

Lightweight Image–Text Triage for Rice Leaf Diseases with Clarify-or-Abstain Decisions and Nepal-Context Advisory Support

Lumeng Han¹, Derek Zhou ^{*2}

Email: lmhan@ucdavis.edu

¹International Agricultural Development, University of California, Davis, CA, USA

²Computer Science, University of California, Berkeley, Berkeley, CA, USA

*Corresponding Author

Abstract

Image-based rice disease screening requires assessment of unfamiliar conditions, degraded photographs, and the evidence used for advice. This study evaluates a modular image–text workflow with separately trained visual and retrieval components. A rotation- and reflection-aware audit of 5,932 rice leaf images identified 343 similarity groups for leakage-reduced training, validation, and testing. Seven lightweight classifiers and a 24,724-parameter CNN were evaluated alongside open-set rejection, image-quality triage, simulated recapture, and external testing on 10,333 exact-unique Paddy Doctor images. Color-texture logistic regression achieved 78.41% grouped-test accuracy and the CNN achieved 81.03%; their accuracies on 4,232 external images from the four known classes fell to 13.33% and 12.55%, respectively. External unknown-detection AUROCs were 0.512 for Mahalanobis scoring and 0.486 for CNN nonconfidence. Across 7,745 scoped CROP records, reranked retrieval recovered the paired answer within five candidates for 934/1,168 test queries (79.97%). Evidence-assisted benchmark accuracy was 65.21% on 365 questions and 41.34% on 179 without an exact question match in the retrieval collection. Four advisory gates enforced disease, location, treatment, and numeric-rate rules, with a trade-off between answer delivery and reference-answer retention. Nepal-context retrieval remained preliminary: the reranker succeeded on 7/8 queries (87.50%; Wilson 95% CI, 52.91–97.76%). Classification, triage, and advisory retrieval provide separate levels of evidence. The findings support further investigation of the workflow, while the external failures preclude claims of reliable field deployment.

Keywords: Rice Leaf Disease; Lightweight Computer Vision; Open-Set Recognition; Agricultural Advisory; Nepal context.

I. INTRODUCTION

Rice disease screening is a natural application for computer vision because visible lesions, discoloration, and venation changes can provide early evidence of disease. Image classifiers have reached high reported accuracy on curated plant-disease collections, and transfer learning has broadened the range of crops and disease classes that can be modeled (Ferentinos, 2018; Mohanty et al., 2016; Too et al., 2019). Rice-specific work has likewise combined learned visual features with support vector machines (Sethy et al., 2020). These advances are useful, but a high closed-set score alone does not establish reliability in field use.

Three deployment gaps are particularly consequential. First, exact copies, crops, recompressions, or transformed versions of a source photograph may cross an ordinary training–test split, exposing models to related acquisition and background cues. Second, a four-class model may encounter nutrient stress, insect injury, mixed infections, healthy leaves, or additional pathogens. Open-set recognition addresses this mismatch by retaining known-class discrimination while rejecting unfamiliar inputs (Bendale & Boulton, 2016; Scheirer et al., 2013). Third, blur, darkness, compression, occlusion, or incomplete framing can obscure symptoms. Out-of-distribution and

misclassification scores can reveal some failures (Hendrycks & Gimpel, 2017), but their operating thresholds and scope require separate evaluation.

A field advisory system also crosses from vision into language. A predicted class can retrieve symptom descriptions, management information, or follow-up questions, yet retrieval introduces another error channel. A relevant-looking passage may concern a different disease or contain a treatment directive that the available evidence does not support. Calibration and selective classification help control automated coverage when uncertainty increases (Geifman & El-Yaniv, 2017; Guo et al., 2017; Zhong et al., 2023). The practical objective is to screen with a bounded candidate set, request a more informative observation, or abstain and escalate. Evidence delivery requires additional checks beyond the confidence of the visual classifier.

This study evaluates a lightweight modular image–text pipeline. Here, vision-language denotes the connection of separately trained visual and textual components through disease hypotheses and evidence identifiers; it does not denote a jointly trained multimodal model. The visual stage uses Rice Leaf Disease Image Samples, Version 2 (Sethy, 2024), with Paddy Doctor providing an external image source (Petchiammal et al., 2023). The language stage uses CROP and CROP Benchmark (Zhang et al., 2024). The images are not treated as Nepalese observations. Nepal appears only as a location filter for advisory evidence. Classification, open-set and quality triage, and advisory retrieval are evaluated separately rather than as an end-to-end agronomic service.

The investigation examines classification after perceptual grouping, sensitivity to ordinary image splitting, rejection of omitted diseases and external conditions, and selective prediction under image degradation. It also tests partially improved second views and the contribution of individual advisory gates. The contributions are a leakage-reduced image evaluation, external testing, uncertainty-aware triage, and an evidence-preserving advisory workflow. Simulated recapture and the small Nepal-context subset provide controlled and preliminary evidence, respectively; they do not establish the behavior of farmers or extension specialists using the system.

II. LITERATURE REVIEW

Plant-disease recognition has evolved from engineered descriptors toward transfer-learned convolutional networks. Mohanty et al. (2016) demonstrated discrimination of numerous crop–disease pairs on a controlled collection, while Ferentinos (2018) compared crops and network families. Too et al. (2019) examined fine-tuning for plant-disease identification. In rice, Sethy et al. (2020) combined deep features with an SVM, illustrating the value of separating representation learning from a lightweight decision layer. These studies motivate the present model comparisons,

while differences in image source and acquisition conditions motivate separate external evaluation.

Engineered features remain informative when symptoms appear in color distributions and local texture. Histograms of oriented gradients summarize local shape (Dalal & Triggs, 2005), support vector machines provide margin-based classification (Cortes & Vapnik, 1995), and extremely randomized trees model nonlinear feature interactions (Geurts et al., 2006). Lightweight neural architectures pursue computational efficiency through compact convolutional blocks. MobileNetV3 uses hardware-aware design (Howard et al., 2019), whereas EfficientNet scales depth, width, and resolution jointly (Tan & Le, 2019). These architectures provide context for a compact CNN baseline; the present comparisons focus on the behavior of inexpensive representations under clean and degraded conditions.

Vision-language and self-supervised pretraining have expanded reusable image representations. CLIP learns aligned image–text embeddings (Radford et al., 2021), DINOv2 learns visual features through self-supervision (Oquab et al., 2023), and sigmoid-loss pretraining supports large-scale image–text learning (Zhai et al., 2023). Such representation learning differs from the modular connection studied here. In the intended workflow, disease hypotheses from the visual classifier guide a separately evaluated retriever. The experiments assess visual prediction, retrieval, and rule compliance separately, without joint image–text optimization or linked image-to-advice evaluation.

Open-set recognition asks whether an observation belongs to the known label space. Scheirer et al. (2013) formalized open-space risk, and Bendale and Boulton (2016) extended open-set methods to deep networks. Practical scores include maximum softmax probability (Hendrycks & Gimpel, 2017), class-conditional Mahalanobis distance (Lee et al., 2018), and energy derived from classifier logits (Liu et al., 2020). These scores rank novelty but require validation thresholds to define rejection. Calibration connects confidence with observed correctness (Guo et al., 2017), while selective classification measures the risk incurred at a chosen coverage (Geifman & El-Yaniv, 2017; Nie & Zheng, 2023).

Conformal prediction offers a complementary set-valued view. Adaptive prediction sets vary their size with instance difficulty while targeting marginal coverage (Romano et al., 2020). Coverage guarantees rely on calibration assumptions, including exchangeability, and do not automatically extend to external or corrupted images (Angelopoulos & Bates, 2023). In a triage interface, a singleton set may support screening, while larger sets can motivate clarification or abstention. Image-quality testing is also necessary because common corruptions can substantially reduce accuracy despite strong clean-test performance (Hendrycks & Dietterich, 2019).

The language stage draws on established retrieval methods (Li, 2024). TF-IDF emphasizes discriminative terms (Salton & Buckley, 1988), BM25 adjusts term frequency for document length and saturation (Robertson & Zaragoza, 2009), and latent semantic analysis projects word statistics into a lower-dimensional space (Deerwester et al., 1990). Reciprocal-rank fusion combines rankings with different score scales (Cormack et al., 2009). ROC and precision–recall analysis characterize ranking behavior, with the latter particularly informative for rare positive classes (Davis & Goadrich, 2006; Fawcett, 2006). For agricultural advice, retrieval accuracy must be distinguished from disease consistency, contextual support, and treatment-rule compliance. None of these automatic checks substitutes for expert assessment of agronomic correctness.



