Intelligent Automation & Soft Computing

Tech Science Press

https://doi.org/10.32604/iasc.2026.091018

ARTICLE

Hierarchical Transportability of Chinese Medicinal Material Recognition from Web Images to a Batch-Limited Pharmacy Acquisition

Jifei Gao[1], Yuchen Zhang[2,*], Heng Aik Teng[3], Fan Yue[4] and Fan Ying[5]

[1]Faculty of Business and Communications, INTI International University, Persiaran Perdana BBN, Putra Nilai, Nilai, Negeri Sembilan, Malaysia
[2]Faculty of Data Science and Information Technology, INTI International University, Persiaran Perdana BBN, Putra Nilai, Nilai, Negeri Sembilan, Malaysia
[3]School of Traditional Chinese Medicine, INTI International University, Persiaran Perdana BBN, Putra Nilai, Nilai, Negeri Sembilan, Malaysia
[4]Beijing Fanshi Traditional Chinese Medicine Technology Development Co., Ltd., Beijing, China
[5]Xiamen Fanshi Wangutang TCM Clinic Co., Ltd., Xiamen, China

*Corresponding Author: Yuchen Zhang. Email: i24026647@student.newinti.edu.my

Received: 14 August 2026; Accepted: 18 September 2026; Published: 20 October 2026

ABSTRACT: Deep learning-based Chinese medicinal material recognition is usually evaluated within image archives, leaving uncertain which reliability properties remain stable under acquisition shift. We evaluated four levels of transportability using NB-TCM-CHM: closed-set performance, class geometry, confidence operating points, and open-set failure behavior. Models were trained on 3384 web images and evaluated on 400 pharmacy photographs from 20 classes. Because the pharmacy set represents 20 tightly constrained class-specific acquisition batches, all target-domain estimates are conditional on this archived acquisition rather than population-level pharmacy settings. Across three pre-specified splits, closed-set Macro-F1 consistently declined from NB-Web to NB-Pharmacy. Broad class geometry remained comparatively reproducible, whereas the association between centroid displacement and class-specific degradation was not identifiable with precision. Legacy confidence thresholds and detailed open-set failure patterns were substantially less stable under pharmacy acquisition. Additional sensitivity analyses using Transformer backbones, adaptation and calibration methods, and six out-of-distribution (OOD) detectors did not eliminate this reliability gap; held-out recalibration restored known acceptance but not unknown rejection; in a fully reproducible public benchmark, the best area under the receiver operating characteristic curve (AUROC) for OOD detection decreased from 0.971 on NB-Web to 0.855 on NB-Pharmacy. Public NB analyses are reproducible from released data and derived artifacts, whereas the original cross-archive legacy analyses remain conditional on an unreleased internal reference archive.

KEYWORDS: Chinese medicinal materials; domain shift; transportability; out-of-distribution detection; calibration; image recognition; reliability; Human health

1 Introduction

Deep learning has increasingly been applied to visual recognition of Chinese medicinal materials, medicinal plants, and herb-derived products [1–5]. Recent work has improved benchmark performance through attentional pyramids, mobile recognition pipelines, paired-domain datasets, knowledge distillation, cross-attention, and more specialized recognition networks [1–5]. These studies demonstrate the technical

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

feasibility of image-based recognition, but most performance claims are still based on one image archive or relatively homogeneous acquisition conditions.

For practical use, the central question is not only whether a classifier performs well on the archive on which it was developed, but which reliability properties remain valid when image acquisition changes. Benchmark performance can decline on newly collected data even when the nominal task is unchanged [6], and predictive uncertainty can degrade under dataset shift in ways that are not captured by accuracy alone [7–9]. In this paper, transportability denotes the extent to which a reliability property measured in one archive or acquisition setting remains valid in another.

We therefore treat reliability as layered rather than scalar. A system may preserve much of its class geometry while losing closed-set performance; retain reasonable classification accuracy while shifting the score distributions used for abstention; or recover mapped-known coverage after recalibration while continuing to accept unfamiliar classes. These outcomes answer different deployment questions and should not be inferred from one aggregate accuracy value.

