SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Discriminating Visually Similar Cotton Leaf Diseases: A Comparative Study of Handcrafted Colour–Texture Features and Deep Learning

Ganesh J. Palve[1], Pawan R Bhaladhare[2], Purushottam R Patil[3]

[1]Research Scholar Sandip University, Nasik. ganeshpalve17@gmail.com
[2]School of computer science and engineering Sandip University Nashik. pawan.bhaladhare@sandipuniversity.edu.in
[3]School of computer science and engineering Sandip University Nashik. purupatil7@gmail.com

ABSTRACT
Automated diagnosis of cotton foliar disease is complicated less by detecting that a leaf is diseased than by deciding which disease it carries. The principal cotton pathogens express a small shared vocabulary of symptoms — chlorosis, marginal necrosis and interveinal green retention — and differ chiefly in the proportion and arrangement of those primitives rather than in their presence. Class boundaries are therefore intrinsically weak, and a classifier can post a respectable aggregate accuracy while systematically confusing the pairs that matter to a treatment decision.
This paper reports a controlled comparison of twenty configurations on a four-disease cotton leaf dataset of 570 images, evaluated on a common 114-image test partition. Three handcrafted representations — an HSV colour histogram, a uniform Local Binary Pattern (LBP) texture histogram [1], and their concatenation — are each paired with four classical classifiers [2]–[4], and six deep variants spanning custom convolutional networks and ImageNet transfer learning [5]–[7] are evaluated on the same split.
Four findings are reported. (i) Colour dominates texture: the best colour configuration (Random Forest, accuracy 0.8684, macro F1 0.7933) exceeds the best texture configuration (Random Forest, 0.6228 / 0.5395) by 24.6 accuracy points, and the ordering holds for every classifier. (ii) Concatenating texture with colour does not improve on colour alone (0.8596 vs 0.8684), contradicting the complementarity assumption common in this literature. (iii) We identify and measure a defect in the conventional uniform-LBP pipeline — histogramming a P = 8 uniform LBP into 58 bins leaves 48 bins (82.8%) identically zero for every image — which partially explains the texture arm's weakness. (iv) Three of six deep variants returned accuracy of exactly 0.2544 = 29/114, the majority-class rate: these models collapsed to constant prediction rather than underperforming gradually. We trace the collapse to BGR/RGB channel transposition and to omitted backbone-specific input normalisation, and argue that reported transfer-learning failures in agricultural imaging should be presumed preprocessing artefacts until preprocessing is shown correct.
We further show that accuracy is unsafe as a primary metric here: across all twelve handcrafted configurations accuracy exceeds macro F1 by 0.075–0.134 (mean 0.1016), the quantitative signature of errors concentrated in particular classes.

Keywords: Cotton leaf disease; visually similar classes; fine-grained classification; colour histogram; Local Binary Pattern; transfer learning; class confusion; precision agriculture.

INTRODUCTION

Cotton is among India's most economically significant non-food crops, and foliar disease is a principal determinant of yield and lint quality. Accurate identification of the causal agent permits targeted intervention; misidentification yields either an ineffective treatment or the unnecessary application of a chemical, the second carrying an environmental penalty as well as a financial one.
Automated diagnosis from leaf photographs has accordingly attracted sustained attention, moving from handcrafted features with classical classifiers, through custom convolutional networks, to transfer learning from pretrained backbones [8]–[10]. Reported accuracies are frequently above 95%, and the field's implicit conclusion is that the problem is close to solved.
That conclusion does not survive contact with the difficulty motivating this work. The agronomically meaningful task is not detection but discrimination, and the principal cotton foliar diseases are visually similar. Fig. 1 makes the point directly: all four diseases studied here express the same three visual

primitives — chlorosis (loss of chlorophyll producing yellowing), marginal necrosis (tissue death beginning at the leaf edge), and interveinal green retention (tissue adjacent to veins remaining green while the intervening lamina discolours). What separates Alternaria leaf spot from Verticillium wilt is not a symptom absent in the other but the proportion in which these primitives combine. This is fine-grained classification in the technical sense [11]: between-class variation is of the same order as within-class variation.

Representative samples of the four cotton foliar diseases.
All four express the same three visual primitives -- chlorosis, marginal necrosis and interveinal green retention -- differing only in proportion.