Figure 1. Experimental workflow and information boundaries.

III. RESEARCH METHOD

A. Study design and data

Figure 1 shows the proposed workflow. The visual component returns a disease distribution, a prediction set, a novelty score, and image-quality flags. These signals determine whether to screen, request another photograph, or abstain. The text component retrieves CROP evidence and applies disease, Nepal-context, treatment, and numeric-rate rules before returning a passage and its evidence identifier. The three evaluation levels—classification, open-set and quality triage, and advisory retrieval—use separate experiments. No linked image–question episodes were available to measure the complete sequence.

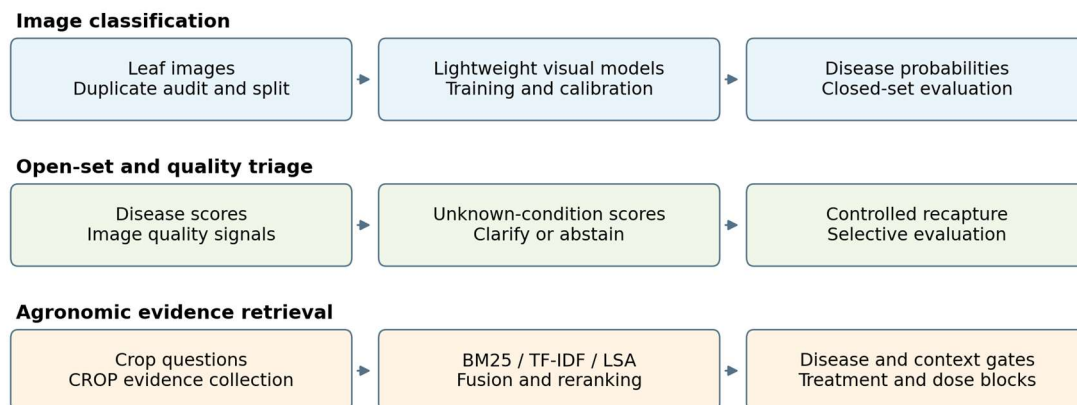


Figure 1. Modular image screening, quality and open-set triage, and context-sensitive evidence retrieval. The arrows describe the intended information flow; the three evidence levels are evaluated separately.

Rice Leaf Disease Image Samples, Version 2, was released in 2024 under CC BY 4.0 and contains 5,932 files in bacterial blight, blast, brown spot, and tungro folders (Sethy, 2024). The external resource was the 10,407-image, ten-class Paddy Doctor competition subset, first released in 2022 (Petchiammal et al., 2023). Its four overlapping disease labels were treated as known classes. Bacterial leaf streak, bacterial panicle blight, dead heart, downy mildew, hispa, and normal were treated as conditions outside the source model's label set. External labels were used only for evaluation; model parameters, calibration, and rejection thresholds remained fixed from the source data.

The 2024 CROP collection and CROP Benchmark supplied the text resources (Zhang et al., 2024). English factual and contextual question–answer records and individual conversational turns were retained when the question and answer together contained rice, paddy, *Oryza*, or Nepal terms and a disease, pest, symptom, management, or Nepal term. Summarization, entity extraction, and event extraction tasks were excluded. This scope yielded 7,745 question–answer pairs with distinct normalized answers, including 531 conversational-turn records. Of 5,045 benchmark items, 365 English questions met the same lexical criteria when applied to the question and its four options. Nepal appeared in 54 evidence answers and eight held-out questions. Table 1 summarizes the resources; the exact selection expressions are provided in the supplementary methods.

Table 1. Data resources and analytic subsets

Resource	Analyzed size	Evaluation role
Rice Leaf Disease Image Samples	5,932 images; four diseases	Source classification, open-set folds, corruption and recapture
Paddy Doctor competition subset	10,407 images; 10,333 exact-unique	External known classes (4,232) and unknown conditions (6,101)
CROP question–answer collection	7,745 English pairs; 7,606 query groups	Evidence retrieval and rule-based advisory
CROP Benchmark	5,045 total; 365 scoped; 179 without exact query overlap	Evidence-assisted multiple-choice answering
Nepal-context subset	54 evidence answers; eight test queries	Preliminary geographic-context assessment

B. Image Grouping and Partitions

Images were decoded with orientation metadata applied and converted to RGB. SHA-256 identified byte-identical files. The initial 64-bit perceptual hash used the low-frequency discrete cosine transform of a 32×32 grayscale image, joining within-label pairs at Hamming distance ≤ 4 . To detect transformations missed by this rule, hashes were also computed over all eight rotations and reflections of the square. Images were joined when their minimum transformed-hash distance was ≤ 4 , and connected components defined the final groups. Filename patterns were examined, but reused or reordered numeric suffixes prevented consistent reconstruction of

original source identities. The resulting groups are similarity groups, not verified acquisition families; crops and stronger transformations may still escape detection.

Table 2. Rotation- and reflection-aware image split

Disease	All images	Train images / groups	Validation images / groups	Test images / groups
Bacterial blight	1584	1,182 / 52	180 / 11	222 / 12
Blast	1440	1,008 / 55	216 / 12	216 / 12
Brown spot	1600	1,112 / 70	240 / 15	248 / 16
Tungro	1308	962 / 61	115 / 13	231 / 14
Total	5,932	4,264 / 238	751 / 51	917 / 54

Within each disease, groups were divided approximately 70/15/15 into training, validation, and test partitions using seed 2024. Table 2 reports the resulting image counts. All images in a group remained in one partition. Ordinary image splitting and single-orientation pHash grouping were retained only for the split-sensitivity experiment. Exposure was measured using the stronger rotation- and reflection-aware groups in all three partitions. The external set was deduplicated by exact hash and checked against the source collection using exact and transformed perceptual hashes.

C. Visual models and Evaluation

Seven lightweight classifiers were trained on fixed representations: logistic regression on low-frequency DCT, color–texture, or HOG–color features; a PCA-based linear SVM; a PCA-RBF SVM; 350 extremely randomized trees; and a whitened PCA prototype classifier. The color–texture vector contained RGB and HSV histograms, channel moments and quantiles, local binary patterns, Sobel and Laplacian statistics, and gradient orientations. The dimensions were 194 for color–texture, 1,958 for HOG–color, and 143 for DCT. Standardization and PCA were fitted on training data. Validation temperature scaling was applied to the SVM and prototype scores. Complete descriptors and classifier settings are given in the supplementary methods.

The neural baseline accepted 32×32 RGB images and used three convolutional blocks with 16, 32, and 64 channels, batch normalization, ReLU, and max pooling, followed by global average pooling, 0.15 dropout, and a four-way output layer. It contained 24,724 parameters. AdamW training used learning rate 0.003, weight decay 0.0001, cosine annealing, class-weighted cross-entropy, and 0.04 label smoothing. Horizontal flips and brightness gains of 0.9–1.1 were applied during training. The best validation-macro-F1 checkpoint was selected from at most 24 epochs, followed by validation temperature scaling.

Evaluation included accuracy, balanced accuracy, macro-F1, Matthews correlation, log loss, Brier score, and 15-bin expected calibration error (ECE). Source-test confidence intervals used 200

bootstrap resamples of the 54 test groups, preserving dependence among transformed images. Per-disease recall intervals used the corresponding disease groups. Timings measured batched CPU prediction on prepared inputs and excluded decoding and feature extraction; they are computational comparisons rather than device response times. Learning curves used five group samples at each of 10%, 25%, and 50% of training groups, with the fixed complete training set at 100%. PCA visualization was fitted on training representations and applied to the held-out test set.

D. Open-set and external Evaluation

Four leave-one-disease-out folds trained a three-class color–texture logistic model after excluding one disease from source training and validation. Only the held-out partition of the omitted disease served as the fold’s unknown set; known test images came from the other three diseases. Scores were one minus maximum probability, predictive entropy, negative log-sum-exp energy, shared-covariance Mahalanobis distance with Ledoit–Wolf shrinkage, and a prototype-distance conformal novelty score. Each rejection threshold was the 90th percentile of the known validation scores. Unknown-class AUROC, AUPR, recall, and known coverage were calculated by fold. Group-bootstrap intervals and variation across the four folds quantify disease dependence. These folds test unfamiliar diseases within one collection, not general unknown-condition detection.

External evaluation used both known and unknown examples from Paddy Doctor, so source identity alone could not distinguish the two groups. All source models were frozen. The four-class color–texture model was paired with Mahalanobis rejection; the CNN used one minus its maximum calibrated probability. External accuracy was evaluated only for the four overlapping diseases, and novelty ranking compared these cases with the six other labels. External proportions use Wilson intervals; AUROC and macro-F1 use 1,000 image-bootstrap samples. These describe variation within the external subset, whose original acquisition-family identifiers were unavailable. Healthy leaves and insect-associated conditions broaden the test, but nutrient deficiencies, verified mixed infections, and unrelated imagery remain outside its coverage.

