Publishing Excellence Inspiring Progress
ARISTON PUBLICATIONS
BioMed and BioSci Advances
ISSN: 3068-7314 www.aristonpubs.com/biomed-and-biosci-advances https://doi.org/10.69626/bba.2026.0202

REVIEW ARTICLE

Grape Leaf Disease Identification Using Image Retrieval and Expert Referral of Doubtful Images

Bhagawan N. Kadlag*, Anil G. Khairnar, Dhananjay N. Kadlag, Chandrakant G. Dighavkar, Arun V. Patil

ABSTRACT: Image-based tools for plant disease identification are typically evaluated on images that resemble their training data, and they produce a diagnosis for every image regardless of whether it resembles anything they have seen before. This paper describes diagnose-or-defer, a content-based image retrieval method for grape leaves that either reports the disease indicated by the most similar labelled reference images or refers the image to an expert. The leaf and its lesions are segmented, and each image is represented by how its lesions differ in color from the healthy tissue of the same leaf, by leaf texture, and by lesion shape. The ten most similar reference images are retrieved, and the method defers when too few of them agree or when the query lies farther from its nearest reference image than 95% of calibration images do. In acquisition-blocked cross-validation on 4,046 Plant-Village images covering black rot, esca, leaf blight, and healthy leaves, accuracy without deferral was 89.0%. With deferral, 60.6% of images received a diagnosis, 97.9% of those diagnoses were correct, and 1.3% of all images received a wrong diagnosis, compared with 11.0% without deferral. On 133 internet images from Plant-Doc, retrieval alone was wrong for 64%; the method referred 96% of them and gave a wrong diagnosis for 1.2%. The method is suited to photographs taken under conditions similar to its reference images, and it recognizes most photographs that are not. Lesion color is measured relative to the healthy tissue of the same leaf, and every diagnosis is presented together with the reference images on which it rests. A two-part rule, calibrated on held-out images, defers to an expert when the retrieved images disagree or when the photograph is unfamiliar. Validation employed acquisition-blocked splits, a separate internet test set, and controlled image changes, and all parameters and code are openly available.

Keywords: Content-based image retrieval; Plant disease diagnosis; Reject option; Selective classification; Out-of-distribution detection; Vitis vinifera

Received: 15 January 2026; Revised: 28 April 2026; Accepted: 08 August 2026; Published Online: 26 August 2026

1. INTRODUCTION

Deep learning has become the dominant approach to image-based plant disease identification, and the reported accuracy figures are often striking. A convolutional neural network trained on the Plant-Village collection reached 99.35% accuracy on held-out Plant-Village images [1]. That figure, however, describes performance on images drawn from the same controlled acquisition conditions as the training data. When the same models were tested on images taken under other conditions, accuracy fell to just above 31%, and models trained on Plant-Village reached only 15% on the internet images of the Plant-Doc benchmark [1]. These gaps are not incidental. Benchmark images carry acquisition signatures that a classifier can learn and exploit, as demonstrated by a model that achieved 49% accuracy using only eight background pixels from Plant-Village images [2]. Training sets rarely cover the variety of field conditions that a deployed system will encounter [3], and random splits of non-independent images, such as several photographs of one leaf, overstate accuracy by allowing near-duplicates to appear in both training and test partitions [4].

Research Center in Electronic Science, MGV's L. V. H. Arts, Science and Commerce College, Panchavati, Nashik 422003, Maharashtra, India.

* Author to whom correspondence should be addressed: kadlag1985@gmail.com (B. N. Kadlag)

© Ariston Publications 2026. All rights reserved.

A further limitation receives less attention in the literature. A conventional classifier returns a diagnosis for every image it receives, including images unlike any it was built from, and the user has no way to distinguish an informed answer from a guess. This is a consequential design flaw in a diagnostic context, because an incorrect diagnosis presented with the same confidence as a correct one can lead to inappropriate treatment decisions. The problem is not merely one of accuracy on a benchmark; it concerns the conditions under which a system should be trusted to speak at all.

Content-based image retrieval offers a principled alternative in this setting [5]. Its answer comes with its evidence: the labelled reference images most similar to the query, which the user can inspect and compare directly with the photograph in hand [6]. The reference database can grow with expert-confirmed images without retraining, allowing the system to adapt to local conditions through accumulated experience. Nearest-neighbor classification, the decision rule underlying retrieval-based diagnosis, also lends itself naturally to a reject option [7]. The method can decline to diagnose when the retrieved images disagree among themselves, or when even the nearest of them lies far from the query, indicating that the photograph is unlike anything in the reference collection.

The novelty of the present work lies in the combination of three elements that have not previously been integrated in a grape leaf diagnostic system. First, lesions are described by how their color differs from the healthy tissue of the same leaf, rather than by absolute color values. The rationale is that a change of illumination or camera, which shifts all colors in a photograph, alters these within-leaf differences less than it alters the colors themselves. Second, the method incorporates an explicit deferral mechanism with two calibrated checks, one for disagreement among retrieved images and one for unfamiliarity of the query relative to the reference database, so that the system states when it cannot answer rather than guessing. Third, the entire method is validated under acquisition-blocked splits that keep images of the same leaf out of the database when they are tested, and it is evaluated on a separate internet collection and under controlled image changes, providing a more realistic assessment than random splits allow. The authors previously applied content-based retrieval to grape and pomegranate disease detection in MATLAB [8]; the present implementation is new, open, and built around these three design commitments. Segmentation settings were chosen by visual inspection of masks on a few dozen Plant-Village images before any accuracy was computed, and no parameter was tuned on accuracy or on the internet images.