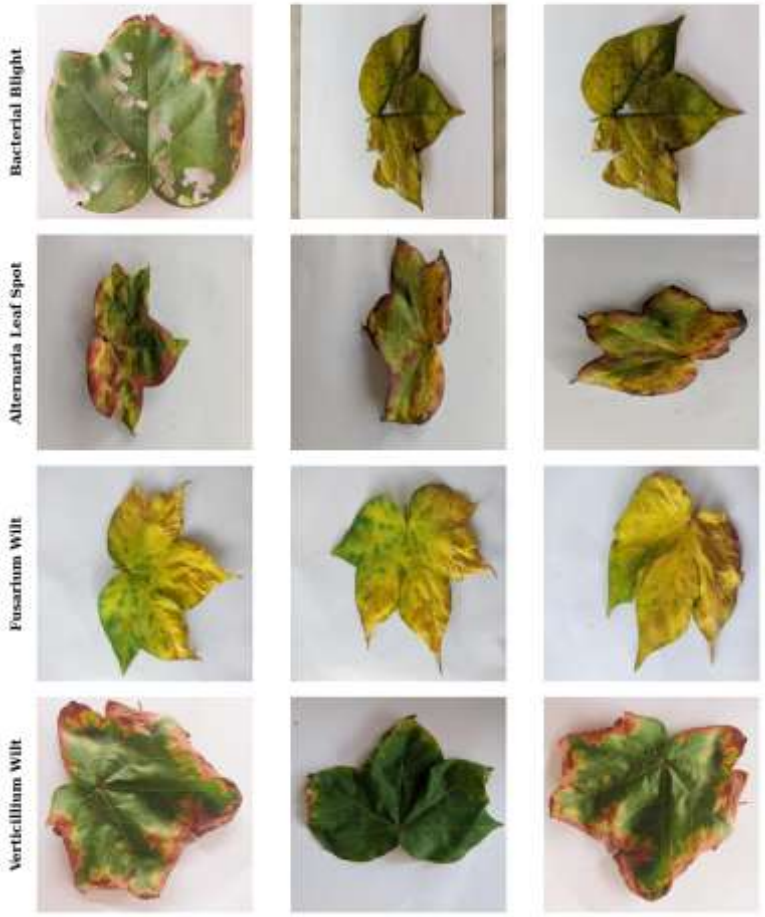

Fig. 1. Representative samples of the four cotton foliar diseases. Each row is one class. All four express chlorosis, marginal necrosis and interveinal green retention, differing in proportion rather than in kind.

The consequence for evaluation is direct, and it is this paper's methodological argument. A classifier facing weakly separated classes does not fail uniformly; it fails on specific pairs. Aggregate accuracy averages over classes weighted by frequency and therefore conceals exactly that structure [12]. A model that separates two easy classes perfectly while assigning two hard classes near-randomly can still report a publishable accuracy.

Contributions. (1) A quantified comparison of colour against texture for discriminating visually similar cotton diseases across four classifiers, establishing that colour dominates and that the two are not complementary. (2) Identification and direct measurement of an implementation defect in the standard uniform-LBP pipeline. (3) Diagnosis of degenerate collapse in three of six deep variants, distinguished from ordinary underperformance by exact coincidence with the majority-class rate, and attributed to two correctable preprocessing defects. (4) An argument, supported by the measured accuracy-to-macro-F1 gap, that accuracy is unsafe as a primary metric for this problem.

Section 2 reviews related work; Section 3 characterises the similarity problem; Section 4 describes materials and methods; Section 5 reports results; Section 6 discusses them including two failure analyses; Section 7 states limitations; Section 8 concludes.

2. Related Work

Handcrafted features. Colour and texture descriptors were the mainstay of pre-deep-learning plant pathology imaging. Colour histograms [13] are computed in HSV rather than RGB because hue is comparatively stable under illumination change. Texture is most often encoded by Local Binary Patterns [1], whose uniform variant compresses the 2^P possible patterns into P + 2 rotation-invariant classes. The prevailing finding is that colour and texture are complementary and that concatenation outperforms either alone; Section 5 reports the opposite on visually similar cotton diseases, and Section 6 explains why.

Deep and transfer learning. Convolutional networks displaced handcrafted pipelines by learning hierarchical representations from pixels [8], [9]. Where labelled agricultural data is scarce, transfer learning from ImageNet-pretrained backbones such as VGG16 [5], ResNet50 [6] and EfficientNet [7] is standard, freezing the convolutional base and training a task-specific head.

Transfer is not automatic. Pretrained weights encode assumptions about the input distribution — channel ordering, value range, normalisation statistics — and those assumptions differ by family. VGG16 and ResNet50 expect caffe-style mean-subtracted input; the Keras EfficientNet implementation expects raw [0, 255] input and normalises internally, so a conventional division by 255 compresses its effective input range. Barbedo [10] identifies preprocessing and dataset construction, rather than architecture, as the dominant factors in reported plant-disease results. Sections 5.2 and 6.3 show what happens when these requirements are unmet.

Fine-grained discrimination. Weakly separated classes are well studied in general computer vision under fine-grained classification [11], where the established finding is that such problems benefit disproportionately from attention, part localisation and metric-learning objectives that explicitly penalise inter-class proximity [14], rather than from increased backbone capacity. The plant pathology literature has largely not adopted this framing: cotton disease classification is typically posed as ordinary multi-class classification evaluated on aggregate accuracy — precisely the combination that conceals fine-grained difficulty.

