation images, MobileNetV2 training schedule, and three seeds. Class weights were computed from each arm's labels. The paired interval resampled the 55 inferred evaluation groups with replacement 2,000 times and averaged the accuracy difference across seeds within each resample.

Table 3
Partition sizes, expressed as images / inferred groups. An inferred group may occur in multiple official partitions

Split	Training	Validation	Test
Group-disjoint	2636 / 263	575 / 54	573 / 55
Official (leaky baseline)	2647 / 371	755 / 322	382 / 231

3.4 Classifiers and training

We fine-tuned four ImageNet-pretrained backbones of different size and design: ResNet50 [30], EfficientNetB0 [27], MobileNetV2 [31], and ConvNeXt-Tiny [32], a recent convolutional network that adopts design elements of vision transformers. Each backbone received its own Keras preprocessing function inside the model, so all models take raw RGB inputs. The classification layer was replaced by global average pooling, dropout (0.4), a 256-unit ReLU layer with L2 regularization (10 − 4), dropout (0.4), and a three-way softmax. Training-time augmentation consisted of horizontal flips, rotations of up to ± 0.08 of a full turn (about ± 29°), zoom and translation of up to 15% and 10%, and contrast changes of up to 10%.

Training had two phases with Adam [33] and cross-entropy with label smoothing of 0.1 [34]. In phase 1 (up to 15 epochs, learning rate 10 − 4) only the new layers were trained. In phase 2 (up to 15 epochs, learning rate 10 − 5) the last 30 backbone layers were unfrozen except batch-normalization layers, which stayed in inference mode. Each phase used early stopping (patience 6) and learning-rate halving (patience 3) on validation loss and kept its best-validation-loss weights; the phase-2 weights were kept only if they improved on phase 1. Class weights inversely proportional to class frequency compensated for the class imbalance. Each backbone was trained with seeds 42, 123, and 2024. The test set was not used for any choice. Table 4 lists the settings.

Table 4
Training configuration

Setting	Value
Input size, batch size	224×224 RGB, 32
Head	GAP → Dropout 0.4 → Dense 256 ReLU (L2 10 − 4) → Dropout 0.4 → Softmax 3
Phase 1	backbone frozen, ≤ 15 epochs, Adam, learning rate 10 − 4
Phase 2	last 30 layers unfrozen (batch normalization frozen), ≤ 15 epochs, learning rate 10 − 5
Loss, imbalance	cross-entropy with label smoothing 0.1; inverse-frequency class weights
Regularization	early stopping (patience 6), learning-rate halving (patience 3), best-validation checkpoint
Seeds	42, 123, 2024
Hardware, software	Tesla T4 (Kaggle), TensorFlow 2.20.0

3.5 Evaluation

On the test set we report accuracy and macro-averaged precision, recall, and F1, as mean ± standard deviation over seeds. Because the eight images of a photograph are not independent, we also report group-level accuracy. We average predicted class probabilities across test images in each inferred group and assign the group the most probable class. A 95% confidence interval for the seed-averaged accuracy was obtained with a bootstrap [35] that resamples inferred test groups, not images (2,000 resamples). Backbones were compared with an exact McNemar test [36] on discordant groups, separately for each seed. The between-protocol contrast is the difference between the mean accuracy of the same configuration on the official and the inferred-group-disjoint test sets. We report the parameter count and mean latency for one 224×224 input on the x86_64 CPU, measured over 30 forward passes after five warm-up passes.

3.6 Case study 2: a dataset with published original photographs

GFDD24 does not release its 473 source photographs, so the groups of Section 3.2 are a reconstruction. To measure leakage against a reference, we used the guava leaf-and-fruit dataset of Rajbongshi et al. [19], as released in its data repository [13], which contains 527 “original” images and 4,899 augmented images in five classes: Disease-free leaves (126 originals / 876 augmented), Phytophthora on fruit (114 / 942), Red rust on leaves (87 / 1,154), Scab on fruit (106 / 864), and Styler-end rot on fruit (94 / 1,063). The accompanying article describes a larger version (681 originals and 8,525 augmented images); we used the files as released.

Tracing augmented images. The augmentations in this dataset include crops with strong zoom, shifts, and rotations by arbitrary angles that leave black corners (examples in Section 4.4). The invariant hash of Section 3.2 linked only 645 of the 4,899 augmented images (13.2%) to an original, and, apart from exact copies, an augmented image was no more similar to its nearest original in embedding space than two different originals were to each other (median cosine similarity 0.879 versus 0.908). We therefore used geometric verification. For each augmented image, the 10 originals of the same class with the most similar embeddings, plus any original within 10 hash bits, were compared using SIFT keypoints [28] (at most 500 per image, computed at 320 pixels on the long side and also on the mirrored image to cover flips), Lowe's ratio test (0.75), and a similarity transform fitted with RANSAC [29] (reprojection threshold 5 pixels). The score is the number of inlier matches. The sampled cross-class pairs, treated as presumed nonmatches, never exceeded 24 inliers (700 pairs; 99th percentile 12), so an augmented image was assigned to its best-scoring original when that original reached at least 25 inliers. The same test, applied to each original and its five most similar originals of the same class, linked originals that show the same leaf or fruit; linked originals were merged into specimen groups.

Protocols. Each image was represented by the pooled features of an ImageNet-pretrained MobileNetV2 [31] at 224×224 without fine-tuning and classified by L2-regularized logistic regression [37], with the regularization strength chosen on a validation partition from {10 − 3, 10 − 2, 10 − 1, 1}. A 1-nearest-neighbour classifier on rotation/flip-averaged embeddings was run alongside as a direct probe of memorization. Three protocols were repeated over 10 random splits. (A) The common practice: the 4,899 augmented images were split 70/15/15 at random by image, and the test set consisted of augmented images. (B) Specimen groups were split 70/15/15 within each class, each partition received the traced augmented images of its groups, and the model was tested both on augmented images and on the original images of the test groups. (C) The test set was the original images of the test groups from (B), but training used every augmented image except those of validation groups, so augmented copies of the test photographs were present in training; as a control for training-set size, a random subset of that set with the same size as the clean training set of (B) was also used. Untraced augmented images (33.5%) were therefore used only in (A) and (C). For each test set we report the share of test images within 10 bits of a training image under the invariant hash. Paired differences on identical test sets were tested with the Wilcoxon signed-rank test [38]. Frozen features isolate the effect of the split from training choices. To check that the conclusions hold for fine-tuned networks, the key comparisons were repeated with MobileNetV2 fine-tuned with the schedule used on GFDD24 (Section 3.4): first the new classification head, then the last 30 backbone layers (inverted-residual blocks 14–16 and the final 1×1 convolution) with batch normalization kept frozen. Because the earlier layers are frozen in both phases, their activations were computed once and cached, which made fine-tuning feasible without a GPU; horizontal flipping was the only training-time augmentation. In case study 2, protocols (B) and (C) were repeated with fine-tuning on the first of the ten splits; protocol (A) was evaluated with frozen features only.