In this paper, we report the development and validation of diagnose-or-defer, a content-based image retrieval method for grape leaf disease identification that refers doubtful images to an expert. The method segments the leaf and its lesions, computes a 56-value descriptor capturing lesion color contrast, color change, leaf texture, and lesion shape, retrieves the ten most similar reference images, and applies two calibrated checks before issuing a diagnosis. We evaluate the method on 4,046 Plant-Village images under acquisition-blocked cross-validation, on 133 internet images from Plant-Doc, and under six controlled image changes, comparing it with an absolute-color descriptor and a global-histogram baseline. We report accuracy without deferral, coverage, the accuracy of diagnoses given, the share of all images receiving a wrong diagnosis, and precision at ten, with confidence intervals and McNemar tests for descriptor comparisons. The results show that deferral reduces wrong diagnoses from 11.0 % to 1.3% on internal images and from 64% to 1.2% on internet images, while maintaining high accuracy among the images that receive a diagnosis. We discuss the conditions under which the method is suitable, its limitations on field images and images with complex backgrounds, and the implications of calibrated deferral for the practical deployment of image-based diagnostic tools.

2. MATERIALS AND METHODS

2.1. Image Datasets

Reference and internal test images were drawn from the color version of the Plant-Village collection [1, 2], used as distributed with its published leaf map [3]. The collection comprised 4,062 grape leaf images distributed across four classes: black rot with 1,180 images, esca or black measles with 1,383, leaf blight or Isariopsis leaf spot with 1,076, and healthy leaves with 423. In the three disease classes each leaf was photographed four times, with one leaf photographed three times, and the images of a given leaf carry consecutive acquisition numbers. The healthy images have no leaf map and originate from a separate acquisition series. The external test set comprised all 133 grape images of Cropped-Plant-Doc [4, 5], which are leaf crops extracted from internet photographs, with 69 images labelled healthy and 64 labelled black rot. The file names of six of the black rot images refer to another disorder, namely downy mildew, anthracnose, Phomopsis, rust, or spray damage; the published labels were retained unchanged to preserve the integrity of the benchmark. A 64-bit perceptual hash based on the discrete cosine transform was used to search for copies of Plant-Village images among the Plant-Doc images. The smallest Hamming distance between the two sets was 10 bits, and visual inspection of the ten closest pairs found no duplicates. Within Plant-Doc, two pairs of near-identical images with distances of 0 and 6 bits were identified and kept in the same fold during the local-image experiment.

2.2. Image Preparation and Segmentation

Images were converted to RGB, the longest side was resized to 256 pixels using bilinear interpolation with anti-aliasing, and the image was centered on a black 256 by 256 canvas whose padding was excluded from all subsequent steps. A 3 by 3 median filter was applied to each channel, and colors

© Ariston Publications 2026. All rights reserved.

were converted to CIELAB with a D65 white point [19]. The leaf was separated from the background by Otsu's threshold [20] on the CIELAB chroma $C^*_{ab} = \sqrt{a^{*2} + b^{*2}}$, computed over pixels whose HSV value exceeds 0.12., computed over pixels whose HSV value exceeded 0.12. The mask was cleaned by morphological closing and opening with a disk of radius 2 pixels, holes were filled, and the largest connected component was retained; if this covered less than 5% of the image, all pixels above the value limit were used instead. A 3-pixel rim was then removed because pixels at the leaf margin mix leaf and background colors.

Lesions were identified by color clustering within the leaf. K-means clustering [21] with K equal to 3, five initializations, and a fixed seed was applied to the a* and b* values of up to 5,000 randomly sampled leaf pixels, and every leaf pixel was assigned to its nearest cluster center. The cluster with the lowest a, the greenest, was taken as healthy tissue. Clusters whose center lay more than 15 units from the healthy center in the ab* plane were taken as lesion tissue; on a healthy leaf all three centers usually lie within this distance, and no lesion was found. The lesion mask was opened with a disk of radius 1 pixel, and components smaller than 10 pixels were removed.

2.3. Descriptor Construction

The descriptor has four blocks (Table 1). The descriptor comprised four blocks totaling 56 values. Block A, color contrast, contained six values: the mean L, a, and b* of the lesion pixels minus those of the healthy pixels of the same leaf, together with the logarithms of the ratios of their standard deviations. Block B, color change, contained 24 values describing the vector from the mean healthy color to the color of each lesion pixel; a 16-bin histogram of its direction in the a*b* plane recorded the hue of the change, and an 8-bin histogram of its length, the CIE76 color difference from 0 to 80, recorded its magnitude. Blocks A and B were set to zero when no lesion was found. Block C, texture, contained 20 values computed from leaf lightness L, rescaled between its 1st and 99th percentiles within the leaf and quantized to 16 levels; five Haralick statistics [22], namely angular second moment, contrast, correlation, inverse difference moment, and entropy, were computed from grey-level co-occurrence matrices at distances of 1 and 3 pixels and averaged over four directions, and a 10-bin histogram of rotation-invariant uniform local binary patterns with 8 neighbors and radius 1 [23] completed the block. Block D, lesion shape, contained six values: the lesion area as a fraction of leaf area, the logarithm of one plus the number of lesions, the mean lesion area as a fraction of leaf area, and the mean eccentricity, solidity, and roughness of the lesions, where roughness was perimeter squared divided by four pi times area. Two comparison descriptors were evaluated with identical retrieval and decision steps: an absolute-color descriptor replacing blocks A and B with the mean and standard deviation of lesion L, a, and b, a 16-bin lesion hue histogram, and an 8-bin lesion L* histogram, and a global-histogram descriptor using 16 hue, 8 saturation, and 8 value bins of the whole image without segmentation.

