

Multi-Crop Disease Pretraining for Black Pepper Leaf Disease Classification: A Leaf-Grouped Evaluation of CNN Encoders

Siji P. D.^{1*}, Prameeja Prasadhan², Ambily Jacob³

^{1,2,3}Assistant Professor, Department of Computer Science, St. Joseph's College (Autonomous), Irinjalakuda, Kerala, India

Article information

Received: 18th April 2026

Received in revised form: 20th May 2026

Accepted: 25th June 2026

Available online: 30th July 2026

Volume: 2

Issue: 3

DOI: <https://doi.org/10.63090/IJITRS/3139.3209.0035>

Abstract

Published work on black pepper (*Piper nigrum* L.) leaf disease reports accuracies close to 100%, yet the public image sets behind such figures are small and full of repeated photographs of the same leaf. We asked a narrower question that matters more to a grower with few labelled images: does a convolutional network trained on many other crops transfer to black pepper better than an ordinary ImageNet network? Two public datasets were used (1,500 and 819 images). SIFT matching with RANSAC verification traced them to about 590 and 121 distinct leaves, and every evaluation kept all photographs of a leaf on one side of the split. We compared two frozen EfficientNet-B3 encoders with identical architecture: the ImageNet initialisation and a multi-task model trained on 38 PlantVillage classes from 14 crops. With all training leaves available, both reached 97% to 99% accuracy and did not differ. With one leaf per class, the multi-crop encoder raised macro F1 from 0.51 to 0.73 on the first dataset ($p < 0.001$, 30 paired draws) and from 0.65 to 0.71 on the second, where the gain was not consistent across training sizes. Used without any black pepper training, its source classifier necessarily named a wrong crop, yet did so with a mean confidence of 0.95. Control inputs showed that image corners with no leaf pixels predicted the disease class of the first dataset with 87.1% accuracy. Error level analysis, a JPEG forensics transform sometimes used as CNN preprocessing, lowered leaf-grouped accuracy on both datasets. Multi-crop pretraining is worth using when labels are scarce, but accuracy on current black pepper benchmarks says little about field performance until acquisition shortcuts are removed.

Keywords:- Black Pepper, Convolutional Neural Network, Data Leakage, Few-Shot Learning, Plant Disease Classification, Shortcut Learning, Transfer Learning.

I. INTRODUCTION

Black pepper is a perennial vine grown mainly in South and South-East Asia, with Kerala among its traditional centres. Several of its diseases first show on the leaves. Foot rot is associated with *Phytophthora* and other soil-borne pathogens [1], while anthracnose, called fungal pollu in Kerala, is caused by *Colletotrichum gloeosporioides* and spots both leaves and spikes [2]. Visible symptoms often come before yield loss. That is why image-based screening of leaves appeals to smallholders who cannot send every sample to a laboratory for confirmation.

Convolutional neural networks (CNNs) are the default tool for that screening. On the PlantVillage collection, CNNs separate 38 crop and disease classes with accuracies above 99% [3], [4], and reviews of

agricultural deep learning report similar figures for many crops [5], [6]. Black pepper studies report similar numbers. Kini et al. reported 99.67% for a ResNet-18 on six leaf classes [7], Sreethu et al. reported 98.72% for a custom four-block CNN [8], and an attention network trained on a public Kaggle set combined with PlantVillage bell pepper images was reported to exceed three earlier detectors [9]. The source images for the first two studies are available only on request.

High accuracy on a small benchmark can have more than one cause. Leakage is the first. When several photographs of one leaf land in both the training and the test split, a model can score well by recognising the leaf rather than the disease, and failures of this kind have distorted published results in many fields [10], [11]. Shortcuts are the second. A model can latch onto a backdrop or lighting that happened to differ between classes [12], [13]; in chest radiography, one network learned to recognise the hospital rather than the pneumonia [14]. Barbedo argued years ago that plant disease image sets are small and narrow enough for both problems to matter [15], [16].

Labelled black pepper images are also scarce. Each confirmed image costs a field visit and an expert's time, so a practical system will usually start from a network trained on something else. ImageNet is the common starting point [17], [18]. A network already trained on tens of thousands of leaf images from other crops, most of them diseased, might carry more useful features. Transfer learning has taken ImageNet weights to a new crop, cassava [19], and few-shot transfer from laboratory to field images has been studied for other plant diseases [20]. We found no study that tests multi-crop pretraining on black pepper while controlling for the duplication and shortcut problems described above.

This paper makes that test on the two black pepper leaf datasets that anyone can download. The contributions are listed below.

- An audit of both datasets, including a SIFT and RANSAC procedure that groups repeated photographs of one leaf. The 1,500 and 819 images reduce to roughly 590 and 121 leaves.
- A matched comparison of two frozen EfficientNet-B3 encoders that differ only in their training data (ImageNet versus 14 PlantVillage crops), under random and leaf-grouped splits, with one to ten training leaves per class, and with no black pepper training at all.
- Control experiments with corner-only, silhouette-only and error level analysis (ELA) inputs, plus group-level permutation nulls, that measure how much of the accuracy on each dataset can be explained without looking at diseased tissue.

II. RELATED WORK

A. Deep Learning for Black Pepper Leaf Disease

Khew et al. photographed 947 leaves from farms in Sarawak under controlled conditions and trained CNNs on two diseases and six nutrient disorders [21]. Kini et al. worked from field photographs taken in Karnataka. Their six classes were anthracnose, slow wilt, early-stage Phytophthora infection, advanced Phytophthora infection, yellowing and healthy leaves, and among four ImageNet-pretrained networks ResNet-18 did best, at 99.67% with a 70:30 split [7]. Sreethu et al. collected 2,786 images in Kerala over two seasons. After augmentation to 18,234 images they reported 98.72% on a held-out tenth [8]. Swathi Chandra and Parthasarathy merged the public Kaggle blight and yellow mottle set, which we also use, with PlantVillage bell pepper classes [9], [22]. None of the published descriptions mentions grouping images by leaf before splitting. None reports a background control either.