E. Metrics Quality Thresholds and Simulated Recapture

The mixed-quality experiment assigned validation and test images cyclically to clean, Gaussian blur (radius 2), dark exposure (brightness $\times 0.55$), JPEG compression (quality 25), or 15° rotation. Corruptions were formed at 96×96 pixels. The color–texture triage model used a separate validation calibration: 24 groups (360 images) fitted a scalar temperature of 3.614, and the remaining 27 groups (391 images) calibrated label-conditional prediction sets. For class c , the nonconformity score was $1 - p(c)$; its threshold was the ordered score at rank $\text{ceil}[0.95(n_c + 1)]$, capped at n_c . A label entered the set when its score did not exceed the class threshold. Thresholds

for bacterial blight, blast, brown spot, and tungro were 0.6403, 0.3959, 0.8373, and 0.3682. Screening required a singleton set and the confidence and quality conditions below. These are empirical calibration sets; group dependence and distribution shift limit nominal coverage guarantees.

Brightness and sharpness flags used the fifth percentiles of mean grayscale intensity and Laplacian variance in the 751 clean validation images: 0.296669 and 0.0005082, respectively, on the $[0, 1]$ intensity scale. Confidence candidates were 0.50–0.95 in steps of 0.05 and 0.99. For each policy, the lowest candidate accepting at least 20 mixed-quality validation cases with empirical error $\leq 2\%$ was selected. Confidence-only rejection had no feasible candidate and therefore abstained on all cases. The full quality–singleton–confidence rule selected 0.95, accepting 53/751 validation cases with no errors. Thresholds were then frozen. Cases failing the full rule requested one second view, and cases still failing the same rule were escalated.

Three second-view conditions tested sensitivity to recapture quality: the clean held-out image of the same leaf; partial improvement using blur radius 1, brightness $\times 0.80$, JPEG quality 60, or 7.5° rotation; and the unchanged first view. A first view already assigned clean remained clean. These controlled transformations do not represent repeated photographs taken by users. The clean condition is a favorable successful-recapture scenario. Coverage is the fraction receiving a class, and selective risk is the error fraction among those accepted. Post-clarification accuracy uses only accepted second views, so its denominator can differ substantially from overall coverage. Confidence-ranked risk–coverage curves and corruption-specific accuracy were also calculated.

F. Evidence Retrieval and Advisory Rules

Normalized identical questions or answers were joined before a seeded 70/15/15 group partition, giving 5,421 training, 1,156 validation, and 1,168 test queries. All 7,745 answer documents remained searchable; only training queries supplied relevance labels for the reranker. Contextual source passages and conversational histories were excluded from queries. Retrieval compared word TF-IDF, character TF-IDF, BM25, 160-dimensional latent semantic analysis, reciprocal-rank fusion, and a logistic lexical reranker. The reranker used 2,500 training queries and reordered the top 60 fused candidates using four retrieval scores, token overlap, length ratio, and disease-set agreement. Corpus-defined relevance was the original paired answer. $\text{Recall}@1/3/5$, reciprocal rank, and $\text{nDCG}@5$ therefore measure answer recovery rather than exhaustive expert-adjudicated relevance. The supplementary methods specify tokenization, feature limits, and all retrieval settings.

Multiple-choice options were scored by word-TFIDF similarity to the question alone or to a reciprocal-log-rank-weighted combination of the five retrieved answers. Exact normalized question matches between the benchmark and the retrieval collection were audited. Matching questions were excluded from the supervised reranker training pool. Accuracy was reported for the full scoped benchmark and for questions without an exact match; both belong to the CROP resource family.

Advisory tests used disease mentions in the text queries as hypotheses; image predictions were not attached to question–answer pairs. Four rules inspected the same five reranked candidates. The disease rule required identical recognized-disease sets when a question named a known disease. The Nepal rule required an explicit Nepal mention when the question requested that context. The treatment rule blocked chemical-treatment expressions when the query named none of the four known diseases. The numeric rule blocked rates or concentrations in a pesticide or fertilizer context. Each policy returned the highest-ranked passing answer or abstained. The ablation compared no gates, all gates, and removal of each gate separately. Delivery coverage, paired-answer delivery, and rule-specific violations measured the effects. The renderer retained the selected answer and its evidence identifier; the matching rules did not adjudicate agronomic correctness. Nepal proportions are reported with raw counts and Wilson 95% intervals.

IV. RESULTS AND FINDINGS

A. Duplicate Audit and Split Sensitivity

The source collection contained 4,794 exact-unique files and 1,138 byte-identical duplicates. Single-orientation pHash grouping yielded 2,135 components, with 5,597/5,932 images (94.35%) in multi-image groups. Accounting for rotations and reflections reduced this to 343 components, all containing multiple images; the largest contained 62. Representative transformed pairs had original-hash distances of 30–36 but transformed distances of zero. The supplementary audit montage documents these pairs. Filename suffixes did not reliably identify source families. A nearest-neighbor inspection after splitting also found a visually similar reframed tungro pair at distance 6, beyond the chosen threshold. Thus, even the stronger grouping remains leakage-reduced, not demonstrably leakage-free.

Table 3. Split sensitivity of HOG–color logistic regression

Partition	Train / test images	Accuracy	Macro-F1	Transformed-group exposure
Rotation/reflection grouping	4,264 / 917	76.12%	0.747	0.00%
Single-orientation pHash	4,126 / 904	86.84%	0.867	100.00%
Ordinary image split	4,152 / 890	98.88%	0.989	100.00%

Table 3 shows 76.12% HOG–color accuracy under rotation/reflection grouping, 86.84% under single-orientation pHash grouping, and 98.88% under ordinary image splitting. Both weaker

partitions exposed every test image to a transformed-group relative in training. The 22.76-percentage-point difference between ordinary and stronger grouping is therefore consistent with substantial dependence on familiar image families. Because the test compositions differ, it should not be interpreted as a controlled causal estimate of leakage alone.

Table 4. Grouped-test comparison of lightweight visual models

Model	Accuracy (95% CI)	Macro-F1	ECE	Model MB
DCT linear probe	34.02% (22.73–45.13)	0.290	0.261	0.009
Color-texture LR	78.41% (64.96–91.47)	0.774	0.182	0.012
HOG-color LR	76.12% (64.74–88.83)	0.747	0.215	0.111
Calibrated linear SVM	73.83% (61.42–86.34)	0.726	0.107	0.561
PCA-RBF SVM	77.97% (66.02–89.92)	0.766	0.019	1.481
Extra Trees	76.23% (64.03–89.93)	0.752	0.150	6.393
Prototype classifier	70.99% (59.74–84.57)	0.697	0.093	0.559
Compact CNN-32	81.03% (69.42–91.52)	0.800	0.132	0.107

An exploratory threshold-8 audit yielded 320 groups and a new 4,227/767/938-image split. Color–texture and HOG–color accuracies were 78.89% and 76.76%, respectively, similar to the principal results. This sensitivity check reduces concern about the selected hash threshold dominating the conclusion, but the changed test composition prevents a paired comparison and does not establish complete source separation.

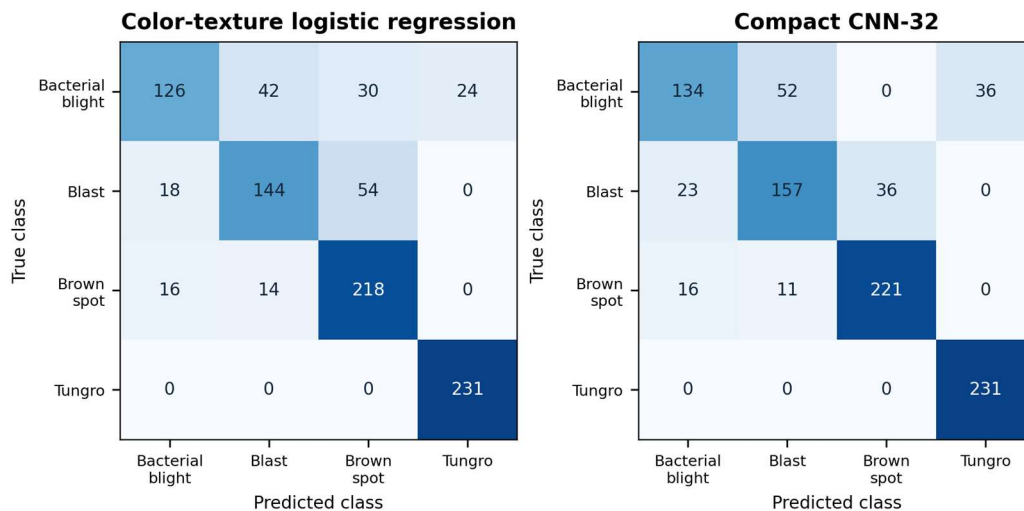


Figure 2. Grouped-test confusion matrices for color–texture logistic regression and the compact CNN. Counts are based on the same 917 held-out images.