2.4. Retrieval and Deferral

Each feature was standardized with the mean and standard deviation of the reference database and weighted by the square root of the size of its block, so that every block contributed equally to the distance on average. The Euclidean distance between the query and every reference image was computed, and the ten nearest reference images were retrieved. The method took the class shown by most of them, breaking a tie in favor of the tied class whose image was nearest, and recorded the agreement as the number of the ten images showing that class. Two checks decided the output. The image was deemed unfamiliar if its distance to the nearest reference image exceeded the 95th percentile of the same distance among calibration images held out from the database; by construction, about 5% of images like the calibration images were deferred as unfamiliar. The retrieved images were deemed to disagree if the agreement fell below a minimum value, defined as the smallest agreement for which no more than 2% of the calibration images that would be diagnosed were diagnosed wrongly; if no value met this target, the minimum was set to 10. The output was either a diagnosis or a referral, and a referral stated which check had failed. New expert-confirmed images could be added at any time by standardizing their descriptors with the existing statistics and appending them to the database, so that distances between existing images and calibrated thresholds remained unchanged and no retraining was required. The retrieved images disagree if m is below $m_{\min}$, the smallest agreement for which no more than 2% of the calibration images that would be diagnosed are diagnosed wrongly; if no value meets this target, $m_{\min} = 10$. The output is:

$$\text{output} = \begin{cases} \hat{c} & \text{if } d_1 \le d_{95} \text{ and } m \ge m_{\min}, \\ \text{refer to an expert} & \text{otherwise,} \end{cases}$$

And a referral states which check failed. The 2% target was set before evaluation and can be changed by the user.

2.5. Validation Design and Metrics

For the internal test, each class was ordered by acquisition number and cut into five contiguous blocks of near-equal size, so that images of the same leaf, and leaves photographed close together in time, were kept out of the database when tested [4].

© Ariston Publications 2026. All rights reserved.

Table 1. Fixed parameters of diagnose-or-defer.

Step	Setting	Value
Image preparation	Size	Longest side 256 px, padded to 256 × 256
	Noise filter	3 × 3 median, each RGB channel
Leaf mask	Threshold	Otsu on CIELAB chroma; pixels with HSV value > 0.12
	Clean-up	Closing and opening (disk, r = 2 px), hole filling, largest component, 3-px rim removed
Lesion mask	Clustering	K-means, K = 3, on a*, b* of ≤ 5,000 leaf pixels; 5 initialisations; seed 0
	Healthy and lesion tissue	Healthy: cluster with lowest a*; lesion: clusters > 15 a*b* units from it
	Clean-up	Opening (disk, r = 1 px); components < 10 px removed
Descriptor (56 values)	A Colour contrast (6)	Lesion minus healthy mean of L*, a*, b*; log ratios of their SDs
	B Colour change (24)	Direction of change in the a*b* plane (16 bins); ΔE^*_{ab} to healthy mean (8 bins, 0–80)
	C Texture (20)	GLCM of L* (1st–99th percentile, 16 levels; distances 1 and 3 px; 4 directions averaged): ASM, contrast, correlation, IDM, entropy; uniform LBP (P = 8, R = 1), 10 bins
	D Lesion shape (6)	Area fraction, log(1 + count), mean area, eccentricity, solidity, roughness
Retrieval	Scaling and distance	z-score on the database; block weight 1/√(block size); Euclidean
	Retrieved images	k = 10
Diagnosis	Class and agreement	Majority class of the 10 (ties: nearest tied image); m = images showing it
Deferral	Unfamiliar	Nearest-image distance > 95th percentile of calibration images
	Disagreement	$m < m_{min}$; m_{min} = smallest m with ≤ 2% error on calibration images (else 10)

ASM, angular second moment; GLCM, grey-level co-occurrence matrix; IDM, inverse difference moment; LBP, local binary pattern; SD, standard deviation.

In the disease classes each cut was moved to the next leaf boundary given by the leaf map; in the healthy class, which has no leaf map, the four images following each cut were excluded, assuming the four-image pattern of the other classes. In round r from 1 to 5, block r served as the test set, the next block, with block 1 following block 5, served as the calibration set, and the other three blocks formed the reference database of 2,421 to 2,434 images, so that each of the 4,046 remaining images was tested once. For the internet images, all 133 Plant-Doc images were diagnosed in every round with that round's database and thresholds, and results were averaged over the five rounds, with a diagnosis of esca or leaf blight counted as wrong. To test whether local images help, the Plant-Doc images were split into five folds stratified by class, with near-identical images and images from the same file-name series kept in one fold; in round r, the images of the other four folds, numbering 106 to 107, were added to the database without changing any threshold, and fold r was tested. To test the expectation behind blocks A and B, six fixed changes were applied to every internal test image before description, with that round's database and thresholds unchanged: darker and brighter exposure with RGB values multiplied by 0.7 and 1.3, a warm and a cool color cast with red multiplied by 1.15 and blue by 0.85 and the reverse,

© Ariston Publications 2026. All rights reserved.

Gaussian blur with a sigma of 1.5 pixels, and JPEG compression at quality 20, in the spirit of common corruption benchmarks [24]. For each test we report accuracy without deferral, the share of images given a diagnosis, the accuracy of the diagnoses given, wrong diagnoses as a share of all images, and precision at 10. Proportions pooled over images are given with Wilson 95% confidence intervals [25], and descriptors were compared on the same images with exact McNemar tests [26].

2.6. Implementation

The method was implemented in Python 3.13 with NumPy [27], SciPy [28], scikit-image [29], and scikit-learn [30], and package versions are listed in the code repository. On one core of a 2.1 GHz Intel Xeon processor, describing an image took a median of 0.13 seconds for Plant-Village images and 0.24 seconds for the larger Plant-Doc images, and retrieval with the decision took 0.5 milliseconds with a database of about 2,400 images. The code was written, and the validation run, with the assistance of a generative AI system under the authors' direction.