3.7 Case study 3: transfer across datasets and acquisition confounding

Even a group-disjoint evaluation can share cameras, sites, and collection practices between training and test sets. We therefore tested transfer between datasets that share classes, using the frozen MobileNetV2 features and logistic regression of case study 2. Two further public datasets were used: the BD Guava Fruit and Leaf Disease Dataset, version 5 [39], with 1,092 field images in ten classes collected in Bangladesh during the monsoon season and per-image metadata (variety, plant part, collection date, location, temperature, humidity, and weather); and a six-class guava leaf dataset [40] with 3,355 images collected in the Savar and Jatrabari areas of Bangladesh (Healthy, Anthracnose, Sooty mould, Senescence, Insect infestation, and Dried). Three comparisons used shared classes: healthy leaf, red rust, and scab between the case-study-2 dataset and BD Guava; the same pair restricted to healthy leaf and red rust, which removes the leaf-versus-fruit cue that identifies scab; and healthy leaf versus insect damage between BD Guava and the six-class dataset. In-dataset accuracy was estimated by five-fold cross-validation that kept near-duplicates together: specimen groups for the case-study-2 dataset, invariant-hash groups for BD Guava, and, for the six-class dataset, whose single leaves on plain backgrounds made the hash collide across different leaves, groups of images with embedding cosine similarity of at least 0.97. The regularization strength was chosen by inner three-fold group cross-validation. Cross-dataset accuracy was obtained by training on all images of the shared classes in one dataset and testing on all of them in the other. To examine acquisition confounding, we cross-tabulated BD Guava classes against collection sessions (date and location), predicted the class from metadata alone with a decision tree under five-fold cross-validation, and, within classes collected in more than one session, predicted the session from image features. Confidence intervals (95%) for the frozen-feature results were obtained by resampling near-duplicate groups of the test dataset (2,000 bootstrap samples), and every comparison was repeated with the fine-tuned MobileNetV2 described in Section 3.6. To locate the signal responsible for session confounding, every BD Guava image was additionally encoded twice more: once with the border ring blanked out, leaving the central disc of radius 0.45 of the image width (a proxy for the plant, which the photographs centre), and once with that disc blanked out, leaving only the border (a proxy for the background). The same classifier was then trained on each of the three inputs. To measure the effect of session confounding directly, the three BD Guava leaf classes that were each collected in two sessions (algal leaf spot, red rust, and healthy leaf) were classified with sessions held out: the model was trained on the main session of each class and tested on the other, and vice versa, and the result was compared with random group cross-validation on the same images.

3.8 Explanations

Explanations were computed for the backbone with the highest mean validation accuracy, using its seed with the best validation accuracy. Grad-CAM [22] used the gradient of the predicted class logit with respect to the last activation layer of the backbone (conv5_block3_out, top_activation, or out_relu; for ConvNeXt-Tiny, the output of its final stage). LIME [23] segmented each image into 50 SLIC superpixels [24], drew 1,000 perturbation samples, and displayed the five superpixels with the largest positive weights. We show two correctly classified test images per class and up to three misclassified test images.

For the deletion test, 60 correctly classified test images were drawn at random. For each image and fraction k ∈ {5%, 10%, 20%}, the k fraction of pixels with the highest Grad-CAM values was replaced by a Gaussian-blurred copy of the image (σ equal to 5% of the image width), and the drop in the predicted-class probability was recorded. As a control, the same procedure was applied with the Grad-CAM map of another, randomly chosen image from the sample, which applies the same nominal masking fraction using a different image's saliency pattern. A faithful explanation should produce larger drops than the control.

Results

4.1 Image grouping and leakage in the official split

The conventional perceptual hash yielded 3,699 groups for 3,784 images, and the invariant hash yielded 2,407. Embedding clustering placed up to 90 Fruit fly images in a single group (Table 5). Geometric verification yielded 372 inferred groups, compared with 473 reported source photographs; some groups therefore appear to merge multiple source photographs. Of these groups, 81.7% contained eight images. Groups larger than eight contained 1,336 images (35.3% of all images), and the largest contained 112. Six Healthy groups contained fewer than eight images. Table 6 gives the group-size distribution by class, and Fig. 1 shows the pooled distribution. Against filename-family membership, the 144 sampled candidate pairs included 32 true positives, 22 false positives, six false negatives and 84 true negatives at 74 inliers. Precision was 59.3% (32/54) and recall was 84.2% (32/38) on this score-stratified sample. Equal allocation by score band changes the prevalence of positives. These estimates describe the sampled candidates and exclude pairs missed by candidate retrieval.

The 62 oversized components contained 167 complete filename families. Each had indices 0–7 and occurred in one component. Across those components, 4,676 of 20,420 within-component pairs (22.9%) shared a filename family. Of the 1,274 consecutive pairs in the exported sheets, 950 shared a family. Component 190 contained 112 Fruit fly images spanning 14 timestamped stems, each with all eight indices. The timestamps ran from 15:42:30 to 15:43:36. The images show similar cut surfaces viewed at different angles. The component merges multiple named captures, and the number of physical fruits remains unknown (Online Resource 1).

Table 5
Groups found per class by each procedure (expected = images ÷ 8). For embedding clustering, the size of its largest cluster is given

Class	Expected	Conventional pHash (τ = 8)	Invariant pHash (τ = 10)	Embedding clustering: largest group	Keypoint-verified (final)	Groups with exactly 8 images
Anthracnose	193	1,473	608	37	164	84.1%
Fruit fly	164	1,310	1,078	90	91	63.7%
Healthy	116	916	721	56	117	92.3%
Total	473	3,699	2,407	—	372	81.7%

Table 6
Inferred group sizes by class. The implied source counts divide the class image totals by eight.

Class	Group sizes (images: groups)	Groups	Implied photographs	Known photographs
Anthracnose	8: 138, 16: 23, 24: 3	164	193	193
Fruit fly	8: 58, 16: 18, 24: 7, 32: 5, 48: 1, 72: 1, 112: 1	91	164	164
Healthy	1: 4, 5: 1, 7: 1, 8: 108, 16: 3	117	116	116
Total	—	372	473	473

All 382 official test images and 754 of 755 validation images belonged to inferred groups represented in the training folder. Membership includes both direct matches and connections through intermediate images. Under the invariant-hash threshold, 227 of 573 grouped-split test images and 155 of 382 official test images had a training neighbour (39.6% and 40.6%, respectively). Median nearest-training-image cosine similarity fell from 0.970 to 0.931 (Fig. 2). The validation files cover 1,461 distinct images, including 139 official test images. For 132 we also observed a training file with the same class and filename stem. This establishes a lower bound of 132/382 (34.6%) on official test images with a filename-family match in training. The export is incomplete for the other test images.