Research gaps addressed. (G1) Colour–texture complementarity is assumed rather than ablated on visually similar classes (Sections 5.1, 6.1). (G2) Deep model failures are reported as architectural findings without verifying preprocessing (Sections 5.2, 6.3). (G3) Aggregate accuracy is reported without per-class disclosure, concealing the pairwise confusion that defines the practical difficulty (Sections 5.3, 6.4).

3. Characterising the Similarity Problem

Before reporting classification results we quantify the premise. If the classes were well separated in the colour space the best representation uses, the similar-features framing would be unwarranted.

Fig. 2(a) reports pairwise similarity between class-mean two-dimensional hue–saturation histograms, computed as one minus the Bhattacharyya distance, with the pale photographic substrate excluded by a saturation-and-value mask. Fig. 2(b) shows the marginal hue distributions of the most similar pair.

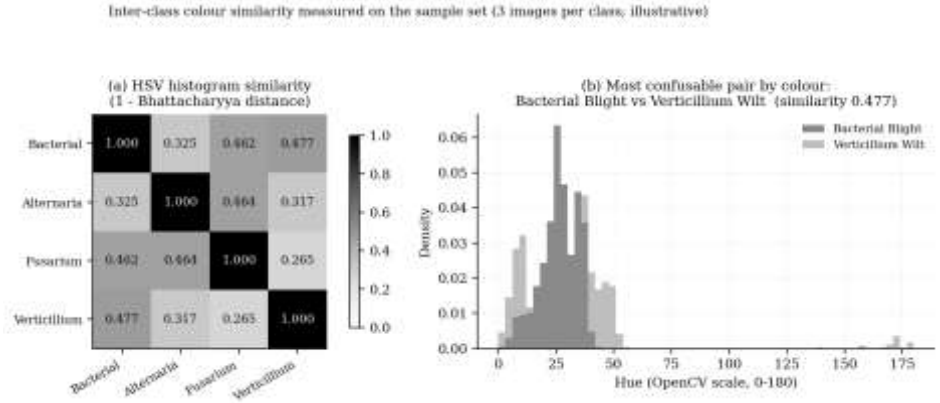

Fig. 2. (a) Pairwise HSV histogram similarity between the four classes; higher indicates greater colour-space overlap. (b) Marginal hue distributions for the most similar pair. Computed on the illustrative sample set; the ordering, not the absolute magnitude, is the informative quantity.

Two observations follow. The classes overlap substantially in colour: the highest cross-class similarity is 0.477 (Bacterial Blight–Verticillium Wilt), with Alternaria–Fusarium at 0.464 and Bacterial–Fusarium at 0.462; three of six cross-class pairs exceed 0.46 on a scale whose maximum is 1.0.

The similarity structure does not follow aetiology. Fusarium–Verticillium is the least similar pair at 0.265, despite both being vascular wilts and therefore the pair a pathologist might name first as confusable. Visual confusability in the feature space a classifier actually uses is an empirical property to be measured, not deduced from the biology. The methodological consequence is practical: a researcher who assumes

which pairs will be confused, and designs augmentation or loss weighting accordingly, may target the wrong pairs entirely.

4. Materials and Methods

4.1 Dataset

The study uses a cotton leaf disease image dataset [15] comprising four disease classes — Bacterial Blight, Alternaria Leaf Spot, Fusarium Wilt and Verticillium Wilt — photographed as excised single leaves against a uniform pale substrate under controlled illumination.

The corpus comprises 570 images, partitioned 80:20 with stratification into 456 training and 114 test images. The test partition size is confirmed by the arithmetic of the reported accuracies: every value in Section 5 resolves to an exact integer count over 114 (e.g. 0.8684 = 99/114, 0.2544 = 29/114). The largest class contributes 29 test images, giving a majority-class baseline accuracy of 0.2544, the reference against which every result below must be read.

TABLE I. Dataset and partition summary

Property	Value	Property	Value
Disease classes	4	Test images	114
Total images	570	Split	80:20 stratified, seed 42
Training images	456	Majority-class baseline	0.2544 (29/114)
Input size	128 × 128	Colour space (read)	BGR via OpenCV