4. RESULTS AND DISCUSSION

3.1. Internal Test on Plant-Village Images

The proposed descriptor identified 89.0% of the internal test images without deferral, with a 95% confidence interval of 87.9 to 89.9%, compared with 86.7% for the absolute-color descriptor and 69.6% for the global histogram. Exact McNemar tests on the same images gave P below 0.001 for both comparisons, indicating that the advantage of the proposed descriptor was not attributable to chance. Confusions between black rot and esca accounted for 58% of the errors, which is consistent with the visual similarity of the two disorders at the lesion level: both produce dark, necrotic patches on the leaf, and the distinguishing cues are often the pattern of lesion distribution and the presence of marginal necrosis rather than lesion color alone. The global histogram performed poorly because it discards all spatial and tissue-relative information, and its errors were dominated by misclassification of healthy leaves as diseased and vice versa, since the background and illumination of a photograph contribute strongly to the whole-image color distribution.

With deferral, 60.6% of the images received a diagnosis, and 97.9% of these diagnoses were correct, with a confidence interval of 97.2 to 98.4%. The share of all images given a wrong diagnosis fell from 11.0% to 1.3%, with a confidence interval of 1.0 to 1.7%, and 88.4% of the images that would have been misdiagnosed were referred instead. Most referrals were due to disagreement among the retrieved images, which accounted for 33.8% of images, while 5.6% of images were unfamiliar, close to the nominal 5% expected from the calibration rule. The calibrated agreement threshold was 10 in three rounds, 9 in one round, and 8 in one round, and in one round no threshold met the 2% calibration target. This variation is informative rather than problematic: it shows that the appropriate level of agreement depends on the composition of the reference database and the calibration set, and it supports the recommendation that users calibrate the method on images from their own setting rather than adopting a single fixed threshold. Coverage ranged from 46.3 % for black rot to 72.4% for leaf blight, and the accuracy of the diagnoses given ranged from 96.2% for black rot to 99.6% for healthy leaves. The lower coverage for black rot reflects the difficulty of separating it from esca, and the lower accuracy among diagnosed black rot images reflects the residual overlap that survives the agreement check. Healthy leaves, by contrast, are distinctive in the descriptor space and are diagnosed with high accuracy when a diagnosis is given.

Table 2 summarizes the internal, internet, and local-image results for all three descriptors. For the proposed descriptor on internal images, precision at 10, the share of the ten retrieved images that show the query's class, was 0.83, compared with 0.81 for absolute color and 0.66 for the global histogram. Precision at 10 is a useful complement to classification accuracy because it measures the quality of the evidence presented to the user, not only the correctness of the final label. A value of 0.83 means that, on average, more than eight of the ten displayed reference images are relevant to the query, which supports the interpretability claim that motivates the retrieval approach.

Figure 1 presents the three representative cases that illustrate the decision logic of the method. In panel (a), a Plant-Village black rot leaf retrieved ten reference images that all showed black rot, and the image lay well within the calibrated distance limit, so a diagnosis was given with the ten supporting images displayed alongside it. In panel (b), a Plant-Village esca leaf retrieved a mixture of esca and black rot references, with only five of the ten showing esca, which fell below the calibrated agreement threshold for that round, so the image was referred to an expert with the reason stated. In panel (c), a Plant-Doc internet image labelled black rot retrieved seven black rot references, which would have satisfied the agreement check, but its distance to the nearest reference image was 1.5 times the calibrated limit, so it was referred as unfamiliar. The three panels together show that the two checks are complementary: the agreement check catches ambiguous queries that resemble several classes, while the familiarity check catches queries that resemble no class closely, and a referral states which condition triggered it. Figure 2(a) plots the error among diagnosed images against the share of images diagnosed in the internal test, as the agreement threshold rises from 1 to 10 with the familiarity check applied.

 © Ariston Publications 2026. All rights reserved.

Table 2. Results on the internal test, on the internet images, and on the internet images after local images were added to the database.

Test set	Descriptor	Correct without deferral (%)	Diagnosed (%)	Correct among diagnosed	Wrong diagnoses (% of images)	Precision at 10
Internal (PlantVillage, n = 4,046)	Proposed	89.0 (87.9–89.9)	60.6 (59.1–62.1)	97.9% (97.2–98.4)	1.3 (1.0–1.7)	0.83
Internal (PlantVillage, n = 4,046)	Absolute color	86.7 (85.6–87.7)	56.5 (55.0–58.1)	97.1% (96.3–97.7)	1.6 (1.3–2.1)	0.81
Internal (PlantVillage, n = 4,046)	Global histogram	69.6 (68.2–71.0)	38.2 (36.7–39.7)	87.2% (85.4–88.8)	4.9 (4.3–5.6)	0.66
Internet (PlantDoc, n = 133)	Proposed	35.9 (31.6–39.8)	3.6 (0.8–8.3)	16 of 24	1.2 (0.8–1.5)	0.33
Internet (PlantDoc, n = 133)	Absolute color	24.1 (19.5–27.1)	1.7 (0.8–2.3)	7 of 11	0.6 (0.0–0.8)	0.24
Internet (PlantDoc, n = 133)	Global histogram	17.3 (8.3–40.6)	1.8 (0.0–3.0)	0 of 12	1.8 (0.0–3.0)	0.21
Internet with local images (n = 133)	Proposed	42.9 (34.8–51.4)	5.3 (2.6–10.5)	5 of 7	1.5 (0.4–5.3)	0.44
Internet with local images (n = 133)	Absolute color	50.4 (42.0–58.7)	4.5 (2.1–9.5)	4 of 6	1.5 (0.4–5.3)	0.47
Internet with local images (n = 133)	Global histogram	48.9 (40.5–57.3)	12.8 (8.1–19.5)	17 of 17	0.0 (0.0–2.8)	0.46

Values in parentheses are Wilson 95% confidence intervals, except where marked. Internet results are means of the five rounds, with the range over rounds in parentheses; correct among diagnosed is pooled over the rounds because few images were diagnosed.