B. Transfer Learning and Few-Shot Plant Disease Recognition

Fine-tuning ImageNet networks is the standard recipe in plant disease work [23], [24]. Its benefit depends on the target data. Kornblith et al. found that networks with better ImageNet accuracy usually, though not always, transfer better [18], while He et al. showed that ImageNet pretraining mostly speeds up convergence when the target data are plentiful and adds little to final accuracy [25]. Small target sets behave differently. There, a linear classifier on a good frozen embedding can match elaborate few-shot methods [26], and Argüeso et al. used a related idea to classify field images of plant diseases from a handful of samples per class [20]. Pan and Yang survey the wider theory of transfer between related domains [27].

C. Leakage, Duplicates and Shortcuts

Kapoor and Narayanan catalogued leakage in machine-learning-based science and found duplicated or non-independent samples across train and test sets to be one of the commonest failures [10]. Barz and Denzler found near-duplicates of test images in the CIFAR training sets [11]. Roberts et al. argued that cross-validation must respect the grouping structure of the data, or error estimates become optimistic [28]. Geirhos et al. reviewed shortcut learning, in which a network solves a benchmark through an unintended cue [12], and hidden subgroups can make average accuracy misleading [29]. In plant pathology, networks trained on laboratory photographs lose

much of their accuracy on field images [3], [30]. Training on individual lesions has been proposed as a way to enlarge and diversify small plant disease sets [31].

III. MATERIALS AND METHODS

Fig. 1 outlines the study. The datasets are described first, then the audit and leaf grouping, the two encoders, and the evaluation settings.

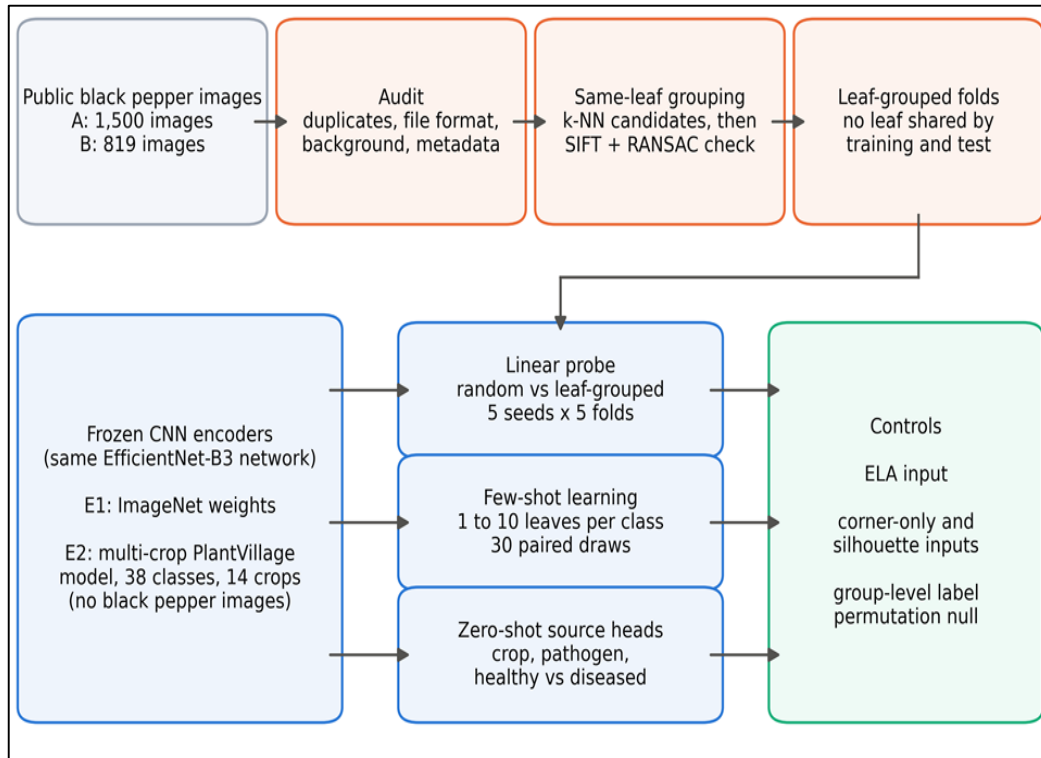


Fig 1: Overview of the study. Both public datasets are audited and grouped by physical leaf before any split is made. Two frozen EfficientNet-B3 encoders with the same architecture are then compared in three settings, and control inputs test for cues that do not come from diseased tissue.

A. Datasets

Dataset A is the Black Pepper Leaf Disease Mini-Dataset [32]. Its description states that the photographs were taken in plantations in the Western Ghats region of Wayanad, Kerala, and that the images were already augmented by rotation and mirroring. It contains 1,500 images at 256×256 pixels, 500 each for foot rot, pollu disease and slow decline. The class names come from the uploader, who notes that they may differ from standard terminology; we use them as given. Dataset B [22] contains 819 images at 500×500 pixels, 273 each for healthy leaves, leaf blight and yellow mottle virus. Its leaves were cut out and placed on a uniform grey backdrop. Fig. 2 shows examples. Neither dataset records the farm, date, device or how labels were confirmed.