B. Closed-set Performance and Calibration

The compact CNN achieved the highest point accuracy, 81.03%, compared with 78.41% for color–texture logistic regression and 77.97% for the PCA-RBF SVM (Table 4). Their broad group-bootstrap intervals overlap, so the single split does not establish a reliable ordering. The color–texture model had ECE 0.182, whereas the PCA-RBF SVM had ECE 0.019. Classification

accuracy and calibration therefore favored different models. The 34.02% DCT result indicates that the low-frequency global representation alone was insufficient. Table 5 and Figure 2 show the uneven class behavior: color–texture recall ranged from 56.76% for bacterial blight to 100% for tungro. These within-source differences do not predict external transfer.

Table 5. Disease-specific color–texture performance

Disease	Correct / n	Precision	Recall (95% CI)	F1
Bacterial blight	126/222	78.75%	56.76% (29.39–88.58)	0.660
Blast	144/216	72.00%	66.67% (41.67–91.67)	0.692
Brown spot	218/248	72.19%	87.90% (71.87–99.62)	0.793
Tungro	231/231	90.59%	100.00% (100.00–100.00)	0.951

Figure 3 compares empirical accuracy with model confidence. Calibration varied across models, and low validation error did not guarantee low test error. In particular, the CNN retained ECE 0.132, and the unscaled color–texture probabilities were often overconfident. These calibration differences motivate evaluating rejection policies on held-out cases rather than assuming that a high maximum probability is sufficient.

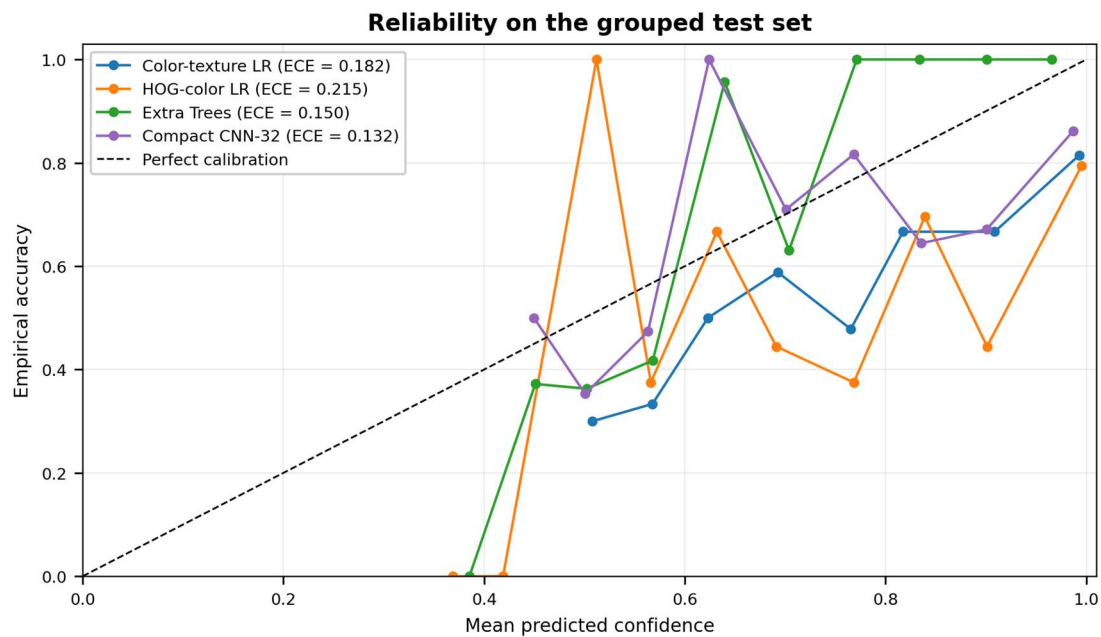


Figure 3. Reliability of the visual models on the grouped test set. The diagonal indicates agreement between confidence and observed accuracy; empty bins are omitted.

C. Within-Source Open-Set Behavior

Mahalanobis distance had the highest mean AUROC, 0.727 ± 0.104 , but its mean unknown recall at the validation threshold was only 21.93%, with 89.67% known coverage (Table 6). Maximum-probability rejection recovered more unknowns on average, 32.55%, despite a lower mean AUROC. Ranking quality and operating-point sensitivity therefore differed. Table 7 and Figure 4 show marked disease dependence. Mahalanobis AUROC was 0.822 for omitted blast and 0.582

for omitted tungro; no tungro unknown was rejected at the selected threshold. The wide fold-specific intervals and four-fold design limit conclusions about other unfamiliar conditions.

Table 6. Open-set performance across four omitted-disease folds

Novelty score	AUROC mean $\pm$ SD	AUPR mean	Known coverage	Unknown recall
Conformal prototype	0.677 $\pm$ 0.127	0.395	87.34%	28.41%
Energy	0.688 $\pm$ 0.117	0.424	90.37%	26.00%
Mahalanobis	0.727 $\pm$ 0.104	0.416	89.67%	21.93%
Maximum softmax	0.688 $\pm$ 0.096	0.428	91.18%	32.55%
Predictive entropy	0.713 $\pm$ 0.107	0.439	91.57%	25.32%

The principal-component projection in Figure 5 illustrates overlap among test color–texture representations. It provides a view of within-source geometry; separation in this projection neither confirms disease-specific biological features nor establishes transfer across acquisition settings.

Table 7. Disease-specific Mahalanobis open-set results

Omitted disease	Unknown images / groups	AUROC (95% CI)	Unknown recall (95% CI)	Known coverage
Bacterial blight	222 / 12	0.724 (0.584–0.875)	0.90% (0.00–2.70)	91.65%
Blast	216 / 12	0.822 (0.671–0.936)	41.67% (16.67–75.00)	87.59%
Brown spot	248 / 16	0.778 (0.618–0.889)	45.16% (25.00–70.97)	87.44%
Tungro	231 / 14	0.582 (0.343–0.747)	0.00% (0.00–0.00)	91.98%

D. Selective Triage and Recapture Sensitivity

Forced classification made 414 errors among 917 mixed-quality first views (45.15% risk). Confidence-only rejection could not meet the validation target with the required support and accepted no test cases. The full quality–singleton–confidence rule accepted 86 first views, of which 73 were correct. With clean second views, coverage increased to 197/917 (21.48%) and risk was 13/197 (6.60%); with partial improvement, coverage was 136/917 (14.83%) and risk was 18/136 (13.24%). An unchanged view added no accepted cases. Table 8 and Figure 6 show that recapture benefit depended on how much image quality improved.

Table 8. Selective triage with different second-view quality

Policy / second view	Accepted / 917	Risk	Requests	Accepted recapture correct / n	Escalation
Forced classification	917 (100.00%)	45.15%	0	—	0.00%
Confidence-only rejection	0 (0.00%)	—	0	—	100.00%
Full rule + clean recapture	197 (21.48%)	6.60%	831	111/111	78.52%
Full rule + partial improvement	136 (14.83%)	13.24%	831	45/50	85.17%
Full rule + unchanged view	86 (9.38%)	15.12%	831	—	90.62%