The filled markers, which include the familiarity check, trace a curve that falls steeply as coverage decreases, and the star marks the calibrated operating point pooled over the five rounds. The open markers, which show the result without deferral, lie above the filled markers at comparable coverage, indicating that the familiarity check contributes to error reduction beyond what the agreement threshold alone provides. For all three descriptors, raising the agreement threshold from 5 to 10 lowered the error among diagnosed images, and at any given coverage the error was lowest for the proposed descriptor. This pattern supports the design choice of a two-part rule: agreement alone reduces errors but at a cost in coverage that depends on the descriptor, whereas the combination of agreement and familiarity achieves a lower error at the same coverage.

Figure 2(b) shows the outcome for every image with and without deferral for the proposed descriptor. Without deferral, 11.0% of internal images and 64.1% of internet images received a wrong diagnosis. With deferral, wrong diagnoses fell to 1.3% of internal images and 1.2% of internet images, and the majority of images that would have been wrong were referred rather than diagnosed. The figure makes the trade-off explicit: deferral converts potential errors into referrals, which is the intended behavior for a diagnostic tool that is meant to support rather than replace expert judgement.

3.2. Internet Images and the Effect of Local Images

Without deferral, retrieval was correct for only 35.9%

© Ariston Publications 2026. All rights reserved.

of the Plant-Doc images, with a range over rounds of 31.6 to 39.8%, compared with 24.1% for absolute color and 17.3% for the global histogram. The proposed descriptor did better than absolute color in every round, with McNemar P between 0.003 and 0.041, which is a consistent though modest advantage on out-of-distribution data. Only 12.5% of the healthy internet images were called healthy, and 38% of all diagnoses named esca or leaf blight, which this set does not contain.

This pattern is a direct consequence of the distance-based retrieval: internet images lie far from the Plant-Village reference images in descriptor space, and their nearest neighbors are therefore drawn from classes that happen to occupy the periphery of the reference distribution rather than from the class that the image actually shows. The method was not designed to diagnose such images, and the low accuracy without deferral confirms that it should not be used for them.

With deferral, 86.6% of the internet images were referred as unfamiliar and 9.8% because the retrieved images disagreed, while 3.6% received a diagnosis. Wrong diagnoses fell from 64.1% to 1.2% of images, and of the 24 diagnoses given over the five rounds, 16 were correct. The high referral rate is the intended behavior for images unlike the reference collection: the method recognizes that it is out of its domain and declines to answer, which is preferable to returning a confident but unsupported label. The small number of diagnoses that were given reflects the few internet images that happen to resemble Plant-Village references closely, and the fact that 16 of 24 were correct suggests that when the method does commit on such images, its evidence is usually meaningful. Adding 106 to 107 Plant-Doc images to the database raised accuracy without deferral to 42.9% but coverage only to 5.3%, and with local images the proposed descriptor no longer did better than the other two.