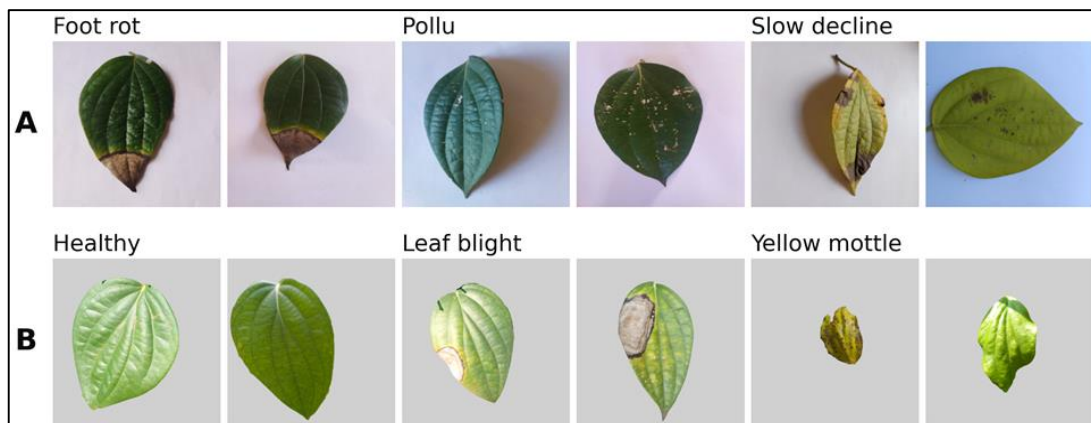


Fig 2: Example images. Row A shows Dataset A (detached leaves photographed on cloth) and row B shows Dataset B (background replaced by uniform grey).

B. Audit and Same-Leaf Grouping

Each file was hashed with MD5 and its container format, colour mode and metadata were recorded. We then looked for repeated photographs of the same physical leaf, which may differ by rotation, mirroring, scale or colour adjustment. A two-stage search was used. In the first stage, every image was embedded with an ImageNet EfficientNet-B0 [33] under the eight flips and quarter turns of the square, and its 20 nearest neighbours by maximum cosine similarity became candidates. The embedding alone was not trusted, because plain healthy leaves on a flat backdrop look alike to it.

The second stage checked each candidate pair geometrically. SIFT keypoints [34] were extracted from the grey-scale leaf region at 384×384 pixels, matched with Lowe's ratio test at 0.75, and fitted with a RANSAC similarity transform [35] with a 6-pixel reprojection threshold; the mirrored image was tried as well. Two images were linked when the fit had at least 10 inliers and a plausible scale between 0.4 and 2.5. We chose the threshold by inspecting sampled pairs in bands of inlier counts. Pairs below 8 inliers were different leaves, and pairs in the 8 to 20 band were mostly the same leaf seen under a different transform. Connected components of the link graph, merged with exact-duplicate files and files sharing a name stem, define a leaf group. A stricter threshold of 20 inliers is reported as a sensitivity check. Fig. 3 shows linked pairs.

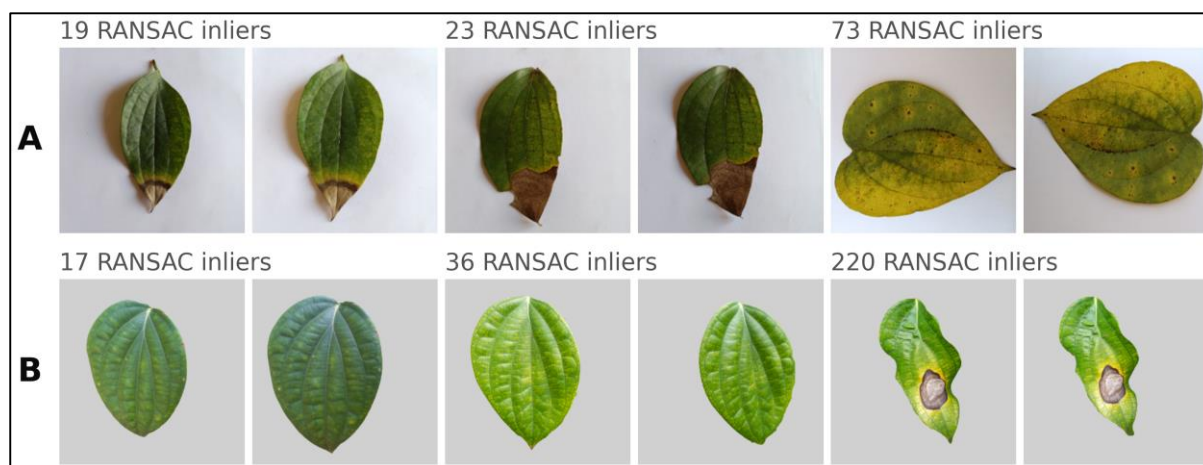


Fig 3: Pairs of images linked as the same physical leaf, with the number of RANSAC inliers above each pair. Rescaled, rotated, mirrored and colour-shifted copies of one leaf are common in both datasets.

C. Encoders

The transfer comparison uses one architecture with two sets of weights. Encoder E1 is EfficientNet-B3 with the ImageNet weights published in the timm library (tag ra2_in1k). Encoder E2 starts from the same timm ImageNet weights (the library default for this architecture) and was trained further, in an earlier, separate doctoral study, as a multi-task model on the PlantVillage colour images: a squeeze-and-excitation and spatial attention block on the final feature map, a 512-unit projection, and heads for 38 disease classes, 14 crops and five pathogen types. Its training log records a best validation accuracy of 99.95% on a random image-level split of PlantVillage. That figure is not re-examined here and plays no role in the analysis. The training code contains no black pepper images. Its documentation mentions gradient reversal for domain adaptation [36], but the trained checkpoint has no domain branch, so E2 is best described as a multi-crop supervised model.

Both encoders process an image the same way: resize the shorter side to 255 pixels, centre-crop 224×224 , normalise with the ImageNet mean and standard deviation. The 1,536-dimensional pooled feature (after the attention block for E2) is the representation used throughout. The encoders are frozen. Keeping them frozen isolates what each set of weights already knows, and it avoids tuning choices that could favour one of them.

D. Evaluation Protocols

A multinomial logistic regression ($C = 0.1$) was fitted on standardised features [37]. We compared three ways of splitting the data. Protocol P1 copies a common practice: each image is first expanded into eight rotated and mirrored copies, and the pooled copies are then split 80:20 at random, so a test image may have its own rotated twin in training. P2 is stratified five-fold cross-validation on the images as distributed. P3 uses stratified group five-fold cross-validation, which places all images of a leaf in one fold [28]. P2 and P3 were each repeated with five random seeds, giving 25 test folds, and for every fold we recorded the share of test leaves that also appear in training.