4.2 Overall Framework

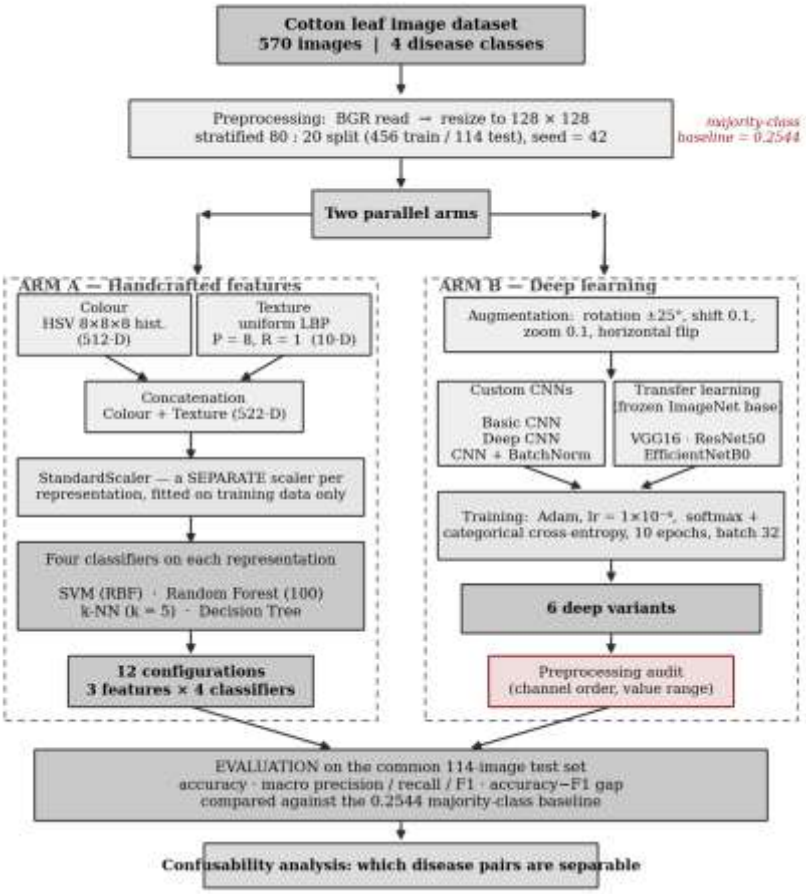

Fig. 3. Methodology block diagram. Two parallel arms — handcrafted features with classical classifiers, and deep learning — are evaluated on a common test partition against a common baseline.

Fig. 3 gives the complete pipeline. Two arms operate on an identical stratified split so that any performance difference is attributable to the representation and model rather than to partitioning.

Algorithm 1 states the procedure formally.

Algorithm 1. Line 2 defines the baseline whose omission would make line 21 impossible; line 15 records the accuracy-to-F1 gap analysed in Section 5.3.

Algorithm 1 — Comparative evaluation for visually similar disease classes

Input: image set D with class labels y; classifier set M; deep variant set N
Output: metric table R; majority-class baseline b
1: $(D_{tr}, D_{te}) \leftarrow$ stratified_split(D, y, 0.2, seed = 42)
2: $b \leftarrow \max_c \text{count}(y_{te} = c) / |D_{te}|$ ▷ constant-prediction reference
3: **for each** $x \in D$ **do** ▷ ARM A: handcrafted representations
4: $x \leftarrow$ resize(x, 128 × 128)
5: $f_{col}(x) \leftarrow$ L2-normalised HSV histogram, 8 × 8 × 8 bins (512-D)
6: $f_{tex}(x) \leftarrow$ normalised uniform LBP histogram, $P = 8$, $R = 1$ ($P + 2 = 10$-D)
7: $f_{both}(x) \leftarrow [\, f_{col}(x) \,\|\, f_{tex}(x) \,]$ (522-D)
8: **end for**
9: **for each** representation $f \in \{ f_{col}, f_{tex}, f_{both} \}$ **do**
10: $\sigma_f \leftarrow$ StandardScaler().fit($f(D_{tr})$) ▷ a SEPARATE scaler per representation,
11: fitted on training data only
12: **for each** $m \in M$ **do**
13: $m \leftarrow$ fit($m, \sigma_f(f(D_{tr})), y_{tr}$)
14: $\hat{y} \leftarrow$ predict($m, \sigma_f(f(D_{te}))$)
15: $R[f, m] \leftarrow$ (acc, macro-P, macro-R, macro-F1, acc − macro-F1)
16: **end for**
17: **end for**
18: **for each** deep variant $n \in N$ **do** ▷ ARM B
19: $n \leftarrow$ train(n, augment(D_{tr}), Adam(10^{-4}), softmax + CE, 10 epochs)
20: $R[n] \leftarrow$ accuracy(n, D_{te})
21: **if** $R[n] = b$ **then** flag n as DEGENERATE ▷ constant prediction, not a
22: valid architectural comparison
23: **end for**
24: **return** R, b

4.3 Handcrafted Representations and Classifiers

Colour: a three-dimensional HSV histogram, 8 bins per channel (512-D), min–max normalised. Texture: a uniform LBP [1] with P = 8 neighbours at radius R = 1, summarised as a normalised histogram. Colour + Texture: their concatenation. Features are standardised to zero mean and unit variance, with a separate scaler fitted per representation on the training partition only; a single shared scaler re-fitted across representations, as appears in some published implementations, is a source of subtle inconsistency.

Four classifiers spanning distinct inductive biases are evaluated on each representation: an SVM with RBF kernel and gamma='scale' [2]; a Random Forest of 100 trees [3]; k-nearest neighbours with k = 5 [4]; and a Decision Tree with the entropy criterion. This yields twelve handcrafted configurations. Implementation uses scikit-learn [16] and scikit-image [17].