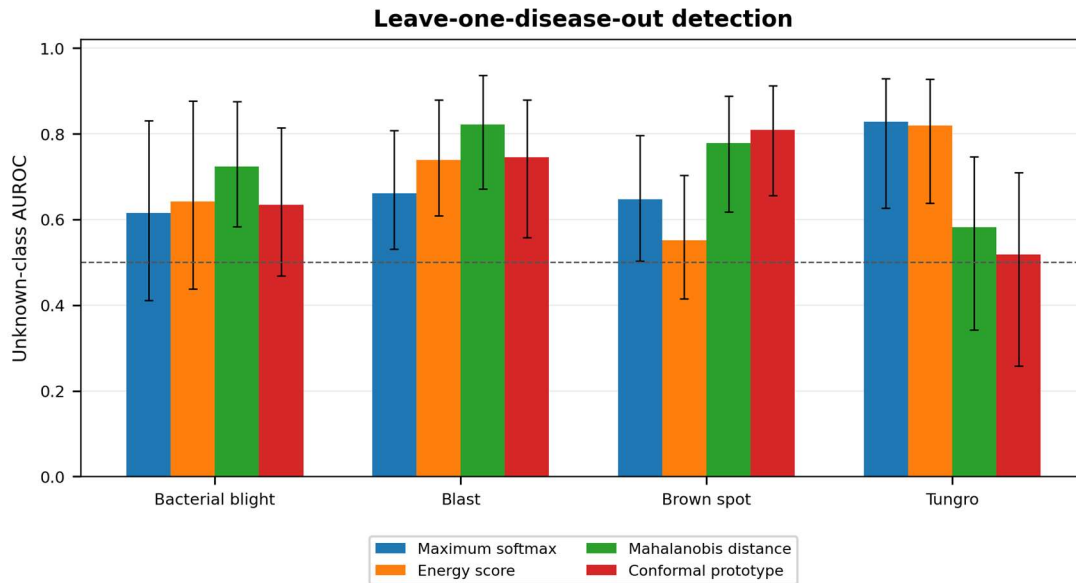


Figure 4. Unknown-detection AUROC across the four omitted-disease folds. Error bars show group-bootstrap 95% intervals. Four scores are plotted; predictive entropy is included in Table 6.

All 111 accepted clean recaptures were correct, compared with 45/50 accepted partially improved recaptures. The 100% clean-recapture value is therefore conditional on selection in a favorable same-image scenario, not overall system accuracy or a measured user-recapture success rate. Before gating, accuracy across all 917 second views was 78.41% for clean images, 70.77% for partial improvement, and 54.85% for unchanged images. Moreover, the 2% validation target did not hold on the test set. For the nominal 95% prediction sets, empirical coverage fell from 77.64% on clean test images to 56.92% on mixed-quality images, reinforcing the limits of validation calibration under shift.

Table 9. Accuracy under image-quality stress

Model	Clean	Blur	Dark	JPEG	Rotation
Color-texture LR	78.41%	34.68%	35.11%	62.70%	64.34%
HOG-color LR	76.12%	52.02%	27.92%	71.32%	68.27%
Extra Trees	76.23%	65.32%	33.15%	76.23%	63.14%
Compact CNN-32	81.03%	56.71%	35.55%	80.59%	76.55%

Table 9 and Figure 7 show different failure patterns across representations. Color-texture accuracy fell to 34.68% under blur and 35.11% under darkness. Extra Trees retained 65.32% under blur, while the CNN reached 56.71%; neither maintained its clean performance. Darkness reduced all four tested models to 27.92–35.55%. These results support examining image quality before interpreting predictions, but they do not establish that routing between the tested models would resolve the errors.

E. Training Coverage and Computational Cost

The learning curve in Table 10 and Figure 8 improved with broader training coverage but did not approach perfect classification. At 10% of the training groups, mean accuracy was 66.96%; the

complete 238-group training set reached 78.41%. Variation across partial subsets indicates sensitivity to which similarity groups are observed. Additional transformations of existing photographs should therefore not be treated as equivalent to collecting independent disease presentations.

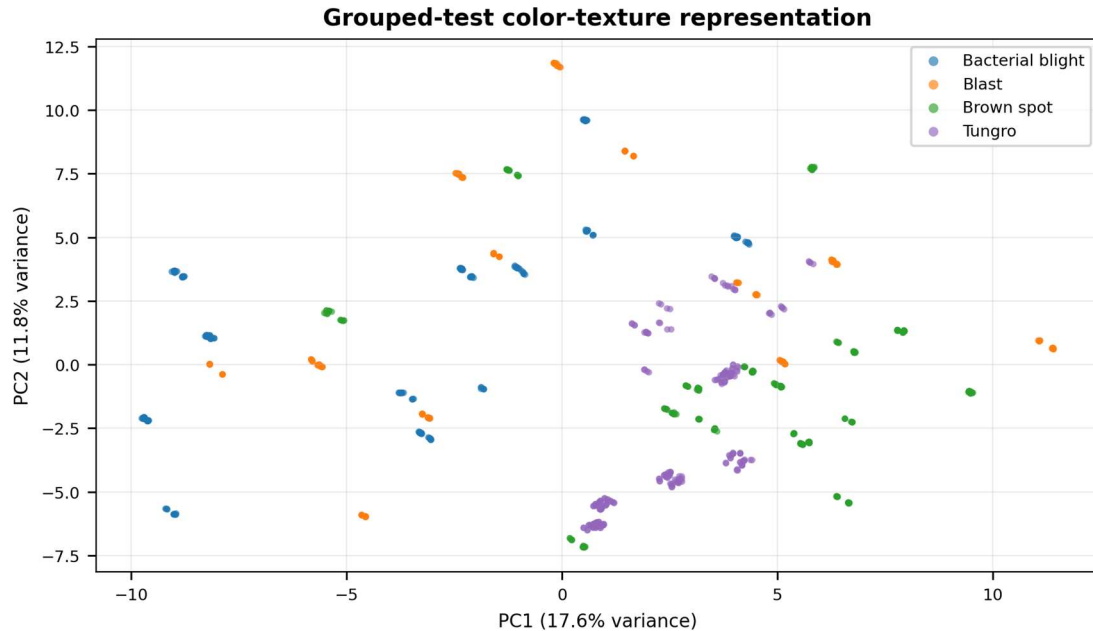


Figure 5. Projection of grouped-test color–texture features onto the first two principal components fitted on source training data.

Figure 9 compares accuracy, stored model size, and batched prediction time. The color–texture model occupied 0.012 MB and the CNN 0.107 MB, while Extra Trees occupied 6.393 MB. Classical timings exclude feature extraction and the CNN timing excludes image preprocessing, so the plot does not establish an end-to-end latency advantage. Small models are computationally accessible, but their low storage cost does not compensate for the generalization failures observed below.

Table 10. Group-based learning curve for color–texture logistic regression

Training fraction	Mean images	Groups	Accuracy mean $\pm$ SD	Macro-F1 mean $\pm$ SD
10%	409.6	24	66.96% $\pm$ 5.02	65.82% $\pm$ 4.34
25%	1100.6	60	73.81% $\pm$ 3.60	72.51% $\pm$ 3.66
50%	2117.8	119	73.96% $\pm$ 0.88	72.96% $\pm$ 1.15
100%	4264.0	238	78.41% $\pm$ 0.00	77.38% $\pm$ 0.00

F. External known classes and unknown conditions

After removing 74 exact duplicates, Paddy Doctor contributed 4,232 known-class and 6,101 unknown-condition images. No exact or transformed-hash match at distance ≤ 4 was found against the source collection. Nevertheless, color–texture accuracy was only 564/4,232 (13.33%) and CNN accuracy was 531/4,232 (12.55%) (Table 11; Figure 10). For color–texture classification,

recall was 417/471 for bacterial leaf blight, 7/1,728 for blast, 55/953 for brown spot, and 85/1,080 for tungro. Most cases were assigned bacterial blight. Across all seven classical comparators, external accuracy ranged from 12.48% to 24.24%.

Table 11. Frozen-model performance on Paddy Doctor

Model / novelty score	Known accuracy	Known macro-F1	Unknown AUROC	Known retained	Unknown rejected
Color–texture / Mahalanobis	564/4,232 (13.33%)	0.105	0.512	4,185/4,232 (98.89%)	337/6,101 (5.52%)
Compact CNN / nonconfidence	531/4,232 (12.55%)	0.111	0.486	3,535/4,232 (83.53%)	942/6,101 (15.44%)

Mahalanobis unknown-detection AUROC was 0.512 (95% CI, 0.501–0.523), and only 337/6,101 unknown images were rejected at the source threshold. The score rejected 112/1,749 normal images, 133/1,589 hispa cases, and 27/1,429 dead-heart cases. CNN nonconfidence was similarly weak in ranking external unknowns (AUROC 0.486). These tests directly demonstrate failure to transfer beyond the source collection. Differences in background, framing, acquisition, and symptom visibility are plausible contributors, but the experiments do not isolate their causal effects.