Few-shot learning was simulated by drawing k leaf groups per class at random for training, with k in $\{1, 2, 3, 5, 10\}$. All images of the remaining leaves formed the test set. We made thirty draws for each k . E1 and E2 always saw exactly the same draws, which allowed a Wilcoxon signed-rank test over the paired macro F1 scores.

Zero-shot use needed no training at all. E2's own heads were applied to every black pepper image, and we recorded the predicted crop, pathogen type and disease class together with the maximum softmax probability. Dataset B includes healthy leaves, so it allows one more test: any PlantVillage class ending in “healthy” was read as a healthy prediction and every other class as diseased, and the balanced accuracy of that rule was compared with the 0.5 obtained by always answering “diseased”.

E. Control Inputs

Four extra inputs were built from a leaf mask. For Dataset A the mask comes from the distance between each pixel and the backdrop colours sampled along the image border, followed by Otsu thresholding and hole filling; for Dataset B it is an exact key on the grey value 208. Leaf only replaces the backdrop by grey. Silhouette only keeps the binary mask and nothing else. Corners only tiles the four 56×56 corner patches of the 256-pixel image after painting any leaf pixel grey, so neither leaf texture nor a usable outline survives. The fourth input is ELA: the absolute difference between the image and its JPEG re-encoding at quality 90, rescaled per image, as used in forgery detection [38], [39]. Control inputs were embedded with ImageNet EfficientNet-B0 and evaluated under P2 and P3.

Chance under a grouped design is harder to judge than one in three suggests. To estimate it, labels were permuted at the level of leaf groups, so that each group kept a single label, and P3 was rerun 50 times; the 95th percentile of those permuted accuracies is the null reference. A random forest (300 trees, depth 6) gave a second check. Its hand-built features ignore leaf texture and colour: file format, alpha channel and metadata, the colour and spread of the backdrop, leaf area, bounding-box aspect ratio and position, letterbox bars, and ELA statistics. Leaf area and shape still carry some disease information, because some diseases stunt or distort leaves. For that reason the forest is read alongside the corner input, which contains no leaf at all.

F. Implementation

Experiments ran in Python with PyTorch, timm, OpenCV and scikit-learn. Feature extraction and all linear probes ran on a four-core laptop CPU.

IV. RESULTS

A. Audit Findings

Dataset A holds 1,456 distinct files among 1,500 images, and 87 images have at least one byte-identical twin. Its 54 PNG files carry a .jpeg extension. Every one of them is in the foot rot class. The 78 letterboxed portrait photographs, framed by white bars, are all slow decline images. Dataset B shows a similar pattern: its backdrop is one grey value throughout, and its 8 genuine JPEG files all belong to the yellow mottle class. Table 1 lists the grouping result. At the main threshold the 1,500 images of Dataset A reduce to 590 leaves and the 819 images of Dataset B to 121; the 273 leaf blight images come from only 28 leaves. No group mixes two classes.

Table 1. Images and Distinct Leaves

Set	Images	Unique files	Leaves (≥ 10 inliers)	Leaves (≥ 20 inliers)	Images per leaf	Leaves per class	Largest group
Dataset A	1,500	1,456	590	849	2.5	162 / 183 / 245	17
Dataset B	819	819	121	181	6.8	38 / 28 / 55	31

B. Effect of the Split Protocol

Fig. 4 and Table 2 show the effect of the split on the ImageNet EfficientNet-B0 probe. Protocol P1 produced 100.0% on both datasets, and every test image had a rotated or mirrored copy of itself in training. With the images as distributed (P2), accuracy was 99.7% on A and 99.7% on B, while 74% and 92% of test leaves were also present in training. Grouping by leaf (P3) removed that overlap and lowered accuracy to 99.0% and 97.6%. The drop on A is small, probably because its three classes look very different; on B the fold-to-fold spread grew from 0.4 to 1.9 points.

ELA hurt in every grouped comparison. On Dataset B, ELA accuracy fell from 98.0% under P2 to 87.4% under P3, a loss of 10.6 points, against 2.1 points for RGB. Much of what ELA appears to learn is therefore tied to individual photographs rather than to disease.

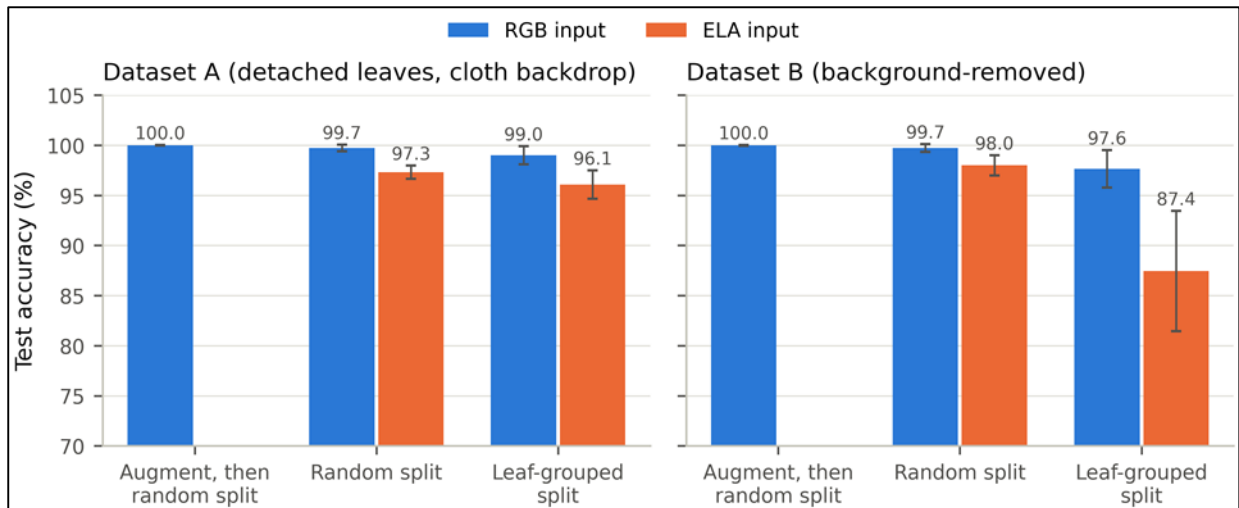


Fig 4: Test accuracy of the ImageNet EfficientNet-B0 linear probe under the three split protocols, with RGB and ELA inputs. Error bars are standard deviations over 25 folds (P2, P3) or five repeated splits (P1). ELA was not evaluated under P1.