4.2 Classification on the group-disjoint split

Table 7 summarizes test performance. EfficientNetB0 achieved the highest mean accuracy, 96.10 ± 0.44% (95% CI [90.24, 99.76]), and group-level accuracy of 98.18 ± 0.00% over 55 inferred test groups. At the group level, the exact McNemar tests found no significant difference between any pair of backbones in any seed (smallest p = 0.25). MobileNetV2 reached 94.07 ± 0.52% with about one-ninth of the parameters of ResNet50 and a CPU latency of 24.2 ms per image. ConvNeXt-Tiny had the lowest mean image accuracy (92.55 ± 0.27%) and the longest CPU inference time (673.0 ms per image). VisionEdge deployed MobileNetV2 on a smartphone for cataract screening [41]. A comparable claim for guava

screening would require timing the model on the intended device and testing photographs acquired with that device.

Table 7
Test results on the group-disjoint split (mean ± SD over three seeds; CI from the group bootstrap)

Backbone	Accuracy (%)	95% CI (%)	Macro-F1 (%)	Group-level accuracy (%)	Parameters (M)	CPU latency (ms)
ResNet50	94.36 ± 0.71	[87.56, 98.95]	93.43 ± 0.95	93.94 ± 2.10	24.11	69.1
EfficientNetB0	96.10 ± 0.44	[90.24, 99.76]	95.78 ± 0.46	98.18 ± 0.00	4.38	38.1
MobileNetV2	94.07 ± 0.52	[88.53, 98.35]	93.16 ± 0.64	93.33 ± 1.05	2.59	24.2
ConvNeXt-Tiny	92.55 ± 0.27	[85.40, 98.17]	92.01 ± 0.29	94.55 ± 1.82	28.02	673.0

4.3 Official split and matched contamination experiment

ResNet50 reached 99.83 ± 0.15% accuracy on the official test folder and 94.36 ± 0.71% on the group-disjoint test set (Table 8). The 5.47-percentage-point difference reflects changes in both the training and the evaluation partitions, so it cannot be attributed to contamination alone.

On the fixed 354-image evaluation set, mean MobileNetV2 accuracy rose from 94.54% to 97.83% after 219 training images were replaced with members of the inferred test groups (Table 8). The paired increase was 3.30 percentage points (95% group-bootstrap CI 0.68–7.36). The three seed-specific increases were 3.11, 3.39, and 3.39 points. The group assignments establish which images the protocol moved. Their photographic relationship remains inferred.

Table 8
Official and group-disjoint ResNet50 results on different test sets, followed by the two MobileNetV2 arms on identical evaluation images. Accuracies are mean ± SD across three training seeds. The MobileNetV2 difference includes a paired 95% group-bootstrap interval.

Model and protocol	Test accuracy (%)	Test images / inferred groups
ResNet50, official	99.83 ± 0.15	382 / 231
ResNet50, grouped	94.36 ± 0.71	573 / 55
Between-protocol difference	5.47 points	—
MobileNetV2, clean	94.54 ± 0.65	354 / 55
MobileNetV2, contaminated	97.83 ± 0.59	354 / 55
Paired difference	3.30 points [0.68, 7.36]	Same images

4.4 Case study 2: leakage measured against published originals

Keypoint verification traced 3,258 of the 4,899 augmented images (66.5%; 60–72% per class) to an original, with a median of 10 traced augmented images per specimen group; 564 augmented images met the near-copy criterion for an original (hash distance of at most 2 bits). The 527 “original” images themselves formed 227 inferred specimen groups: 109 were unique, and the others belonged to groups of 2 to 13 images showing the same leaf or fruit after rotation, zoom, or re-framing (Fig. 3). The “original” folder is therefore not a set of independent photographs.

Table 9 and Fig. 4 compare the protocols. The image-level split gave 98.31 ± 0.53% accuracy, and the 1-nearest-neighbour classifier, which succeeds only by finding a similar training image, reached 99.77%. Splitting by specimen lowered accuracy to 93.38 ± 3.42% on augmented test images and to 91.22 ± 5.35% on the original test photographs. On identical original test photographs, including their augmented copies in training raised accuracy from 91.22% to 98.52 ± 1.97% (+ 7.30 points; higher in all 10 splits; Wilcoxon $p = 0.002$) and 1-nearest-neighbour accuracy from 89.65% to 100.00%. The size-matched control still reached 96.90 ± 3.22% (+ 5.68 points; higher in all 10 splits; $p = 0.002$), showing that the difference persisted after matching training-set size. A sensitivity analysis using a lower threshold of 15 inliers preserved the direction of the protocol differences. Because 4% of the clean test images still had a near-copy in training under the hash criterion, the residual impact of related training images remains uncertain. On the first split, the fine-tuned model classified 61 of 63 original test photographs correctly under clean training and all 63 under each protocol that allowed test copies. This comparison uses one split.

Table 9
Case study 2: test results by protocol. Frozen: logistic regression on frozen MobileNetV2 features, mean ± SD over 10 random splits. Fine-tuned: MobileNetV2 with its last 30 layers fine-tuned, first split only. Near-copy: test images within 10 bits of a training image under the invariant hash

Protocol	Training images	Test images	Near-copy in training (%)	Accuracy, frozen (%)	Accuracy, fine-tuned (%)	1-NN accuracy (%)
(A) Image-level split of augmented set	3,429	734 augmented	11.83 ± 1.02	98.31 ± 0.53	—	99.77 ± 0.16
(B) Specimen-level split, augmented test	2,336 (mean)	455 augmented (mean)	4.84 ± 1.01	93.38 ± 3.42	96.96	92.54 ± 2.52
(B) Specimen-level split, original test photographs	2,336 (mean)	76 originals (mean)	3.98 ± 2.32	91.22 ± 5.35	96.83	89.65 ± 4.50
(C) Same test photographs, copies in training	4,432 (mean)	76 originals (mean)	99.08 ± 1.39	98.52 ± 1.97	100.00	100.00 ± 0.00
(C) Same, size-matched training set	2,336 (mean)	76 originals (mean)	56.42 ± 6.13	96.90 ± 3.22	100.00	99.88 ± 0.39