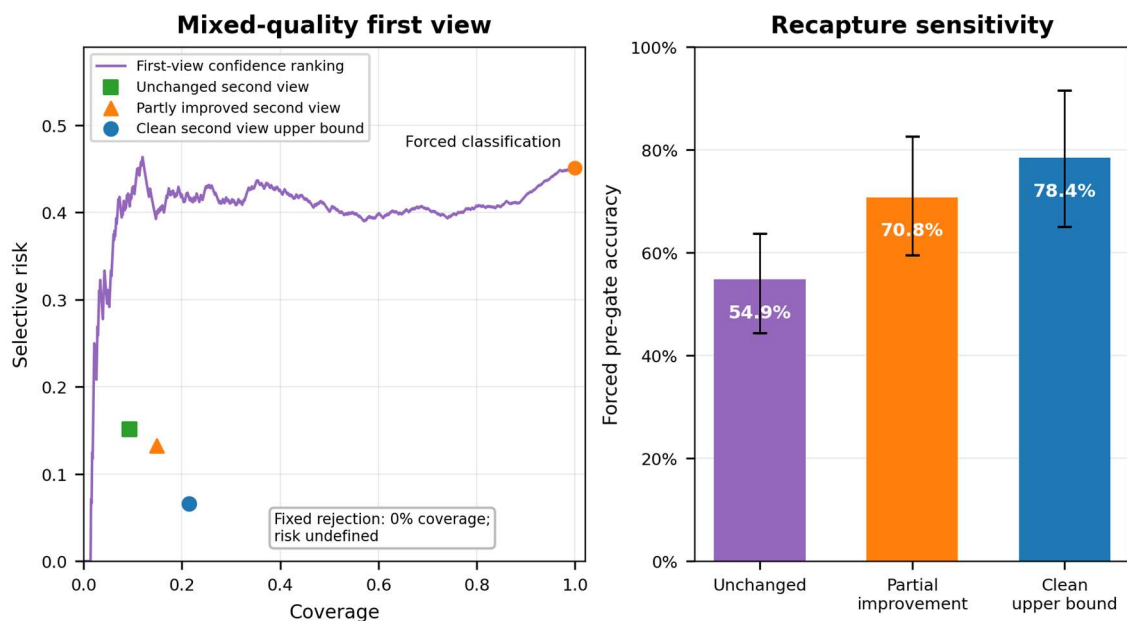


Figure 6. First-view confidence-ranked risk–coverage behavior and recapture sensitivity. Markers show the full-rule operating points. Bars show forced second-view accuracy before gating with group-bootstrap 95% intervals; the clean condition is a favorable recapture simulation.

G. Evidence retrieval and Nepal context

The lexical reranker recovered the paired answer among five candidates for 934/1,168 queries (79.97%), compared with 919/1,168 (78.68%) for BM25 (Table 12; Figure 11). BM25 had slightly higher rank-one recall, 60.87% versus 60.45%. LSA had the lowest Recall@5, 46.66%, despite high agreement on recognized disease names. Disease consistency and retrieval of the

paired evidence are therefore distinct measures. Among eight Nepal queries, BM25, word TF-IDF, and reciprocal-rank fusion each achieved 8/8; character TF-IDF, LSA, and the reranker each achieved 7/8. With only 54 Nepal-bearing answers and eight held-out queries, these estimates remain preliminary and do not establish broadly representative Nepal-context performance.

Table 12. Evidence retrieval over 1,168 held-out CROP queries

Retriever	R@1	R@3	R@5 (count)	MRR	nDCG@5	Disease agreement	Nepal R@5
BM25	60.87%	74.83%	78.68% (919/1,168)	0.690	0.707	96/115	8/8
Word TF-IDF	53.25%	68.92%	73.54% (859/1,168)	0.626	0.643	102/115	8/8
Character TF-IDF	52.91%	70.03%	75.77% (885/1,168)	0.633	0.654	97/115	7/8
LSA	27.74%	40.58%	46.66% (545/1,168)	0.371	0.377	113/115	7/8
RRF	53.17%	70.29%	76.03% (888/1,168)	0.636	0.656	109/115	8/8
Lexical reranker	60.45%	75.51%	79.97% (934/1,168)	0.693	0.711	104/115	7/8

CROP evidence retrieval on 1,168 held-out queries

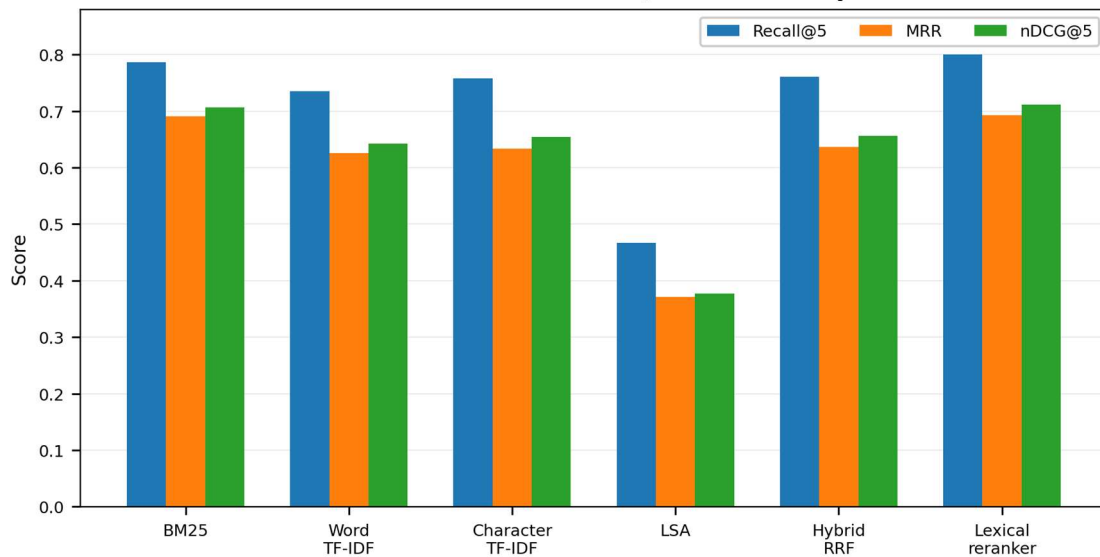


Figure 11. CROP retrieval effectiveness. Recall@5, reciprocal rank, and nDCG@5 evaluate recovery and ranking of the paired answer; they do not measure treatment appropriateness.

H. Benchmark overlap and advisory ablation

The overlap audit found exact normalized matches for 186 of 365 benchmark questions. On the complete set, the lexical baseline answered 90 correctly (24.66%), BM25 and fused evidence each answered 233 correctly (63.84%), and reranked evidence answered 238 correctly (65.21%). On the 179 questions without an exact match, the counts were 53 (29.61%), 75 (41.90%), 79 (44.13%), and 74 (41.34%), respectively (Table 13). The lower nonmatching-subset scores show

the importance of corpus overlap when interpreting the complete benchmark. Evidence remained useful relative to the lexical baseline, but the reranker did not outperform simpler retrieval on that subset.

Table 13. Evidence-assisted multiple-choice performance

Evidence method	All benchmark questions	Accuracy	No exact question match	Accuracy
Question-option lexical	90/365	24.66%	53/179	29.61%
BM25 evidence	233/365	63.84%	75/179	41.90%
RRF evidence	233/365	63.84%	79/179	44.13%
Lexical reranker evidence	238/365	65.21%	74/179	41.34%

Without gates, all 1,168 queries received evidence, with 11 disease mismatches, 1 missing Nepal mention, 69 out-of-scope chemical-treatment flags, and 14 numeric-rate flags. The full policy delivered 1161/1,168 answers (99.40%) with no violations of its own four rules. Removing the disease, Nepal, treatment, or numeric gate separately produced 11, 1, 68, and 11 corresponding violations (Table 14). However, paired-answer delivery decreased from 706/1,168 (60.45%) to 660/1,168 (56.51%). High delivery coverage therefore concealed some loss of reference evidence.

Table 14. Contribution of the four advisory gates

Gate policy	Delivered / 1,168	Paired answer / 1,168	Disease flags	Nepal flags	Treatment flags	Numeric flags
No gates	1168 (100.00%)	706 (60.45%)	11	1	69	14
Full	1161 (99.40%)	660 (56.51%)	0	0	0	0
Full minus disease	1162 (99.49%)	666 (57.02%)	11	0	0	0
Full minus Nepal	1161 (99.40%)	661 (56.59%)	0	1	0	0
Full minus treatment	1167 (99.91%)	688 (58.90%)	0	0	68	0
Full minus numeric	1162 (99.49%)	667 (57.11%)	0	0	0	11

All eight Nepal queries received a passage mentioning Nepal under the full gate, but only five received their paired answer. Three paired answers did not repeat the location stated in the question. This small example illustrates the difference between matching geographic wording and selecting locally appropriate evidence. Zero violations follow from the blocking definitions and should not be interpreted as zero agronomic error.