Table 2. Accuracy (%) by Split Protocol

Dataset, input	P1 augment then split	P2 random	P3 leaf-grouped	Test leaves seen in training, P2 / P3 (%)
A, RGB	100.0 ± 0.0	99.7 ± 0.3	99.0 ± 0.9	74 / 0
A, ELA	not run	97.3 ± 0.7	96.1 ± 1.4	74 / 0
B, RGB	100.0 ± 0.0	99.7 ± 0.4	97.6 ± 1.9	92 / 0
B, ELA	not run	98.0 ± 1.0	87.4 ± 6.0	92 / 0

Values are mean ± standard deviation. Under P1 every test image had a transformed copy in training.

C. Transfer with All Training Leaves

Table 3 compares E1 and E2 with all training leaves available. Under the leaf-grouped protocol, macro F1 on Dataset A was 99.1% for E1 and 99.2% for E2. On Dataset B the two scored 97.1% and 97.1%, and on both datasets the paired differences stayed below half a point. None was significant. The 512-unit projection of E2 did slightly worse than its pooled feature (98.4% and 94.9% under P3), probably because the projection was shaped by the source labels.

Table 3. Macro F1 (%) of Frozen Efficientnet-B3 Encoders with all Training Data

Dataset, split	E1 ImageNet	E2 multi-crop	E2 minus E1 (points)	Wilcoxon
A, random	99.7 ± 0.3	99.7 ± 0.3	-0.00	p = 0.958
A, leaf-grouped	99.1 ± 0.7	99.2 ± 0.7	+0.12	p = 0.520
B, random	99.9 ± 0.3	99.6 ± 0.5	-0.24	p = 0.057
B, leaf-grouped	97.1 ± 3.6	97.1 ± 2.1	-0.03	p = 0.475

Mean ± standard deviation over 25 folds (five seeds × five folds). The test compares paired fold scores.

D. Transfer with Few Training Leaves

Scarce labels change the picture (Fig. 5, Table 4). With one training leaf per class on Dataset A, macro F1 was 0.507 for E1 and 0.728 for E2. E2 stayed ahead at every k. With five leaves per class the scores were 0.838 and 0.948, and with ten they were 0.913 and 0.973, with $p < 0.001$ in the paired test at each k. In practical terms, E2 given two leaves per class (0.865) did slightly better than E1 given five (0.838).

On Dataset B the gain depended on k. E2 was ahead with one leaf per class (0.712 against 0.653, $p = 0.007$) and with ten (0.950 against 0.936, $p = 0.003$), but the two encoders were indistinguishable at two, three and five leaves. Dataset B has fewer distinct leaves, and its healthy and diseased leaves probably differ in colour and size in ways an ImageNet network already encodes, which would leave less room for disease-specific features to help.

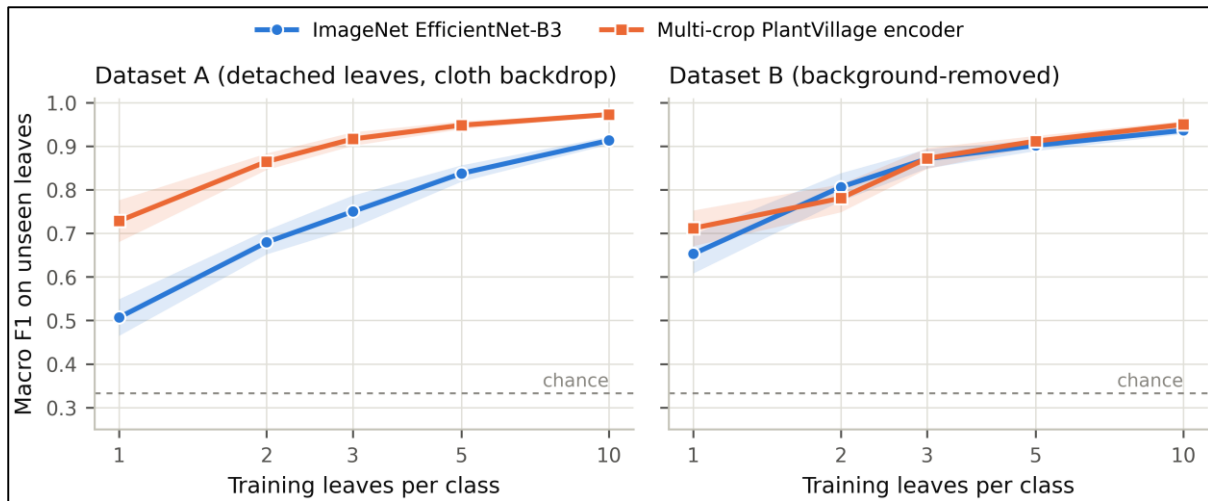


Fig 5: Macro F1 on unseen leaves when k leaves per class are available for training. Lines are means over 30 paired draws and shaded bands are 95% confidence intervals of the mean. The dashed line marks chance for three balanced classes.