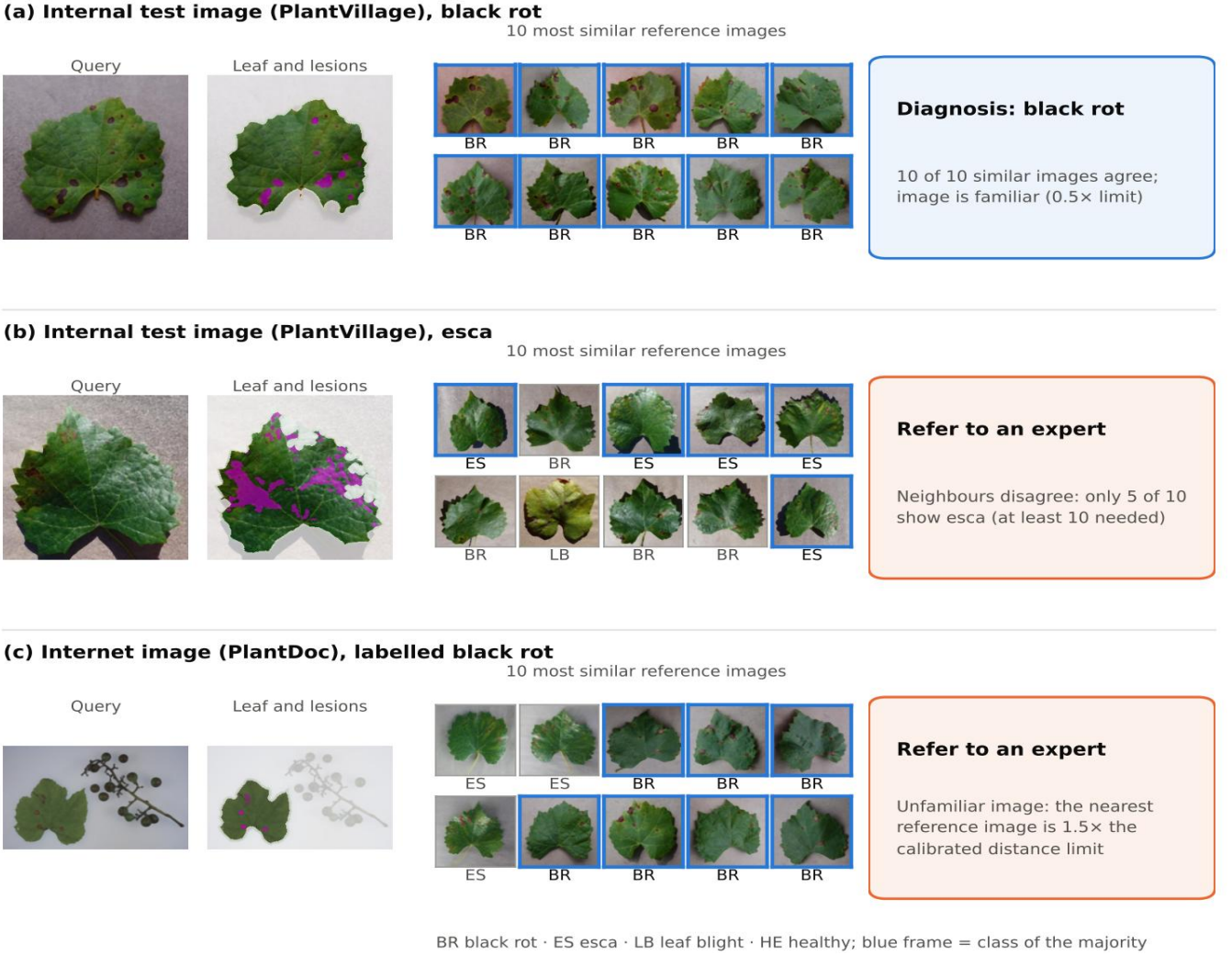


Fig. 1. Diagnose-or-defer applied to three test images. (a) A Plant-Village black rot leaf: all ten retrieved reference images show black rot and the image is familiar, so a diagnosis is given. (b) A Plant-Village esca leaf: only five of the ten retrieved images show esca, fewer than the ten required in this round, so the image is referred. (c) A Plant-Doc internet image labelled black rot: seven of the ten retrieved images show black rot, but its distance to the nearest reference image is 1.5 times the calibrated limit, so the image is referred as unfamiliar. Lesion pixels are shown in magenta and pixels outside the leaf mask are faded.

 © Ariston Publications 2026. All rights reserved.

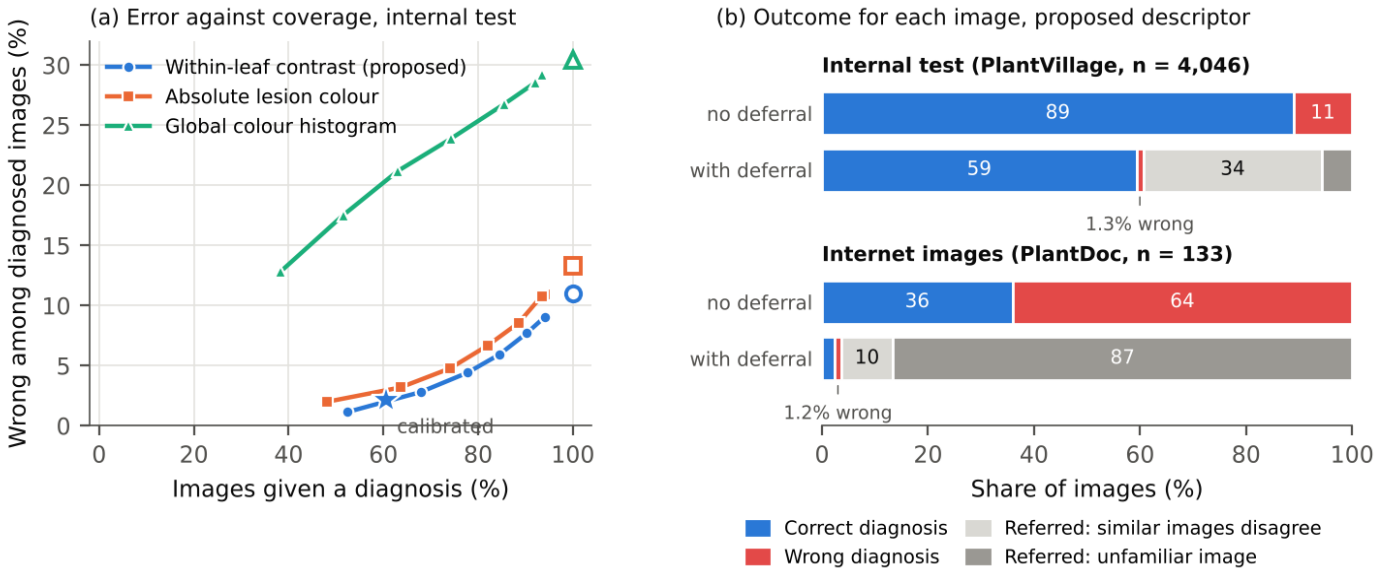


Fig. 2. Deferral behavior and per-image outcomes. (a) Error among diagnosed images against the share of images diagnosed in the internal test, as the agreement threshold rises from 1 to 10 with the familiarity check applied (filled markers). Open markers show the result without deferral, and the star marks the calibrated thresholds of the proposed method, pooled over the five rounds. (b) Outcome for every image with and without deferral, proposed descriptor; internet results are means of the five rounds.

The improvement in accuracy without deferral indicates that local references do help retrieval on local images, but the small gain in coverage indicates that one hundred images are not enough to make the reference database representative of internet variability. The loss of the proposed descriptor's advantage is also instructive: when the database contains images from the same source as the query, the absolute-color descriptor benefits disproportionately, because the acquisition conditions are now matched and absolute color becomes informative rather than misleading. This result supports the view that the within-leaf color contrast in blocks A and B is most valuable when the database and the query come from different acquisition conditions, which is precisely the setting the method was designed for.

3.3. Robustness to Controlled Image Changes

Table 3 reports the effect of six controlled image changes on the internal test images, with the database and thresholds unchanged. Under the four photometric changes, accuracy without deferral fell by 3.7 to 9.5 percentage points for the proposed descriptor, by 5.4 to 24.3 points for absolute color, and by 18.7 to 37.4 points for the global histogram. The smaller degradation of the proposed descriptor under photometric change is the expected consequence of measuring lesion color relative to the healthy tissue of the same leaf: a global color shift affects both the lesion and the healthy tissue, so the difference between them is partly preserved, whereas absolute color and whole-image histograms shift directly with the illumination. The warm and cool casts produced the largest degradation for absolute color and the global histogram, which is consistent with the fact that these changes displace the color distribution most strongly, while the proposed descriptor retained 79.5% and 84.9% accuracy under the cool and warm casts respectively. Blur, which alters the texture features shared by both segmentation-based descriptors, reduced their accuracy to 38%, but every blurred image was referred as unfamiliar, so no wrong diagnosis was given. This result is important because it shows that the familiarity check responds to a change that the descriptor itself does not tolerate: the descriptor's distance from the reference database increases under blur, and the method declines to diagnose rather than returning a degraded label. The global histogram, which does not rely on texture in the same way, was less affected by blur but was already the weakest descriptor overall, and its wrong-diagnosis rate under blur was 4.9%, the same as under no change, because it continued to diagnose images it could not reliably classify. JPEG compression reduced the accuracy of the proposed descriptor to 68.6% and its coverage to 14.0%, which reflects the loss of fine texture and lesion-boundary detail at quality 20. Across the six changes, wrong diagnoses with deferral stayed at or below 2.4% of images for the proposed descriptor, against up to 4.2% for absolute color and 8.0% for the global histogram. The proposed descriptor therefore maintained the lowest worst-case wrong-diagnosis rate under every change tested, which is the property that matters most for a diagnostic tool intended to support decisions rather than replace them.