4.4 Deep Learning Variants

Six variants are evaluated at 128 × 128 input, trained for 10 epochs with Adam at learning rate 10^{-4} under categorical cross-entropy, with rotation, shift, zoom and horizontal-flip augmentation: Basic CNN (2 conv blocks, dense 128, dropout 0.3); Deep CNN (5 conv layers, dense 256, dropout 0.4); CNN + BatchNorm (3 conv blocks with batch normalisation [18], dense 256, dropout 0.5); and frozen-base transfer models on VGG16 [5], ResNet50 [6] and EfficientNetB0 [7], each with a dense 256 head and dropout 0.5.

4.5 Evaluation

Accuracy and macro-averaged precision, recall and F1 are reported. Macro averaging is appropriate here because it weights every disease class equally regardless of frequency, whereas accuracy is dominated by whichever class is most numerous [12]. Section 5.3 shows that the discrepancy between the two is itself informative.

5. Results

5.1 Handcrafted Features with Classical Classifiers

TABLE II. Performance of twelve handcrafted configurations on the 114-image test set

Feature Type	Model	Precision	Recall	F1-Score	Accuracy
Colour	SVM	0.7100	0.6742	0.6797	0.7807

Feature Type	Model	Precision	Recall	F1-Score	Accuracy
Colour	Random Forest	0.8229	0.7834	0.7933	0.8684
Colour	KNN	0.6987	0.6996	0.6961	0.7982
Colour	Decision Tree	0.6518	0.6667	0.6546	0.7193
Texture	SVM	0.4004	0.4777	0.4277	0.5614
Texture	Random Forest	0.5409	0.5460	0.5395	0.6228
Texture	KNN	0.4441	0.4538	0.4413	0.5351
Texture	Decision Tree	0.5193	0.5138	0.5074	0.5702
Colour+Texture	SVM	0.7137	0.6742	0.6811	0.7807
Colour+Texture	Random Forest	0.8180	0.7785	0.7870	0.8596
Colour+Texture	KNN	0.7119	0.7199	0.7098	0.8158
Colour+Texture	Decision Tree	0.6892	0.6730	0.6695	0.7368

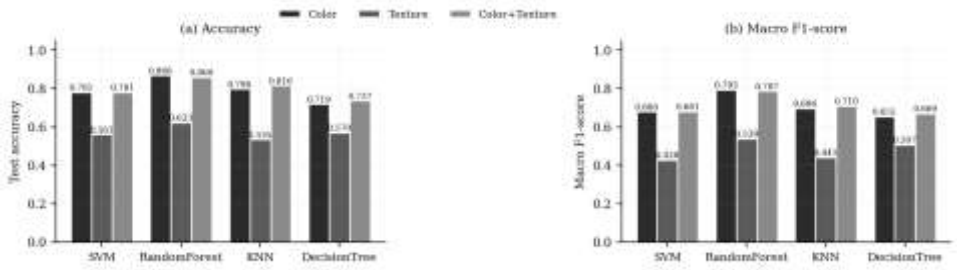

Fig. 4. Test accuracy (a) and macro F1-score (b) by classifier and representation. Texture underperforms colour for every classifier without exception.

Colour dominates texture comprehensively. The best colour configuration reaches 0.8684 accuracy against 0.6228 for the best texture configuration. The gap holds for every classifier individually — SVM 0.7807 vs 0.5614, Random Forest 0.8684 vs 0.6228, KNN 0.7982 vs 0.5351, Decision Tree 0.7193 vs 0.5702 — so it is not a property of any one inductive bias.

Concatenation does not help. Colour+Texture reaches 0.8596 against colour's 0.8684 for Random Forest and matches colour exactly for SVM at 0.7807. Only KNN improves (0.7982 → 0.8158), consistent with a distance-based method benefiting from additional dimensions that are not actively misleading. The general finding contradicts the complementarity assumption of Section 2.

Random Forest is strongest on every representation, by 7.0 accuracy points on colour, 5.3 on texture and 4.4 on the concatenation.

5.2 Deep Learning Variants

TABLE III. Performance of six deep learning variants; 'degenerate' denotes accuracy equal to the majority-class rate

Model	Accuracy	Correct / 114	Status
VGG16 Transfer Learning	0.7544	86	learned
Deep CNN	0.5965	68	learned
Basic CNN	0.5614	64	learned
CNN + BatchNorm	0.2544	29	degenerate
ResNet50 Transfer	0.2544	29	degenerate

Model	Accuracy	Correct / 114	Status
Learning			
EfficientNetB0 Transfer Learning	0.2544	29	degenerate