4.5 Case study 3: transfer across datasets and acquisition confounding

Table 10 compares in-dataset and cross-dataset accuracy. Every transfer lost accuracy, with frozen and with fine-tuned features. For healthy leaf, red rust, and scab, frozen-feature accuracy was 99.1% within the case-study-2 dataset and 93.9% within BD Guava, but 78.5% when a model trained on the first was tested on the second and 84.6% in the opposite direction; with fine-tuning the transfers gave 76.4% and 76.5%. Without the leaf-versus-fruit cue that identifies scab, healthy leaf versus red rust was classified with 100.0% and 89.1% accuracy within the two datasets, but with 72.8% and 86.8% across them (74.1% and 76.1% fine-tuned). For healthy leaf versus insect damage, within-dataset accuracy was 95.4% for BD Guava and 99.6% for the six-class dataset, whereas transfer gave 72.4% and 73.8% (75.9% and 78.4% fine-tuned); the fine-tuned model trained on BD Guava recognized 31.5% of the insect-damaged leaves of the six-class dataset.

The 1,092 BD Guava images span 15 date–location sessions. Six of ten classes were each collected in a single session, and each session included at most two classes. Assigning each image the most frequent class of its session would classify 87.0% correctly. A decision tree using date, location, temperature, humidity, and weather reached 86.9% in cross-validation. Among the three leaf classes observed in two

sessions each (336 images), the central-disc classifier reached 84.8% balanced accuracy under grouped cross-validation and 45.9% on held-out sessions. For the border ring, corresponding values were 60.1% and 25.8% (Table 11). Both regions therefore carried acquisition-related information. The central disc can include background and does not isolate specimen appearance.

Table 10
Case study 3: accuracy (%) within and across datasets on shared classes. Each cell gives frozen-feature accuracy [95% group-bootstrap CI] / fine-tuned accuracy. In-dataset values are five-fold cross-validation with near-duplicates kept together; A → B means trained on dataset A and tested on dataset B. Images per dataset: 319 / 326 (first row), 213 / 239 (second), and 259 / 1,272 (third)

Shared classes	Dataset A / dataset B	Within A	Within B	A → B	B → A
Healthy leaf, red rust, scab	Case study 2 / BD Guava	99.1 [97.8, 100.0] / 99.4	93.9 [91.8, 98.8] / 89.9	78.5 [74.1, 89.2] / 76.4	84.6 [79.7, 89.4] / 76.5
Healthy leaf, red rust	Case study 2 / BD Guava	100.0 [100.0, 100.0] / 99.1	89.1 [86.7, 97.6] / 86.6	72.8 [69.1, 86.5] / 74.1	86.8 [80.5, 92.1] / 76.1
Healthy leaf, insect damage	BD Guava / six-class	95.4 [90.2, 97.1] / 95.4	99.6 [99.2, 99.9] / 99.7	72.4 [70.0, 74.8] / 75.9	73.8 [70.1, 83.8] / 78.4

Table 11
Balanced accuracy (%) for the central disc and border ring in BD Guava. Session identification is evaluated within a single class. The disease classification task uses three leaf classes

Task	Chance	Central disc	Border ring
Disease, 3 leaf classes, random group CV	33.3	84.8	60.1
Disease, 3 leaf classes, session held out	33.3	45.9	25.8
Session, within algal leaf spot (2 sessions)	50.0	79.2	68.1
Session, within red rust (2 sessions)	50.0	84.4	85.3
Session, within healthy fruit (3 markets)	33.3	69.7	55.8

Near-duplicates are also common: in BD Guava, 599 of the 1,092 images have an invariant hash within 10 bits of another image, many of them consecutive shots of the same leaf, and 11 files are byte-identical copies; in the six-class dataset, 89 images have an embedding near-duplicate.

4.6 Per-class performance and errors

Figure 5 shows the seed-42 confusion matrices. Across three seeds, ResNet50 had its lowest mean recall on Healthy fruit (82.51%). At seed 42 it assigned 26 of 141 Healthy images to Fruit fly and two to Anthracnose. The illustrated errors include Healthy fruit assigned to Fruit fly and one Fruit fly image assigned to Healthy.

4.7 Explanations

Figure 6 shows the seed-42 explanations for ResNet50. The correctly classified Anthracnose examples highlight areas with visible discoloration. In row 7, the highlighted area includes the dark calyx of a Healthy fruit classified as Fruit fly. In row 8, the model classified a Fruit fly image as Healthy, with Grad-CAM emphasizing the central fruit surface. These examples illustrate individual decisions and lack expert lesion masks.

Blurring the most salient 10% of pixels reduced the predicted-class probability by 0.101 on average, compared with 0.042 for the control mask. The Grad-CAM mask produced a larger reduction than the control for 71.7% of the 60 correctly classified test images (Table 12).

Table 12
Deletion test on 60 correctly classified test images: mean drop in predicted-class probability

Pixels blurred	Grad-CAM mask	Control mask	Images where Grad-CAM drop > control (%)
5%	0.046	0.026	63.3
10%	0.101	0.042	71.7
20%	0.215	0.077	83.3

Discussion

5.1 What accuracy on GFDD24 measures

GFDD24's official split permits extensive overlap between inferred groups, making its accuracy difficult to interpret as performance on new fruit. The fixed grouped split tests generalization to 55 held-out groups. Differences between official and grouped accuracies combine changes in the training and test sets. The matched GFDD24 comparison holds the evaluation set and training-set size fixed and finds a 3.30-point increase under the replacement protocol. In case study 2, the same original test photographs were evaluated under both conditions. Frozen-feature accuracy increased by 5.68 points when a size-matched training set permitted augmented copies of test photographs. The session and cross-dataset evaluations identify additional sources of variability even when related images are grouped.

5.2 Checklist for evaluating a plant-disease benchmark

1. Group the files before splitting. Keep every augmentation of a photograph together. For claims about new specimens, group repeated views of each specimen as well.

2. Audit the groups. Inspect the original-image folder, compare accepted and rejected candidate pairs, and examine oversized groups.

3. Report independent test units. Give both image and group counts, and estimate uncertainty by resampling groups.

4. Test acquisition and dataset shifts. Hold out complete collection sessions and evaluate shared labels on an independently collected dataset.

5. Release the split manifest. Include image, source, specimen, session and partition identifiers wherever available.

5.3 Explanations

Blurring the regions ranked highest by Grad-CAM reduced predicted-class probability more, on average, than a control mask transferred from another image. This comparison measures the classifier's sensitivity to those regions. Spatial agreement with a lesion requires an annotated reference. For example, Alam et al. evaluated impacted-tooth segmentation on panoramic radiographs using intersection over union [42]. An analogous test here would compare highlighted regions against lesion masks drawn by a plant pathologist.