© Ariston Publications 2026. All rights reserved.

Table 3. Effect of controlled image changes on the internal test images (n = 4,046), with the database and thresholds unchanged.

Image change	Correct without deferral, proposed (%)	Correct without deferral, absolute color (%)	Correct without deferral, global histogram (%)	Diagnosed, proposed (%)	Wrong diagnoses, proposed (%)	Wrong diagnoses, absolute colour (%)	Wrong diagnoses, global histogram (%)
None	89.0	86.7	69.6	60.6	1.3	1.6	4.9
Darker (× 0.7)	85.3	81.3	50.9	54.9	2.4	2.6	4.4
Brighter (× 1.3)	81.8	70.7	44.7	45.3	1.6	1.1	0.8
Warm cast	84.9	79.0	35.0	58.2	2.4	4.2	1.7
Cool cast	79.5	62.4	32.2	48.5	1.9	2.0	0.0
Blur (σ = 1.5 px)	37.9	38.0	69.7	0.0	0.0	0.0	4.9
JPEG (quality 20)	68.6	63.6	60.9	14.0	2.0	3.9	8.0

Warm cast: red × 1.15 and blue × 0.85; cool cast: the reverse.

3.4. Discussion

The results support three conclusions about the design of the method. First, describing lesions by their color difference from the healthy tissue of the same leaf confers a consistent advantage over absolute color and whole-image histograms, both on internal images and under photometric change. The advantage is modest on internal images, where acquisition conditions are uniform, but it becomes larger under color casts and on internet images, which is precisely where a retrieval-based diagnostic tool is most likely to be used. Second, the two-part deferral rule is effective at converting potential errors into referrals. On internal images, 88.4% of the images that would have been misdiagnosed were referred instead, and on internet images the wrong-diagnosis rate fell from 64.1% to 1.2%. The agreement check and the familiarity check are complementary, as the three cases in Figure 1 show, and neither alone would achieve the same reduction in errors at the same coverage. Third, the method's behavior under controlled image changes is predictable and conservative: it degrades in accuracy when the descriptor is stressed, but it also increases its referral rate, so that wrong diagnoses remain low even when the images are unlike the reference collection.

The limitations of the study qualify these conclusions in several ways. The method was validated only on grape leaves of four classes, from one benchmark collection and one internet collection, and no field images collected by the authors were used. Plant-Village images were taken under controlled conditions on plain backgrounds, and its healthy images come from a different acquisition series from the diseased ones, so the internal results are probably optimistic despite the blocked splits [2, 4]. On internet images the method referred almost every image, and in its present form it is not a field tool; it should be used with single-leaf photographs on a plain background, taken like its reference images. Adding about one hundred internet images to the database did not change this, probably because these images differ widely among themselves, and a database of local photographs taken in a fixed way may behave differently, but this was not tested. The segmentation assumes one leaf on a background less colorful than the leaf, so other leaves, soil, or fruit in the picture are segmented wrongly, and blur alters the texture features on which the descriptor partly relies. The Plant-Doc labels were used as published, although some file names suggest other disorders. A referral is only as useful as the expert who receives it, and with the calibrated thresholds about 40% of Plant-Village images would still need one. Finally, the calibration target was missed in one round and the calibrated threshold varied between rounds, so users should calibrate the method on images from their own setting. These limitations point to clear directions for further work. A reference database assembled from photographs taken in a fixed local protocol, with the same camera and background, would test whether the method can operate as a field tool when its reference images match its queries. A larger and more diverse reference collection might reduce the familiarity referral rate on internet images without compromising the wrong-diagnosis rate, although the present results suggest that one hundred additional images are not

 © Ariston Publications 2026. All rights reserved.

enough. Extending the descriptor to handle multiple leaves or complex backgrounds would require a different segmentation strategy, and the texture features would need to be made more robust to blur if the method is to be used with hand-held photographs taken without careful focus. Finally, a user study with domain experts receiving the referrals would establish whether the displayed reference images and the stated reason for referral are useful in practice, which is the ultimate test of a diagnostic tool designed to support rather than replace human judgement.

4. CONCLUSION

This paper presented diagnose-or-defer, a content-based image retrieval method for grape leaf disease identification that either reports the disease supported by the most similar labelled reference images or refers the image to an expert when its evidence is weak. The method describes lesions by how their color differs from the healthy tissue of the same leaf, combines this with leaf texture and lesion shape in a 56-value descriptor, retrieves the ten most similar reference images, and applies two calibrated checks before issuing a diagnosis. On 4,046 Plant-Village images under acquisition-blocked cross-validation, accuracy without deferral was 89.0%, and with deferral 60.6% of images received a diagnosis, 97.9% of which were correct. Wrong diagnoses fell from 11.0% to 1.3% of all images, and 88.4% of the images that would have been misdiagnosed were referred instead. On 133 internet images from Plant-Doc, retrieval alone was wrong for 64%, whereas the method referred 96% of them and gave a wrong diagnosis for 1.2%. Under six controlled image changes, wrong diagnoses with deferral remained at or below 2.4% of images for the proposed descriptor, the lowest worst-case rate among the three descriptors compared. The method is suited to single-leaf photographs taken on a plain background under conditions similar to its reference images, and it recognizes most photographs that are not. Its main limitation is that it referred almost every internet image, so it is not yet a field tool. Future work should test a reference database of local photographs taken in a fixed protocol, extend the segmentation to complex backgrounds, and evaluate the usefulness of referrals with domain experts.