Table 4. Few-Shot Macro F1 on Unseen Leaves

Leaves per class	A: E1	A: E2	A: Wilcoxon	B: E1	B: E2	B: Wilcoxon
1	0.507	0.728	$p < 0.001$	0.653	0.712	$p = 0.007$
2	0.680	0.865	$p < 0.001$	0.806	0.781	$p = 0.158$
3	0.750	0.917	$p < 0.001$	0.871	0.872	$p = 0.730$
5	0.838	0.948	$p < 0.001$	0.902	0.912	$p = 0.152$
10	0.913	0.973	$p < 0.001$	0.936	0.950	$p = 0.003$

Means over 30 random draws of training leaves; both encoders saw identical draws.

E. Using the Source Classifier Without Black Pepper Training

PlantVillage contains no black pepper, so the crop head of E2 could not give a correct answer. Its answers clustered on a few crops. On Dataset A it named bell pepper for 24% of images and blueberry for 24%; on Dataset B it named tomato for 47% and blueberry for 31% (Fig. 6). Confidence stayed high. The mean maximum softmax probability was 0.95 on A and 0.93 on B, and 85% of Dataset A images received a probability above 0.9.

Healthy-versus-diseased predictions carried over to Dataset B but not to Dataset A. On Dataset B the rule reached a balanced accuracy of 0.939, with sensitivity 0.885 for diseased leaves and specificity 0.993 for healthy ones, far above the 0.5 of a constant answer. Dataset A contains only diseased leaves. There, 45% of images were called healthy. Cut-out leaves on a plain backdrop look much more like PlantVillage photographs than leaves on cloth do, which is the likely reason the source head worked better on them.

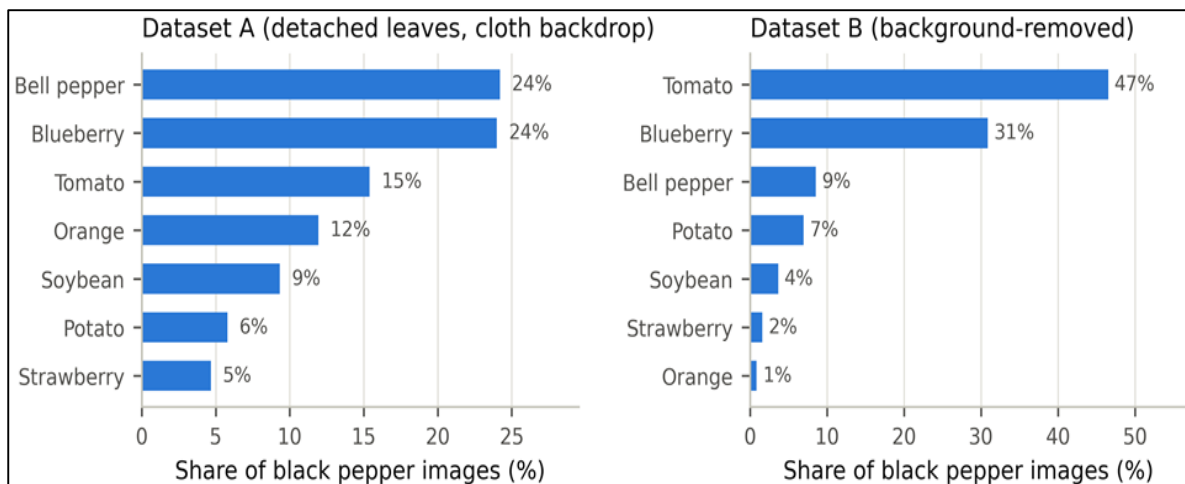


Fig 6: Crop predicted by the multi-crop source model for black pepper images, which it never saw during training. Only the seven most frequent answers are shown.

Table 5. Zero-Shot Behaviour of the Multi-Crop Source Model

Measure	Dataset A	Dataset B
Most frequent predicted crop	Bell pepper (24%)	Tomato (47%)
Mean maximum softmax probability	0.95	0.93
Images with probability above 0.9 (%)	85	79
Diseased leaves called healthy (%)	45	12
Healthy vs diseased, balanced accuracy	not defined (no healthy class)	0.939

F. Accuracy Obtained Without Diseased Tissue

Fig. 7 and Table 6 report the control inputs. On Dataset A, the four image corners, with every leaf pixel painted out, predicted the disease class with 87.1% accuracy under the leaf-grouped protocol. The 95th percentile of the group-permutation null was 38.5%. The binary silhouette alone reached 72.9%, and the random forest on file, backdrop and leaf-geometry features reached 81.2%. The three classes of Dataset A were apparently photographed against backdrops and lighting that differ systematically, so the dataset rewards a model for recognising the photo session.

Dataset B has a single synthetic grey backdrop. Its corner input reached 51.2% against a null 95th percentile of 45.4%, and that small signal comes from large leaves reaching into the corners rather than from the scene. The silhouette did better, at 76.1%, since yellow mottle leaves in this set tend to be small and distorted. The random forest reached 86.5%. That figure is not a pure shortcut. Without the geometry features it fell to 60.0% on file properties and 58.6% on ELA statistics, and leaf size is partly a symptom. On both datasets the leaf-only input scored close to the full image, so the leaf carries most of the usable signal; on Dataset A, though, a second route to the same answer exists that never looks at the leaf.