5.4 Limitations

The filename-based audit establishes augmentation-family membership for the exported paths. It cannot establish the identity of individual fruit across captures. The audit covers all oversized components and a score-stratified candidate sample. Dataset-wide retrieval recall and source-level overlap in the group-disjoint split require a complete file manifest. All three training seeds used the same split with 55 test groups, so the bootstrap intervals do not quantify between-split variation. The three-class GFDD24 fruit classifier lacks an independent external test. Cross-dataset experiments covered separate shared-label tasks, whose label definitions may differ. In case study 2, one third of augmented images could not be traced. Fine-tuning covered one of ten splits and the classes partly coincide with plant organ. Session holdouts in BD Guava involved just two sessions per selected leaf class. The central disc and border ring approximate plant and background regions. Their results cannot establish what an accurate plant segmentation would remove. In the size-matched comparison in case study 2, the pool that allowed test copies also included augmented images whose source could not be traced. The 5.68-point contrast therefore describes the change in training protocol. It cannot isolate the effect of verified copies alone. Repeated splits share specimens, which limits inference from split-level signed-rank tests.

Conclusion

On GFDD24, all official test images belonged to inferred groups represented in training. Within the validation export, 132 official test images also had a training image from the same filename family. In the size-matched GFDD24 experiment, replacing 219 training images with members of inferred test groups increased accuracy on the same 354 images by 3.30 percentage points. Keeping groups together

yielded an EfficientNetB0 image-level accuracy of 96.10 ± 0.44% across three training seeds. In the dataset with published originals, permitting augmented test copies in an equally large training set raised frozen-feature accuracy on the same test photographs from 91.22% to 96.90%. Holding out acquisition sessions also reduced balanced accuracy, and transfer across datasets with shared labels yielded frozen-feature accuracies of 72.4–86.8%. Future benchmarks should document image grouping, publish identifiable partitions, and evaluate generalization across sessions and data sources.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable. This study used only previously published, publicly archived image datasets and involved no human participants, human data, or animals.

Consent for publication

Not applicable.

Funding

This research received no funding.

Author contributions

Abdul Ahad developed the methodology and supervised the research. Mirza Farhan Shahriar developed the software and ran the experiments. Md. Naimur Rahman curated and validated the data. Nourash Azmine Chowdhury prepared the figures and drafted the manuscript. All four authors approved the final manuscript and its submission to SN Computer Science.

Data availability

The four datasets are archived on Mendeley Data [3, 13, 39, 40]. The authors used the released image files and dataset versions specified in Materials and Methods.

Code availability

The analysis code is available from the corresponding author upon request.

References

1. Mohanty SP, Hughes DP, Salathé M (2016) Using deep learning for image-based plant disease detection. Front Plant Sci 7 Article 1419. https://doi.org/10.3389/fpls.2016.01419
2. Ferentinos KP (2018) Deep learning models for plant disease detection and diagnosis. Comput Electron Agric 145:311–318. https://doi.org/10.1016/j.compag.2018.01.009
3. Amin MA, Mahmud MI, Rahman AB, Parvin MA, Mamun MAA (2024) Guava fruit disease dataset. Mendeley Data, V1. https://doi.org/10.17632/bkdkc4n835.1
4. Kilci O, Koklu M (2024) Classification of guava diseases using features extracted from SqueezeNet with AdaBoost and gradient boosting. In Proceedings of the 4th Int. Conf. Frontiers in Academic Research, pp. 578–586
5. Rijon TA, Arafath Y (2025) A novel CNN–gradient boosting ensemble for guava disease detection. arXiv:2512.19989.
6. Ahmed M et al (2025) Automated guava disease detection using transfer learning with ResNet-101. Food Sci Nutr 13(12) Article e71348. https://doi.org/10.1002/fsn3.71348
7. Ghosh S, Anika SA, Ahmed UH, Alam BMS, Noor MT, Niloy NT (2025) Healthy harvests: A comparative look at guava disease classification using InceptionV3. In Proceedings of the 16th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT), arXiv:2602.10967
8. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4(9) Article 100804. https://doi.org/10.1016/j.patter.2023.100804
9. Gerz F, Jelali M (2026) Towards reliable data augmentation in machine learning: Practices to prevent data leakage. Smart Agricultural Technol Article 102462. https://doi.org/10.1016/j.atech.2026.102462
10. Noyan MA (2022) Uncovering bias in the PlantVillage dataset. arXiv:2206.04374
11. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In Proceedings of the Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 9505–9515
12. Petsiuk V, Das A, Saenko K (2018) RISE: Randomized input sampling for explanation of black-box models. In Proceedings of the British Machine Vision Conf. (BMVC)
13. Rajbongshi A et al (2022b) Guava leaves and fruits dataset for guava disease recognition through machine learning and deep learning. Mendeley Data, V1. https://doi.org/10.17632/x84p2g3k6z.1
14. Mostafa AM, Kumar SA, Meraj T, Rauf HT, Alnuaim AA, Alkhayyal MA (2022) Guava disease detection using deep convolutional neural networks: A case study of guava plants. Appl Sci 12(1) Article 239. https://doi.org/10.3390/app12010239
15. Almadhor A, Rauf HT, Lali MIU, Damaševičius R, Alouffi B, Alharbi A (2021) AI-driven framework for recognition of guava plant diseases through machine learning from DSLR camera sensor based high resolution imagery. Sensors 21(11). Article 3830. https://doi.org/10.3390/s21113830

16. Kılcı O, Koklu M (2025) Guava fruit disease classification using deep learning and machine learning models. Res Agric Sci 56(3):217–226. https://doi.org/10.17097/agricultureatauni.1665941
17. Shahade AK, Shrimant GV, Phadke M, Meshram VV, Meshram VA, Wankhede DS, Shahade MR (2026) Deep learning for early detection of guava fruit anthracnose and fruit fly infestation. BMC Plant Biol 26:1266. https://doi.org/10.1186/s12870-026-09125-1
18. Biswas J et al (2026) GuavaVision AI: An explainable deep learning framework for automated classification, lesion localization, and segmentation of guava diseases. https://doi.org/10.64898/2026.06.18.733093. bioRxiv
19. Rajbongshi A, Sazzad S, Shakil R, Akter B, Sara U (2022a) Data Brief 42., Article 108174. https://doi.org/10.1016/j.dib.2022.108174. A comprehensive guava leaves and fruits dataset for guava disease recognition
20. Shihab MR, Saim NI, Mojumdar MU, Raza DM, Siddiquee SMT, Noori SRH, Chakraborty NR (2025) Data Brief 59:111378. https://doi.org/10.1016/j.dib.2025.111378. Image dataset for classification of diseases in guava fruits and leaves
21. Zauner C (2010) Implementation and benchmarking of perceptual image hash functions. M.S. thesis, Upper Austria Univ. Appl. Sci., Hagenberg, Austria
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE Int. Conf. Comput. Vis. (ICCV), pp. 618–626
23. Ribeiro MT, Singh S, Guestrin C (2016) 'Why should I trust you?': Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 1135–1144
24. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
25. Fahmid A et al (2025) Multi-label thoracic disease detection and weakly supervised localization in chest X-rays using DenseNet121 and probabilistic CAM pooling. In Proceedings of the IEEE RAAICON, pp. 450–455. https://doi.org/10.1109/RAAICON69033.2025.11502273
26. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 248–255
27. Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the Int. Conf. Mach. Learn. (ICML), pp. 6105–6114
28. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
29. Fischler MA, Bolles RC (1981) Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778

31. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4510–4520
32. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 11976–11986
33. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In Proceedings of the Int. Conf. Learn. Represent. (ICLR)
34. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2818–2826
35. Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman & Hall, New York, NY, USA
36. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2):153–157
37. Pedregosa F et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
38. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull 1(6):80–83
39. Shuvo SKB (2026) BD Guava fruit and leaf disease dataset. Mendeley Data, V5, Aug. 2026. https://doi.org/10.17632/rgp9xy5gy2.5
40. Islam MA, Das B (2026) A multi-class guava (Psidium guajava L.) leaf diseases dataset for precision agriculture and AI-based crop health monitoring. Mendeley Data, V1, Mar. 2026. https://doi.org/10.17632/ng85zg4xm4.1
41. Chowdhury NA, Ahad A, Alam AM, Fahim A (2024) VisionEdge: Cataract detection via smartphone. In Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT). https://doi.org/10.1109/ICCIT64611.2024.11022415
42. Alam SS, Ahad A, Ahmed S, Dudley J, Farook TH (2025) Using deep learning to segment impacted molar teeth from panoramic radiographs. Digit Dentistry J 1(1). https://doi.org/10.1016/j.ddj.2025.100007. Article 100007

Figures

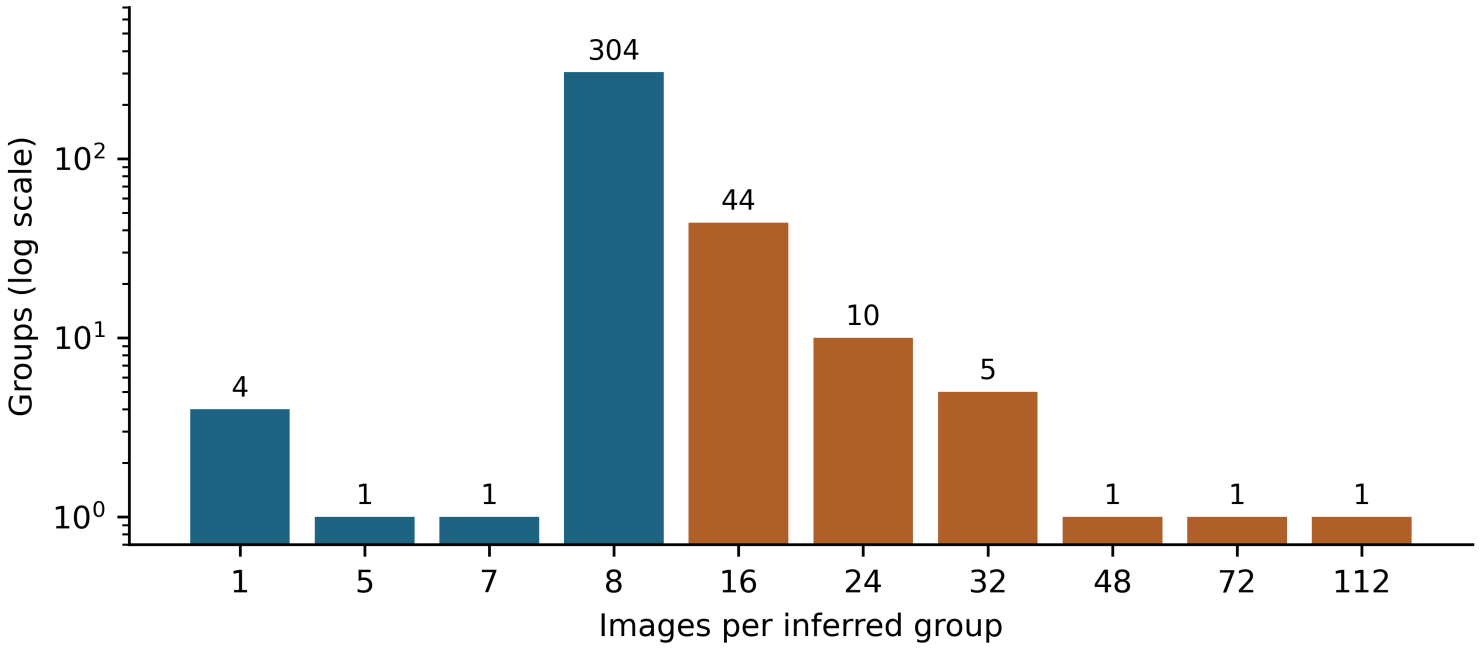

Figure 1

Size distribution of the 372 inferred GFDD24 groups. Orange bars mark groups larger than one eight-image family. Counts above the bars give numbers of groups.