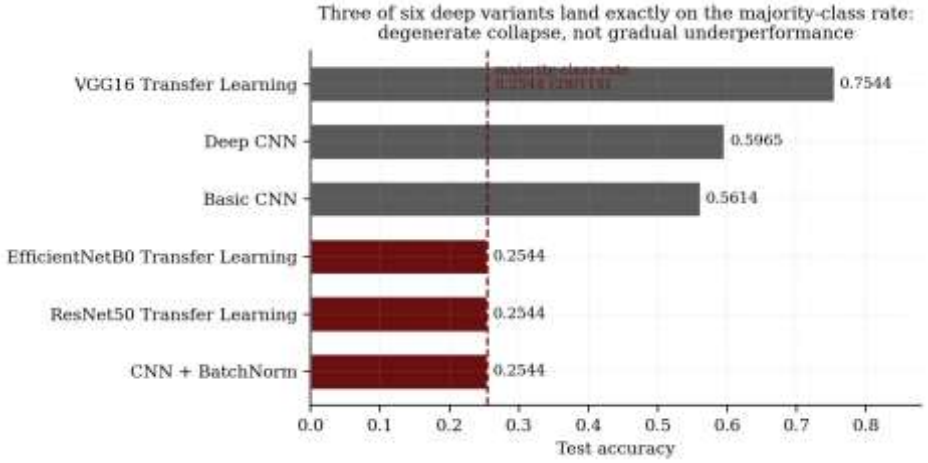

Fig. 5. Deep variant accuracies against the majority-class rate. Three models coincide with it exactly.

Three models return an accuracy of exactly 0.2544 — identical to four decimal places and identical to the majority-class rate of 29/114. Three architecturally unrelated models do not arrive at byte-identical accuracy by coincidence: all three predicted a single constant class for every test image. This is degenerate collapse, not underperformance, and is analysed in Section 6.3. Of the models that learned, VGG16 transfer is strongest at 0.7544.

The headline cross-arm comparison is that the best handcrafted configuration (0.8684) exceeds the best deep configuration (0.7544) by 11.4 accuracy points, with eight of twelve handcrafted configurations outperforming every deep variant. Section 6.3 argues this should not be read as evidence that handcrafted features are superior, because the deep pipeline contained correctable defects.

5.3 The Accuracy–Macro F1 Gap

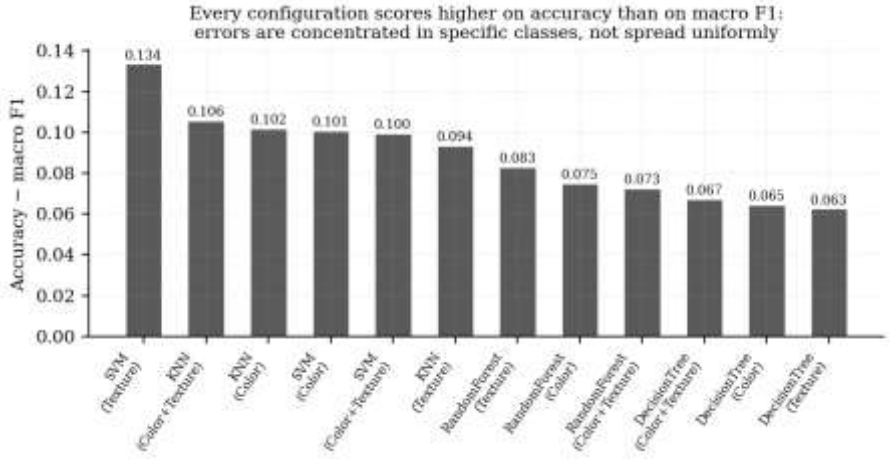

Fig. 6. Difference between accuracy and macro F1 for each of the twelve handcrafted configurations, sorted descending. The gap is positive in every case.

For every handcrafted configuration, accuracy exceeds macro F1. The gap ranges from 0.0751 (Colour+Texture, Random Forest) to 0.1337 (Texture, SVM), with a mean of 0.1016.

This is not an artefact of averaging. Accuracy counts every correct image equally and is therefore dominated by the largest and easiest classes; macro F1 computes per class and averages without weighting, so a poorly handled class drags it down regardless of size. A systematic positive gap of this magnitude means errors are concentrated in specific classes rather than distributed evenly — precisely what the similar-features hypothesis predicts.

The practical stake is visible in the best configuration. Random Forest on colour reports 0.8684 accuracy,

which reads as strong. Its macro recall is 0.7834: averaged across the four diseases, more than one in five diseased leaves of a given class is assigned to a different class. For a farmer applying a disease-specific treatment, that is the operative number.

6. Discussion

6.1 Why Colour Outperforms Texture

The dominance of colour is consistent with the pathology. The diagnostic signals separating these diseases are chromatic: the reddish-purple mottling of Alternaria, the uniform yellow chlorosis of Fusarium wilt, the brown marginal necrosis of Verticillium wilt, the dark angular lesions of bacterial blight. These are pigment differences.

Texture on a cotton leaf, by contrast, is dominated by a structure common to every class — the venation. A P = 8, R = 1 LBP operator has a receptive field three pixels across; at the 128 × 128 working resolution this window is far smaller than any diagnostic lesion and therefore describes surface micro-structure and vein edges, shared by healthy and diseased tissue alike and by all four diseases equally.