I. Discussion

The findings require three separate interpretations. First, image classification measures discrimination within a leakage-reduced source split; residual relatives and broad group-bootstrap intervals limit precision. Second, open-set and quality triage measure rejection and selection under specified unfamiliar-disease and corruption conditions. These tests do not validate nutrient-deficiency recognition, mixed-infection diagnosis, or actual recapture behavior. Third, retrieval and advisory experiments measure corpus-defined answer recovery and compliance with explicit text rules. Because these datasets do not contain linked plant, image, question, and treatment episodes, the results cannot be combined into an end-to-end estimate of agronomic accuracy or safety.

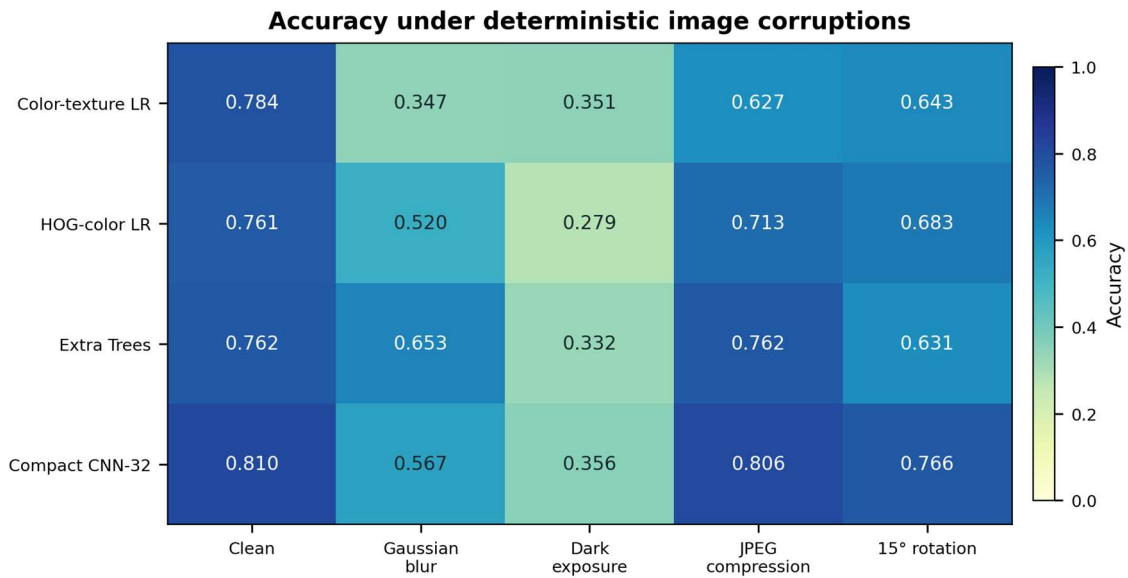


Figure 7. Grouped-test accuracy across clean and corrupted conditions for the four models used in the quality experiment.

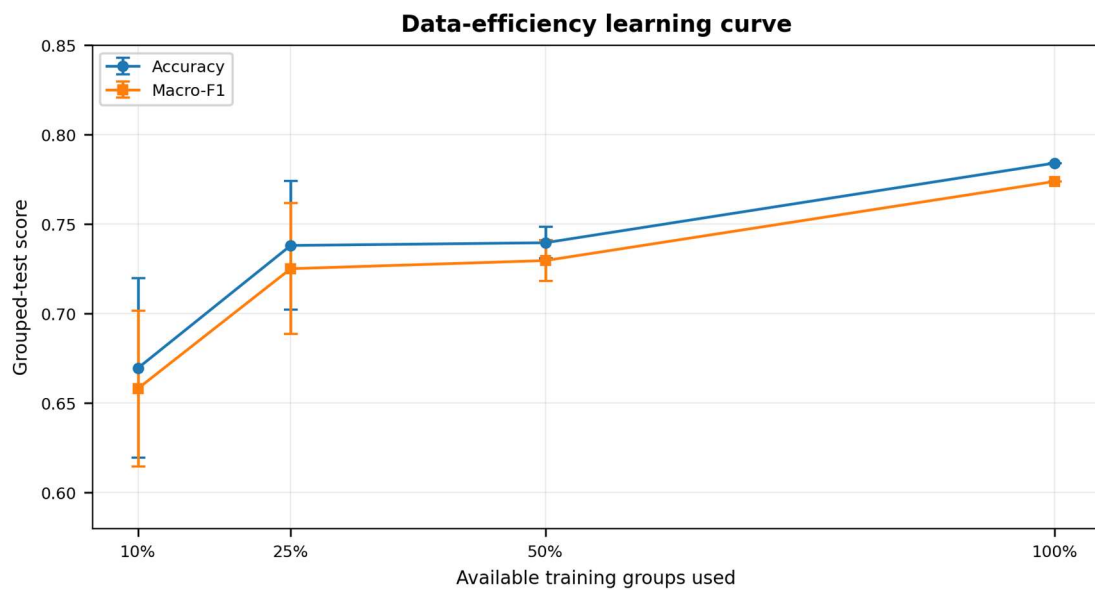


Figure 8. Accuracy and macro-F1 as the number of training groups increases. Error bars summarize five sampled subsets at each partial fraction; the full-data result is a single fixed training set.

The external experiment is the clearest limit on generalization. Both the color–texture classifier and the CNN failed on the external known classes, and their novelty scores provided little protection against the six other conditions. Strong within-source separation can therefore coexist with weak field-image transfer. Acquisition and background cues are plausible contributors, particularly for global color–texture features, but additional controlled image analysis would be

needed to attribute the errors to a specific cue. The results favor collecting independent farms, devices, seasons, and source identifiers before considering field deployment.

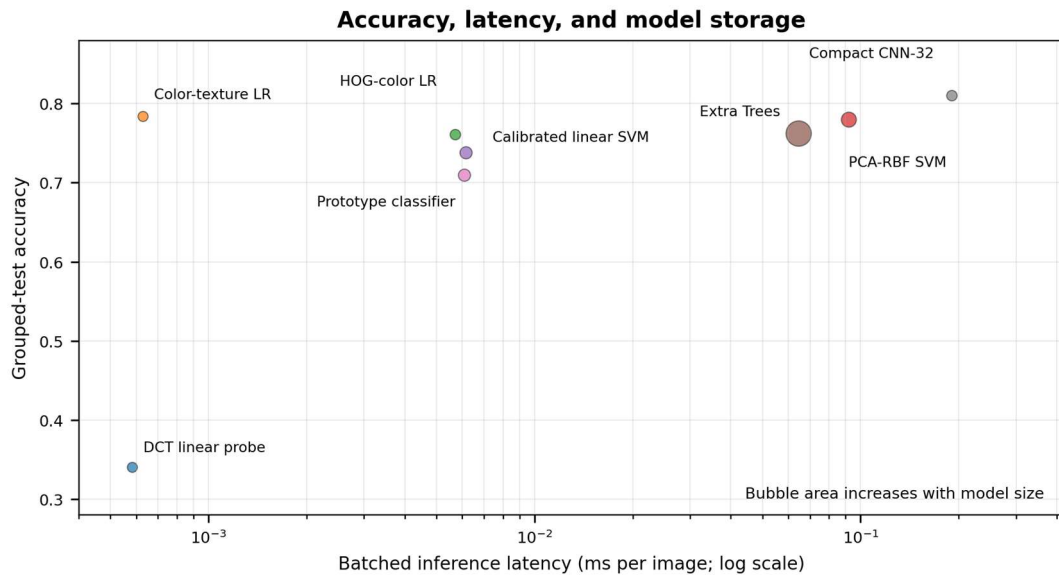


Figure 9. Grouped-test accuracy, batched CPU prediction time, and serialized model size. Timing excludes preprocessing and uses different prepared inputs for feature-based models and the CNN.