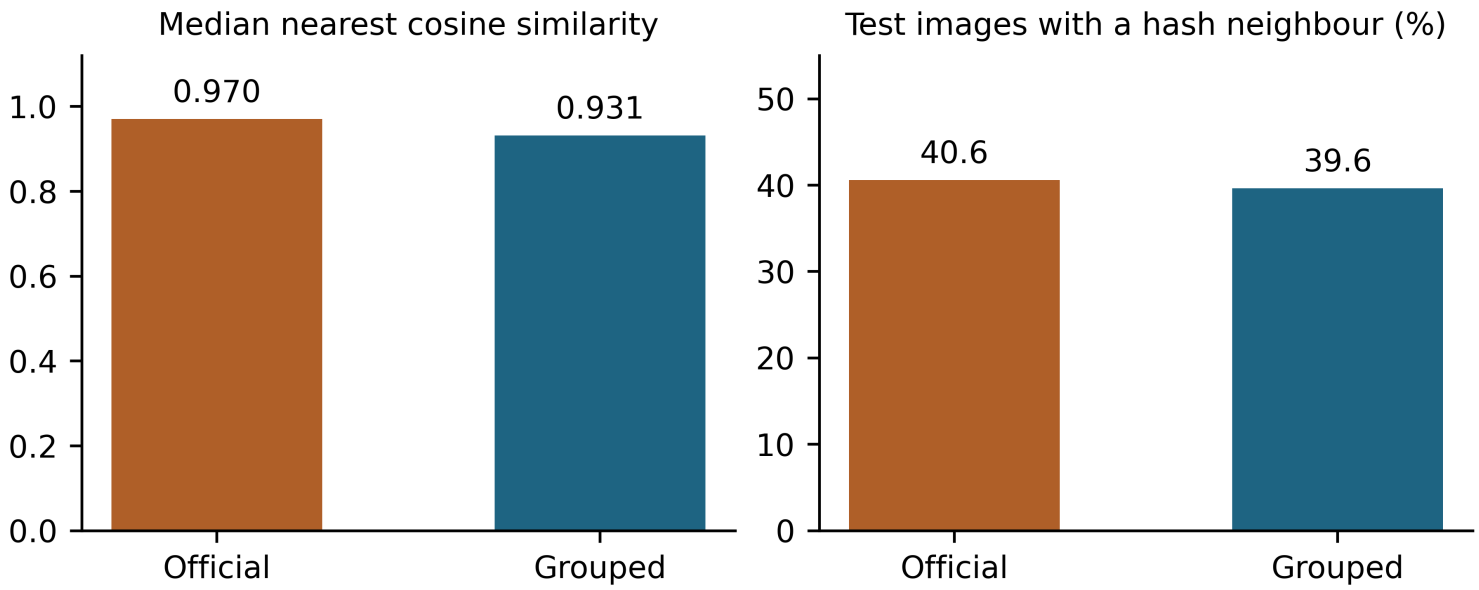

Figure 2

Nearest-training-image similarity on the official and group-disjoint splits. Left, median cosine similarity. Right, percentage of test images with a training neighbour within ten bits under the invariant hash.



Figure 3

Styler-end rot examples from Rajbongshi et al. [13]. Top row, originals 68–72 in one inferred specimen group. Bottom row, traced images 150, 154, 914, 913, 934, and 938 with 426, 218, 101, 64, 45, and 38 inliers, respectively. Panels are cropped from the original comparison and arranged in the same order. Source images licensed under CC BY 4.0.

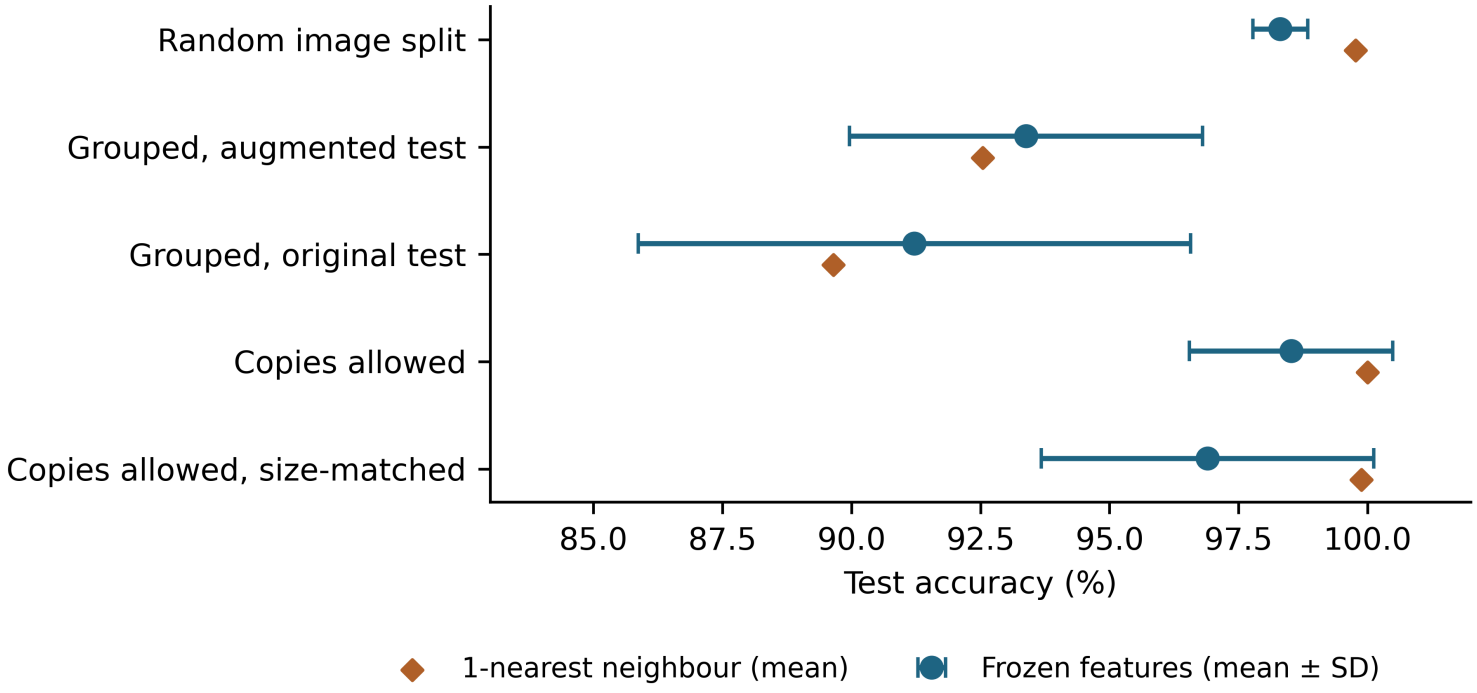

Figure 4

Test accuracy across five evaluation protocols in case study 2. Circles and whiskers show the mean and one standard deviation of frozen-feature accuracy over ten splits. Diamonds show mean 1-nearest-neighbour accuracy. Test images differ between the first three protocols. The last three share their original test photographs within each split.