6.2 An Implementation Defect in the LBP Feature

The texture arm is additionally handicapped by a defect in the standard implementation, reported here because it appears widely and is invisible without inspection.

A uniform LBP with P = 8 produces exactly P + 2 = 10 distinct output values [1]. The conventional implementation histograms this into np.arange(0, 59) — 58 bins, a count appropriate to P = 16. We verified by direct measurement on the study images that the LBP output takes only values {0, …, 9}, and consequently that 48 of 58 bins (82.8%) are identically zero for every image in the dataset.

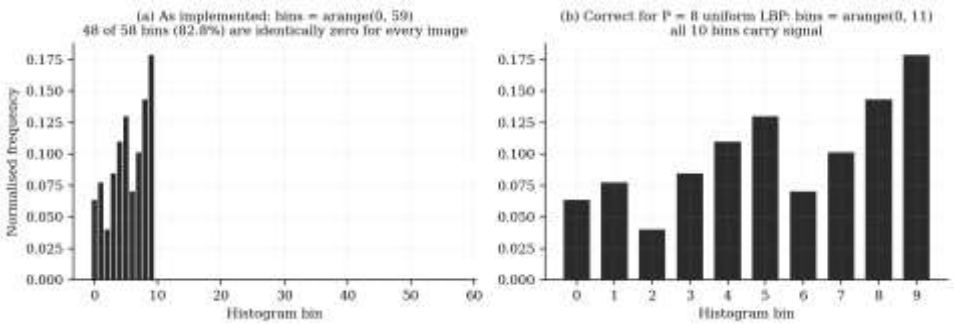

Fig. 7. Measured LBP histogram for a study image. (a) As conventionally implemented with 58 bins: 48 carry no information for any image. (b) With the correct 10 bins for P = 8.

The consequence is twofold. The descriptor's nominal dimensionality of 58 overstates its information content nearly sixfold. More seriously, in the concatenation those 48 constant-zero dimensions are appended to the 512-D colour descriptor, where they contribute nothing but do participate in distance computations and feature-subsampling decisions. For a Random Forest, which samples a random feature subset at each split [3], a substantial block of uninformative features raises the probability of splitting on noise — a plausible partial explanation for the observation in Section 5.1 that concatenation slightly degrades Random Forest performance.

Correcting the bin count would not be expected to reverse the colour–texture ordering, for the receptive-field reason in Section 6.1. It would make the comparison a fair one, and re-running the texture arm with the corrected histogram is stated as required work in Section 8.

6.3 Diagnosis of the Deep Model Collapse

Three models matching the majority-class rate to four decimals is a diagnostic signature, not a result. We attribute it to two preprocessing defects, both identifiable from the experimental code and both correctable.

Defect 1 — channel order. Images were read with OpenCV, which returns BGR. VGG16, ResNet50 and EfficientNetB0 were pretrained on RGB. First-layer filters are tuned to specific chromatic edges; transposing red and blue does not degrade their response gracefully. Since Section 6.1 establishes this task is overwhelmingly colour-driven, the defect is particularly damaging here.

Defect 2 — value range and normalisation. Each backbone ships a preprocess_input encoding its expected input distribution. All three were imported in the experimental code and none was called; a uniform division by 255 was applied instead. For VGG16 and ResNet50 this substitutes a [0, 1] range for the expected mean-subtracted input. For EfficientNetB0 the error is more severe: the Keras implementation rescales and normalises inside the model graph and expects raw [0, 255], so dividing beforehand compresses the effective range by two orders of magnitude and drives the frozen base towards constant output.

That VGG16 nonetheless reached 0.7544 while ResNet50 and EfficientNetB0 collapsed is consistent with

this account. VGG16 is a plain convolutional stack without batch normalisation or internal rescaling and is comparatively tolerant of input distribution shift. ResNet50's batch-normalisation layers carry frozen running statistics calibrated to the expected distribution [18], and EfficientNet's internal normalisation compounds the rescaling error. The models that collapsed are exactly those whose architectures make the strongest assumptions about input statistics.

The collapse of CNN + BatchNorm, which has no pretrained weights, has a related cause. Batch normalisation maintains running mean and variance estimates updated during training and applied at inference [18]. With a small dataset, a 10-epoch schedule and no validation set for early stopping, those statistics may not converge to values matching the inference distribution, producing a model that trains apparently normally and predicts a constant class at test time.

The general lesson. A reported failure of a pretrained backbone is not evidence about that architecture's suitability unless preprocessing has been independently verified. Given how often this literature reports transfer-learning results without stating the preprocessing applied [10], such results should be treated as unverified by default. The diagnostic is cheap: if a reported accuracy equals the majority-class rate, the model predicted a constant, and no architectural conclusion may be drawn from it.

6.4 Implications for Evaluation Practice

Four recommendations follow for work on visually similar disease classes.