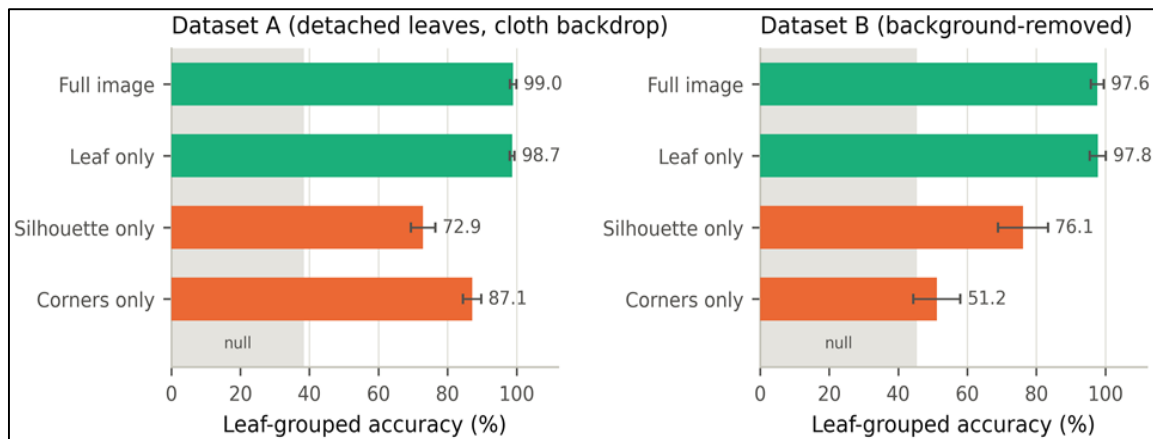


Fig. 7. Leaf-grouped accuracy of the ImageNet EfficientNet-B0 probe on control inputs. Green bars contain leaf tissue; orange bars contain none. The grey band marks the 95th percentile of the group-permutation null for the corner input.

Table 6. Leaf-Grouped Accuracy (%) on Control Inputs

Input	Dataset A	Dataset B
Full image	99.0 ± 0.9	97.6 ± 1.9
Leaf only	98.7 ± 0.7	97.8 ± 2.3
Silhouette only	72.9 ± 3.6	76.1 ± 7.2
Corners only	87.1 ± 2.6	51.2 ± 6.8
ELA image	96.1 ± 1.4	87.4 ± 6.0
Random forest on file, backdrop and leaf-geometry features	81.2 ± 2.0	86.5 ± 5.0
Group-permutation null, 95th percentile (corners)	38.5	45.4

Chance is 33.3% for three balanced classes. Mean ± standard deviation over 25 leaf-grouped folds.

V. DISCUSSION

For black pepper the practical news is good on one point. A network that has already learned what lesions, chlorosis and necrosis look like on fourteen other crops needs far fewer black pepper leaves to reach a given accuracy, at least on Dataset A. That matters to any group that can afford laboratory confirmation for only a few leaves per disease. Starting from a plant disease model instead of ImageNet costs nothing extra. The gain shrank as leaves were added and was absent when every training leaf was used, in line with earlier findings that the choice of pretraining matters most for small target sets [25], [26].

The second message concerns the benchmarks themselves. Near-perfect accuracy on them is easy to obtain and says little about disease recognition. On these data, augmenting before splitting produced 100% by construction. Splitting at the image level, when each leaf was photographed several times, inflates it. And on Dataset A a model that never sees the leaf still reaches close to 90%, because each class carries its own backdrop and lighting. A reported accuracy for black pepper leaf disease should therefore state how images were grouped before splitting, and ideally report a background-only control beside the main result. The same applies to private collections such as those in [7] and [8]; their published descriptions do not say whether repeated photographs of a leaf were kept together, and the images are available only on request.

ELA needs a separate comment, because it is occasionally proposed as a preprocessing step for CNN classifiers [39]. It measures how an image responds to JPEG re-compression, and that response depends on the camera, on earlier saves and on the file format [38], none of which has anything to do with disease. In Dataset A the PNG files all belong to one class. In Dataset B the only real JPEG files do. ELA therefore has differences of exactly that kind to exploit, and its grouped accuracy was below that of plain RGB on both datasets, with the largest loss of any input when the split moved from random to grouped. We see no reason to use it for disease classification.

The zero-shot results argue for caution before a large plant disease model is reused as a ready-made diagnostic tool. A classifier trained on a closed set of crops cannot say that a leaf belongs to none of them, and here it gave confident wrong answers instead. Its notion of “healthy” carried over to cut-out leaves on grey. It failed for leaves on cloth. Frozen features from the model were useful for training a new classifier, while its original labels and confidence scores were not.

A. Limitations

The two datasets are small, come from single uploaders and lack acquisition records, so the leaf groups are inferred rather than known; a group may occasionally join two similar leaves, which makes the grouped estimates slightly conservative, and may miss some repeats that were not among the 20 nearest neighbours. The class labels were not confirmed by a plant pathologist for this study. The multi-crop encoder is a single model trained on laboratory photographs, and a model trained on field images might transfer differently. Encoders were compared frozen; full fine-tuning could narrow or widen the few-shot gap. Most importantly, neither dataset contains leaves photographed on the vine, so none of the accuracies here should be read as field performance.

VI. Conclusion

We tested whether a CNN trained on many crops helps classify black pepper leaf diseases, using both public datasets and splits that keep every photograph of a leaf on one side. With plenty of training leaves, multi-crop and ImageNet pretraining performed the same. With one to ten leaves per class, the multi-crop encoder was better at every step on Dataset A, and on Dataset B at one and at ten leaves. The audit also showed that augmenting before splitting yields 100% accuracy by construction and that ELA lowers grouped accuracy. On Dataset A, background corners alone predicted the disease class with 87.1% accuracy. Future black pepper datasets should record the leaf, plant, farm and device behind each photograph and include healthy leaves imaged in the same way as diseased ones, so that grouped evaluation can become routine.

A. Data Availability

Both datasets are publicly available on Kaggle at the addresses given in [32] and [22]. The leaf-group assignments, fold-level results and analysis scripts are available from the corresponding author on request.

REFERENCES

- [1] Jeevalatha, F. Zumaila, C. N. Biju, and M. F. Peeran, "Multiplex PCR assay for simultaneous detection of Phytophthora, Pythium and Fusarium associated with foot rot and yellowing diseases of black pepper," *J. Spices Aromat. Crops*, vol. 31, no. 1, pp. 92–96, 2022, doi: 10.25081/josac.2022.v31.i1.7739.
- [2] N. Biju, R. Praveena, S. J. Ankegowda, C. N. Darshana, and K. C. Jashmi, "Epidemiological studies of black pepper anthracnose (*Colletotrichum gloeosporioides*)," *Indian J. Agric. Sci.*, vol. 83, no. 11, 2013.