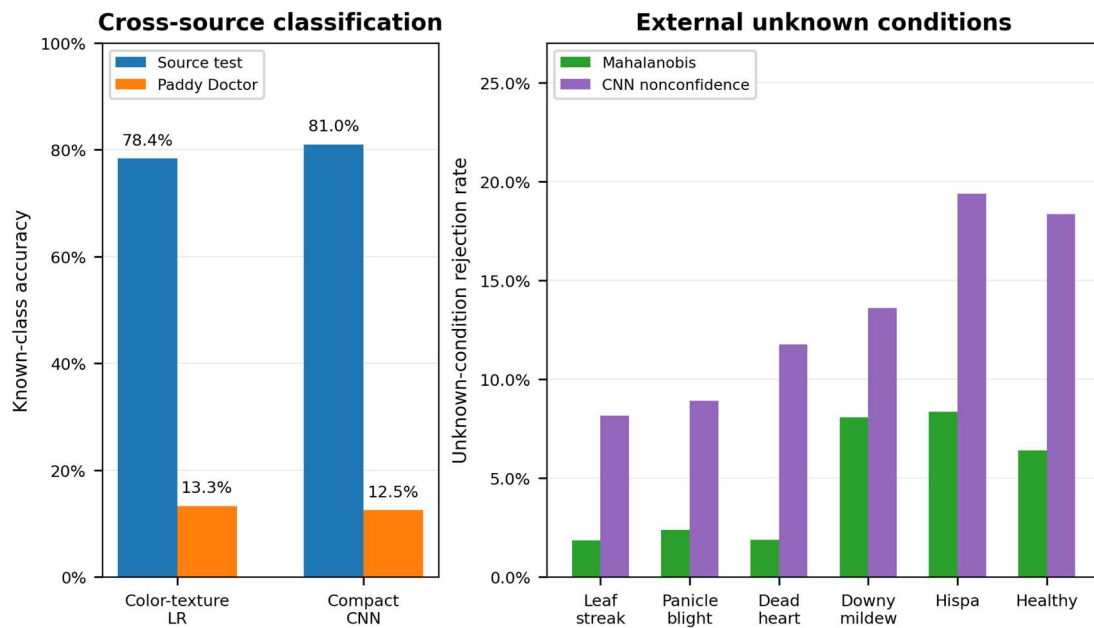


Figure 10. Transfer from the source collection to Paddy Doctor and external unknown-condition performance. Models and rejection thresholds were frozen before external evaluation.

The recapture experiment likewise supports a conditional conclusion. A clean second view helped some cases that the quality and uncertainty rules initially withheld, but partial improvement accepted fewer cases and increased selective risk. The 111/111 correct accepted clean recaptures describe an upper-bound successful-recapture scenario under selection. They do not estimate how

often a farmer would obtain such a photograph. Validation-selected thresholds also failed to maintain the intended risk on held-out groups, so confidence calibration and prediction sets should be treated as operating aids whose transfer must be tested.

Advisory retrieval introduces another distinct limitation. CROP is a broad crop-science resource, and a lexical disease or Nepal match does not establish that a passage supports a particular treatment in a local setting. The ablation shows what each blocking rule removes and the accompanying loss of paired-answer delivery. Benchmark overlap and the eight-query Nepal evaluation further constrain interpretation. Authoritative local guidance and independent review by extension specialists would be necessary to evaluate treatment suitability and the consequences of wrong image hypotheses.

Several limitations remain. The rotation/reflection audit does not guarantee invariance to crops or strong reframing, and original acquisition-family identifiers could not be reconstructed reliably. Paddy Doctor broadens the unknown set to normal and insect-associated conditions but does not cover all relevant stresses, mixtures, or unrelated scenes. External intervals treat images as sampling units because reliable source families were unavailable, so residual correlation may narrow them. Recapture uses deterministic alterations of the same source image. Text relevance is the original paired answer rather than a comprehensive expert judgment, and the advisory rules can miss paraphrased treatment or dose instructions. No field study in Nepal, linked image-to-advice trial, or evaluation of farmer outcomes was conducted.

V. CONCLUSION AND RECOMMENDATION

Lightweight rice disease models showed moderate discrimination after a stronger transformation-aware audit, with 78.41% accuracy for color–texture logistic regression and 81.03% for a compact CNN. Ordinary image splitting produced a much more favorable result while exposing test images to transformed training relatives. The external accuracies of 13.33% and 12.55%, together with weak novelty ranking, show that the source results do not support reliable field use.

Quality and uncertainty rules reduced automated coverage and allowed evaluation of when an improved view might help. Their effectiveness depended on recapture quality, and validation risk did not transfer consistently to test groups. Reranked evidence retrieval achieved 79.97% Recall@5, while benchmark overlap, paired-answer losses under gating, and the small Nepal subset limited advisory claims. The study thus provides a framework for examining failure across classification, triage, and evidence delivery rather than a validated diagnostic or treatment service. Further development should prioritize independent image acquisition with retained source identities, broader unknown conditions, real repeated captures, and linked evaluation of image

errors and advisory decisions. The text component should use authoritative Nepal-specific guidance and expert assessment of treatment support. These steps are necessary to determine whether the modular workflow can provide useful screening and referral support in practice.

DATA AVAILABILITY STATEMENT

Rice Leaf Disease Image Samples, Version 2, is available from Mendeley Data (Sethy, 2024; <https://doi.org/10.17632/fwcj7stb8r.2>). Paddy Doctor is available at <https://paddydoc.github.io/>; the 10,407-image competition subset is also distributed at <https://huggingface.co/datasets/anthony2261/paddy-disease-classification> (Petchiammal et al., 2023). CROP and CROP Benchmark are available at <https://huggingface.co/AI4Agr> and https://github.com/RenqiChen/The_Crop (Zhang et al., 2024). Analysis scripts, partition manifests, prediction files, and tabulated results accompany the supplementary materials.

AI USAGE AND GENERATIVE AI DECLARATION

OpenAI ChatGPT was used to assist with language editing, literature checking, analysis-code development, and figure preparation. Responsibility for the study, its interpretation, and the final manuscript remains with the authors.

REFERENCES

- Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/22000000101>
- Bendale, A., & Boulton, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1563–1572). <https://doi.org/10.1109/CVPR.2016.173>
- Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). <https://doi.org/10.1145/1571941.1572114>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 886–893). <https://doi.org/10.1109/CVPR.2005.177>
- Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). <https://doi.org/10.1145/1143844.1143874>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII1>3.0.CO;2-9)

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318. <https://doi.org/10.1016/j.compag.2018.01.009>
- Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30). <https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1903.12261>
- Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1610.02136>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314–1324). <https://doi.org/10.1109/ICCV.2019.00140>
- Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 7167–7177). https://proceedings.neurips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html
- Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems*, 4(7), 65–82. <https://doi.org/10.69987/JACS.2024.40706>
- Liu, W., Wang, X., Owens, J., & Li, Y. (2020). Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 21464–21475). https://proceedings.neurips.cc/paper_files/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, Article 1419. <https://doi.org/10.3389/fpls.2016.01419>
- Nie, J., & Zheng, D. (2023). Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal. *Journal of Advanced Computing Systems*, 3(1), 66–80. <https://doi.org/10.69987/JACS.2023.30105>
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., . . . Bojanowski, P. (2023). DINOv2:

- Learning robust visual features without supervision. arXiv. <https://doi.org/10.48550/arXiv.2304.07193>
- Petchiammal, Kiruba, B., Murugan, & Arjunan, P. (2023). Paddy Doctor: A visual image dataset for automated paddy disease classification and benchmarking. In Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (pp. 203–207). Association for Computing Machinery. <https://doi.org/10.1145/3570991.3570994>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748–8763). PMLR. <https://proceedings.mlr.press/v139/radford21a.html>
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- Romano, Y., Sesia, M., & Candès, E. (2020). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 3581–3591). https://proceedings.neurips.cc/paper_files/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7), 1757–1772. <https://doi.org/10.1109/TPAMI.2012.256>
- Sethy, P. K. (2024). Rice leaf disease image samples (Version 2) [Data set]. Mendeley Data. <https://doi.org/10.17632/fwcj7stb8r.2>
- Sethy, P. K., Barpanda, N. K., Rath, A. K., & Behera, S. K. (2020). Deep feature based rice leaf disease identification using support vector machine. *Computers and Electronics in Agriculture*, 175, Article 105527. <https://doi.org/10.1016/j.compag.2020.105527>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 6105–6114). PMLR. <https://proceedings.mlr.press/v97/tan19a.html>
- Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>
- Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11975–11986). https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html
- Zhang, H., Sun, J., Chen, R., Liu, W., Yuan, Z., Zheng, X., Wang, Z., Yang, Z., Yan, H., Zhong, H., Wang, X., Ouyang, W., Yang, F., & Dong, N. (2024). Empowering and assessing the utility of large language models in crop science. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information*

Processing Systems (Vol. 37, pp. 52670–52722). Curran Associates.
<https://doi.org/10.52202/079017-1669>

Zhong, Z. S., Ma, R., & Zhao, H. (2023). Human-uncertainty distillation for calibrated vision models on CIFAR-10H. *Journal of Advanced Computing Systems*, 3(2), 77–89.
<https://doi.org/10.69987/JACS.2023.30206>