DECLARATIONS

CrediT authorship contribution statement

All authors contributed equally to the preparation of this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Data availability

The images are publicly available from the repositories cited [4, 5]. The code, the fixed parameters, the data splits, the per-image predictions and the scripts that reproduce every number and figure in this article are available at [repository link to be added before submission].

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude (Anthropic) to design and write the code of the method, to run the validation, to draft the code that draws the figures from the images and results, and to draft and revise the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

[1] Mohanty, S.P., Hughes, D.P. and Salathé, M., **2016**. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7*, p.215232. https://doi.org/10.3389/fpls.2016.01419

[2] Noyan, M.A., **2022**. Uncovering bias in the PlantVillage dataset. *arXiv preprint arXiv:2206.04374*. https://doi.org/10.48550/arXiv.2206.04374

[3] Barbedo, J.G., **2018**. Factors influencing the use of deep learning for plant disease recognition. *Biosystems engineering*, *172*, pp.84-91. https://doi.org/10.1016/j.biosystemseng.2018.05.013

[4] Kapoor, S. and Narayanan, A., **2023**. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9). https://doi.org/10.1016/j.patter.2023.100804

[5] Smeulders, A.W., Worring, M., Santini, S., Gupta, A. and Jain, R., **2000**. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(12), pp.1349-1380. https://doi.org/10.1109/34.895972

© Ariston Publications 2026. All rights reserved.

[6] Rudin, C., **2019**. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), pp.206-215. https://doi.org/10.1038/s42256-019-0048-x

[7] Chow, C., **1970**. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, *16*(1), pp.41-46. https://doi.org/10.1109/TIT.1970.1054406

[8] Patil, J.K. and Kumar, R., **2017**. Analysis of content based image retrieval for plant leaf diseases using color, shape and texture features. *Engineering in agriculture, environment and food*, *10*(2), pp.69-78. https://doi.org/10.1016/j.eaef.2016.11.004

[9] Cover, T. and Hart, P., **1967**. Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), pp.21-27. https://doi.org/10.1109/TIT.1967.1053964

[10] Geifman, Y. and El-Yaniv, R., **2017**. Selective classification for deep neural networks. *Advances in neural information processing systems*, *30*.

[11] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S. and Batra, N., **2020**. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (pp. 249-253). https://doi.org/10.1145/3371158.3371196

[12] Hughes, D. and Salathé, M., **2015**. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*. https://doi.org/10.48550/arXiv.1511.08060

[13] Otsu, N., **1979**. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), pp.62-66. https://doi.org/10.1109/TSMC.1979.4310076

[14] Lloyd, S., **1982**. Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), pp.129-137. https://doi.org/10.1109/TIT.1982.1056489

[15] Haralick, R.M., Shanmugam, K. and Dinstein, I.H., **1973**. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610-621. https://doi.org/10.1109/TSMC.1973.4309314

[16] Ojala, T., Pietikainen, M. and Maenpaa, T., **2002**. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), pp.971-987. https://doi.org/10.1109/TPAMI.2002.1017623

[17] Karthikeyan, M. and Raja, D., **2023**. Deep transfer learning enabled DenseNet model for content based image retrieval in agricultural plant disease images. *Multimedia Tools and Applications*, *82*(23), pp.36067-36090. https://doi.org/10.1007/s11042-023-14992-z

[18] Hendrycks, D. and Dietterich, T., **2019**. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*. https://doi.org/10.48550/arXiv.1903.12261

[19] Gonzalez, R.C., **2018**. Richard e. woods digital image processing, pearson.

[20] Reddi, S.S., Rudin, S.F. and Keshavan, H.R., **1984**. An optimal multiple threshold scheme for image segmentation. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), pp.661-665. https://doi.org/10.1109/TSMC.1984.6313341

[21] Algazi, V., **1966**. Useful approximations to optimum quantization. *IEEE Transactions on Communication Technology*, *14*(3), pp.297-301. https://doi.org/10.1109/TCOM.1966.1089333

[22] Peleg, S., Naor, J., Hartley, R. and Avnir, D., **1984**. Multiple resolution texture analysis and classification. *IEEE transactions on pattern analysis and machine intelligence*, (4), pp.518-523. https://doi.org/10.1109/TPAMI.1984.4767557

[23] Ojala, T., Pietikäinen, M. and Mäenpää, T., **2001**, March. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. In *International conference on advances in pattern recognition* (pp. 399-408). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44732-6_41

[24] Hendrycks, D. and Dietterich, T.G., **2018**. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*. https://doi.org/10.48550/arXiv.1807.01697

[25] Wilson, E.B., **1927**. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), pp.209-212. https://doi.org/10.1080/01621459.1927.10502953

 © Ariston Publications 2026. All rights reserved.

[26] McNemar, Q., **1947**. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), pp.153-157. https://doi.org/10.1007/BF02295996

[27] Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J. and Kern, R., **2020**. Array programming with NumPy. *nature*, *585*(7825), pp.357-362. https://doi.org/10.1038/s41586-020-2649-2

[28] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. and Van Der Walt, S.J., **2020**. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), pp.261-272. https://doi.org/10.1038/s41592-019-0686-2

[29] Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E. and Yu, T., **2014**. scikit-image: image processing in Python. *PeerJ*, *2*, p.e453. https://doi.org/10.7717/peerj.453

[30] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., **2011**. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

© Ariston Publications 2026. All rights reserved.