- [3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, Art. no. 1419, 2016, doi: 10.3389/fpls.2016.01419.
- [4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, 2018, doi: 10.1016/j.compag.2018.01.009.
- [5] Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agric.*, vol. 147, pp. 70–90, 2018, doi: 10.1016/j.compag.2018.02.016.
- [6] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: A review," *Plant Methods*, vol. 17, no. 1, Art. no. 22, 2021, doi: 10.1186/s13007-021-00722-9.
- [7] S. Kini, K. V. Prema, and S. N. Pai, "Early stage black pepper leaf disease prediction based on transfer learning using ConvNets," *Sci. Rep.*, vol. 14, no. 1, Art. no. 1404, 2024, doi: 10.1038/s41598-024-51884-0.
- [8] P. T. Sreethu, M. M. Paul, P. P. Gopinath, I. L. Shahana, and N. S. Radhika, "Foliar symptom-based disease detection in black pepper using convolutional neural network," *Phytopathol. Res.*, vol. 7, no. 1, Art. no. 21, 2025, doi: 10.1186/s42483-024-00305-1.
- [9] T. R. Swathi Chandra and V. Parthasarathy, "Enhanced leaf disease diagnosis in black-pepper plants using a dense back-projection joint attention network," *J. Phytopathol.*, vol. 174, no. 1, Art. no. e70223, 2026, doi: 10.1111/jph.70223.
- [10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
- [11] Barz and J. Denzler, "Do we train on test data? Purging CIFAR of near-duplicates," *J. Imaging*, vol. 6, no. 6, Art. no. 41, 2020, doi: 10.3390/jimaging6060041.
- [12] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.
- [13] Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 1521–1528, doi: 10.1109/CVPR.2011.5995347.
- [14] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, Art. no. e1002683, 2018, doi: 10.1371/journal.pmed.1002683.
- [15] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Comput. Electron. Agric.*, vol. 153, pp. 46–53, 2018, doi: 10.1016/j.compag.2018.08.013.
- [16] J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," *Biosyst. Eng.*, vol. 172, pp. 84–91, 2018, doi: 10.1016/j.biosystemseng.2018.05.013.
- [17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
- [18] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2656–2666, doi: 10.1109/CVPR.2019.00277.
- [19] Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," *Front. Plant Sci.*, vol. 8, Art. no. 1852, 2017, doi: 10.3389/fpls.2017.01852.
- [20] Argüeso, A. Picon, U. Irusta, A. Medela, M. G. San-Emeterio, A. Bereciartua, and A. Alvarez-Gila, "Few-shot learning approach for plant disease classification using images taken in the field," *Comput. Electron. Agric.*, vol. 175, Art. no. 105542, 2020, doi: 10.1016/j.compag.2020.105542.
- [21] Y. Khew, Y. Q. Teow, E. T. Lau, S. S. Hwang, C. H. Bong, and N. K. Lee, "Evaluation of deep learning for image-based black pepper disease and nutrient deficiency classification," in *Proc. 2nd Int. Conf. Artif. Intell. Data Sci. (AiDAS)*, 2021, pp. 1–6, doi: 10.1109/AiDAS53897.2021.9574346.
- [22] udi17live, "Black pepper leaf blight and yellow mottle virus," Kaggle, 2022. [Online]. Available: <https://www.kaggle.com/datasets/udi17live/black-pepper-leaf-blight-and-yellow-mottle-virus>
- [23] C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Comput. Electron. Agric.*, vol. 161, pp. 272–279, 2019, doi: 10.1016/j.compag.2018.03.032.
- [24] M. H. Saleem, J. Potgieter, and K. M. Arif, "Plant disease classification: A comparative evaluation of convolutional neural networks and deep learning optimizers," *Plants*, vol. 9, no. 10, Art. no. 1319, 2020, doi: 10.3390/plants9101319.
- [25] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 4917–4926, doi: 10.1109/ICCV.2019.00502.
- [26] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?," in *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science)*. Cham, Switzerland: Springer, 2020, pp. 266–282, doi: 10.1007/978-3-030-58568-6_16.
- [27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
- [28] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Aroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
- [29] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proc. ACM Conf. Health, Inference, Learn. (CHIL)*, 2020, pp. 151–159, doi: 10.1145/3368555.3384468.
- [30] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Solving current limitations of deep learning based approaches for plant disease detection," *Symmetry*, vol. 11, no. 7, Art. no. 939, 2019, doi: 10.3390/sym11070939.
- [31] J. G. A. Barbedo, "Plant disease identification from individual lesions and spots using deep learning," *Biosyst. Eng.*, vol. 180, pp. 96–107, 2019, doi: 10.1016/j.biosystemseng.2019.02.002.

- [32] adithyantg, "Black pepper leaf disease mini-dataset," Kaggle, 2025. [Online]. Available: <https://www.kaggle.com/datasets/adithyantg/black-pepper-leaf-disease-mini-dataset>
- [33] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), ser. Proc. Mach. Learn. Res., vol. 97, 2019, pp. 6105–6114.
- [34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
- [35] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981, doi: 10.1145/358669.358692.
- [36] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), ser. Proc. Mach. Learn. Res., vol. 37, 2015, pp. 1180–1189.
- [37] Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
- [38] Farid, "Exposing digital forgeries from JPEG ghosts," IEEE Trans. Inf. Forensics Secur., vol. 4, no. 1, pp. 154–160, 2009, doi: 10.1109/TIFS.2008.2012215.
- [39] B. K. Sudiatmika, F. Rahman, Trisno, and Suyoto, "Image forgery detection using error level analysis and deep learning," TELKOMNIKA, vol. 17, no. 2, p. 653, 2019, doi: 10.12928/telkomnika.v17i2.8976.