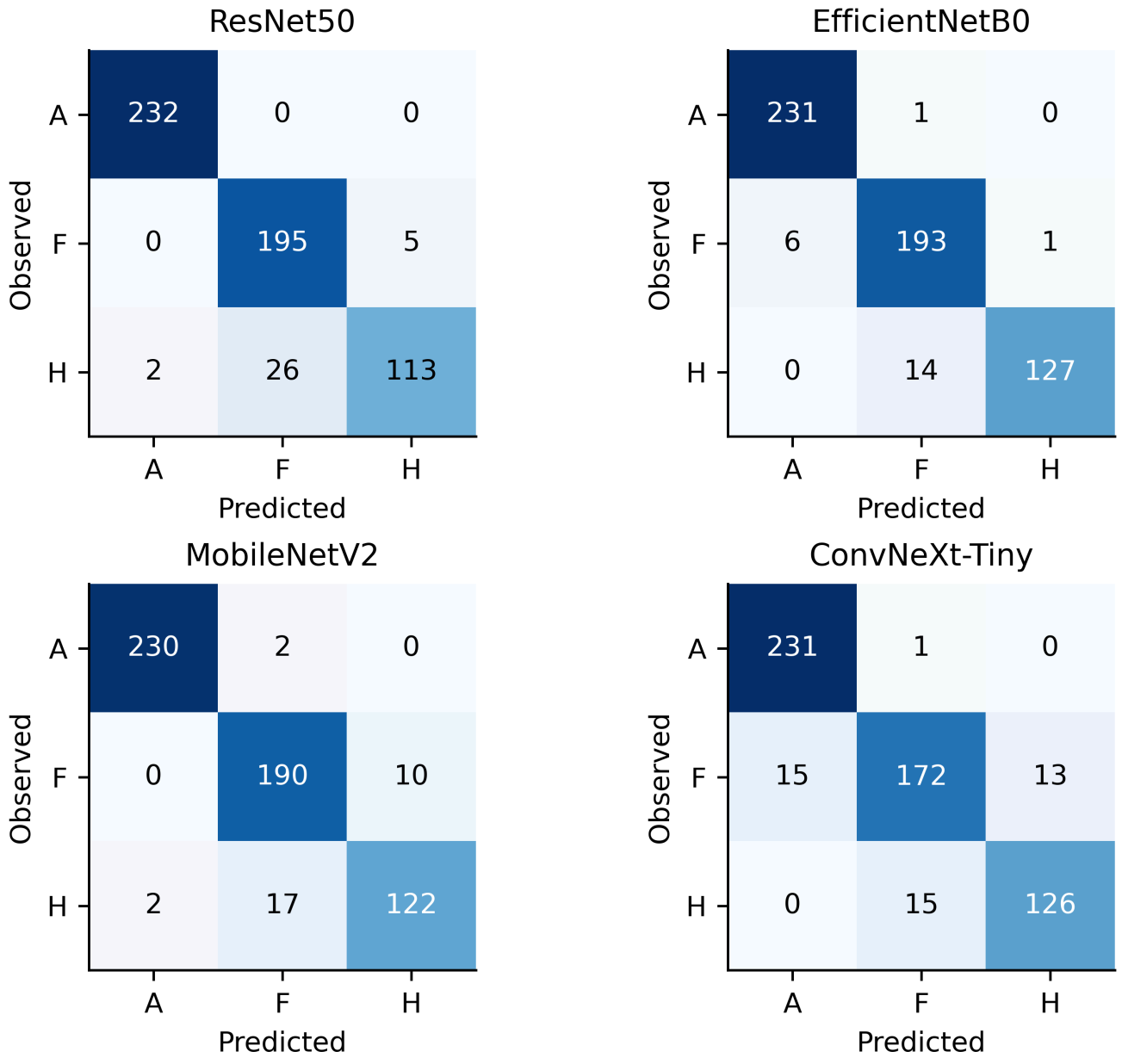

Figure 5

Confusion matrices on 573 group-disjoint test images (seed 42). A denotes Anthracnose, F denotes Fruit fly, and H denotes Healthy. Cells contain image counts.

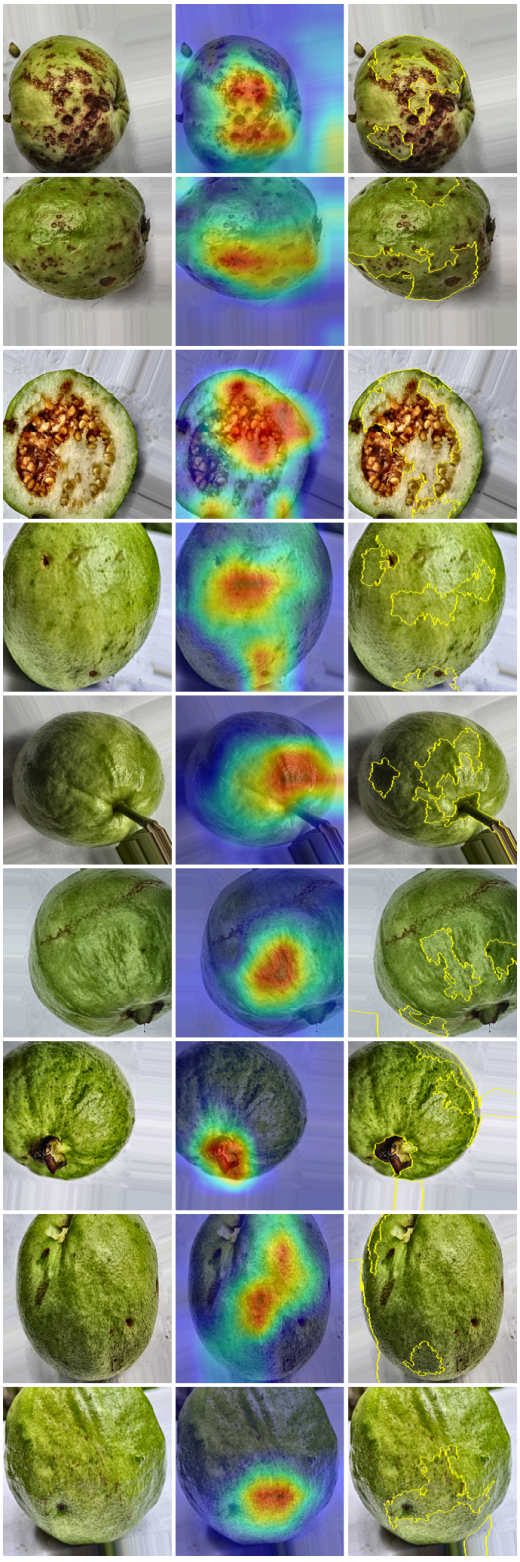

Figure 6

ResNet50 predictions on GFDD24 images [3]. Columns show the input, Grad-CAM overlay, and five highest-weight positive LIME regions. Rows 1–6 show two correct predictions each for Anthracnose, Fruit fly, and Healthy, in that order. Rows 7 and 9 show Healthy fruit assigned to Fruit fly. Row 8 shows Fruit fly assigned to Healthy. Source images by Amin et al. [3], CC BY 4.0. Overlays are model outputs.