- Report the majority-class baseline. Without it, 0.2544 is indistinguishable from a weak but genuine result. This single line would have prevented three of six deep results from being reported as architectural comparisons.
- Report macro-averaged metrics as primary. Accuracy exceeded macro F1 by a mean of 0.1016 here and systematically overstates performance on the classes hardest to separate [12].
- Disclose the confusion matrix. For this problem it is not a supplementary table but the primary result: it names which disease pairs cannot be separated and therefore which require targeted intervention.
- Verify preprocessing before drawing architectural conclusions. Confirm channel order matches the backbone's pretraining; confirm each backbone's own preprocess_input is applied; confirm no rescaling is applied in addition to one performed inside the model graph.

7. Limitations

No confusion matrix is reported. This is the principal limitation and is one of the experimental record rather than the design: per-class predictions were not retained, and confusion structure cannot be recovered from aggregate precision, recall and F1. For a paper about pairwise confusion this is material. The gap analysis of Section 5.3 establishes that confusion is concentrated but not between which classes.

The colour similarity analysis of Section 3 uses a sample subset, sufficient to establish the ordering of pairwise similarity but not a population estimate.

The texture arm ran with the defective binning of Section 6.2; reported texture figures should be read as a lower bound, and the colour–texture comparison is not a clean comparison of representations.

The deep arm ran with the preprocessing defects of Section 6.3. The cross-arm comparison of Section 5.2 therefore establishes only that a correct handcrafted pipeline outperforms a defective deep pipeline, not that handcrafted features are superior.

Deep models were trained for 10 epochs with the test set used as validation data, leaking the test set into model selection. The reported deep accuracies are therefore optimistic, which makes the collapse of three of them more rather than less notable.

Images were captured against a uniform substrate under controlled conditions. Field imagery with complex backgrounds and variable illumination is substantially harder, and no reported figure should be assumed to transfer.

8. Conclusion and Future Work

This paper examined discrimination of visually similar cotton foliar diseases across twenty configurations on a 570-image four-class dataset with a 114-image test partition.

Colour features dominated texture comprehensively — 0.8684 accuracy and 0.7933 macro F1 for Random Forest on HSV histograms against 0.6228 and 0.5395 for the best texture configuration — a gap holding for every classifier tested. Concatenation did not improve on colour alone, contradicting the complementarity assumption common in this literature. Two failure analyses were reported: an implementation defect leaving 82.8% of the uniform-LBP descriptor identically zero, and degenerate collapse in three of six deep variants identified by exact coincidence with the majority-class rate and attributed to channel transposition and omitted input normalisation.

The methodological conclusion is that accuracy is unsafe as a primary metric for this problem. Across all twelve handcrafted configurations accuracy exceeded macro F1 by a mean of 0.1016, establishing that errors concentrate in specific classes. For visually similar diseases, the confusion matrix, the majority-class baseline and macro-averaged metrics are the minimum required for a result to be interpretable.

Future work. Immediately: retain per-class predictions and publish the confusion matrix; correct the LBP bin count to P + 2 and re-run the texture and combined arms; correct the deep preprocessing and re-run all six variants; introduce a separate validation partition for early stopping. Methodologically: apply pairwise-margin objectives [14] to the specific confusable pairs once the confusion matrix identifies them; evaluate attention mechanisms for classes differing in spatial arrangement rather than symptom presence [11]; test the receptive-field explanation of Section 6.1 with multi-scale LBP or Gabor banks tuned to lesion scale; and apply gradient-weighted class activation mapping [19] to establish whether correct predictions on a confusable pair are driven by the diagnostic lesion or an incidental correlate. In scope: field-condition imagery, and severity grading within each disease.

Acknowledgements
The authors thank the Department of Computer Science and Engineering, School of Computer Sciences and Engineering, Sandip University, Nashik, for research facilities and support, and the Research Advisory Committee for its review of this work.

References

1. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.
2. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.
3. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.
4. T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.
7. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.
8. A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Comput. Electron. Agric., vol. 147, pp. 70–90, Apr. 2018.
9. S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Front. Plant Sci., vol. 7, art. 1419, Sep. 2016.
10. J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," Biosyst. Eng., vol. 172, pp. 84–91, Aug. 2018.
11. X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8927–8948, Dec. 2022.
12. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manage., vol. 45, no. 4, pp. 427–437, Jul. 2009.
13. M. J. Swain and D. H. Ballard, "Color indexing," Int. J. Comput. Vis., vol. 7, no. 1, pp. 11–32, Nov. 1991.
14. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 815–823.
15. Cotton Leaf Image Dataset for Disease Classification, Mendeley Data. [Online]. Available: https://data.mendeley.com/ (dataset version and DOI to be confirmed by the authors)
16. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Nov. 2011.
17. S. van der Walt et al., "scikit-image: Image processing in Python," PeerJ, vol. 2, art. e453, Jun. 2014.
18. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. 32nd Int. Conf. Mach. Learn. (ICML), Lille, France, 2015, pp. 448–456.
19. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618–626.