The present study uses the public NB-TCM-CHM benchmark [3], which contains the same 20 Chinese medicinal fruit/seed categories in web and pharmacy acquisition settings, together with a frozen internal 29-class legacy reference used only for cross-archive geometry and open-set replication. The primary analysis asks three questions. RQ1: Does closed-set performance degrade consistently from NB-Web to NB-Pharmacy? RQ2: Is class-specific degradation associated with frozen feature centroid displacement between the two domains? RQ3: Are global class geometry, confidence operating points, and open-set failure structure transportable across archives and acquisition domains?

The contributions are fourfold. First, we provide a pre-specified, leakage-aware evaluation of web-to-pharmacy transportability while explicitly treating the pharmacy images as batch-limited evidence rather than a population-level sample. Second, we separate closed-set performance, class geometry, confidence operating points, and open-set failure behavior. Third, additional robustness analyses evaluate whether the findings persist across Transformer backbones, domain-generalization/adaptation methods, test-time adaptation, probability calibration, and six OOD detectors. Fourth, we make the reproducibility boundary explicit: the public NB analyses and new NB-only sensitivity benchmark are reproducible from public data and derived artifacts, whereas the original legacy cross-archive component remains a fixed-reference analysis.

2 Related Work

2.1 Chinese Medicinal Material and Medicinal Plant Image Recognition

Computer vision for Chinese medicinal materials has progressed from proof-of-concept classification toward larger datasets, mobile-oriented models, and lightweight recognition systems [1–5]. Xu et al. [1] used multiple attentional pyramid networks for Chinese herbal recognition, while Sun et al. [2] demonstrated a mobile deep-learning application. Tian et al. [3] introduced NB-TCM-CHM, including 3,384 web images and a 400-image smartphone acquisition of the same 20 classes. Hou et al. [4] subsequently proposed Shuffle-CANet using knowledge distillation and cross-attention, and Tang et al. [5] extended recent medicinal-plant recognition with dual-branch attention fusion.

Published within-archive scores on NB-TCM-CHM are higher than the deliberately leakage-aware source-domain results reported here. The original NB-TCM-CHM study reported a Dataset 1 F1 of 0.9658 for MobileNetV3 and a best Dataset 2 accuracy of 0.8975 for MobileNetV2 [3], whereas Hou et al. reported F1 = 0.988 for ShuffleCANet on Dataset 1 only [4]. The present primary EfficientNet-B0 NB-Web Macro-F1 is approximately 0.94. These figures should not be interpreted as direct head-to-head comparisons because the

studies differ in data partitions, model-selection procedures, augmentation, metrics, leakage controls, and whether Dataset 2 contributes to the evaluation protocol. Our objective is therefore not to establish a new within-archive state of the art, but to characterize which reliability properties survive acquisition change.

2.2 Dataset Shift and Transportability of Visual Recognition

Dataset shift is a well-established challenge in machine learning. Recht et al. [6] showed that ImageNet classifiers can lose accuracy on newly collected but nominally matched test images, Ovadia et al. [7] demonstrated that uncertainty estimates can become unreliable under shift, and WILDS formalized in-the-wild distribution shifts as a benchmark problem [8]. Calibration can also fail under covariate shift [9]. These findings motivate an evaluation that separates recognition performance from confidence behavior and that avoids treating a single target acquisition as representative of an entire deployment environment.

2.3 Open-Set Recognition, OOD Scoring, Calibration, and Trustworthy Evaluation

Open-set and out-of-distribution (OOD) recognition are commonly operationalized through confidence or feature-space scores. Common post-hoc approaches include maximum softmax probability (MSP), class-conditional Mahalanobis distance, energy-based scoring, and ReAct. OpenOOD [10] provides a unified implementation and emphasizes the importance of consistent benchmark protocols for these approaches. Other recent methods include ViM, which combines residual feature information with logits [11], and NNGuide, which incorporates nearest-neighbor geometry [12].

Probability calibration answers a related but distinct question. Temperature scaling, vector scaling, and Dirichlet calibration are standard post-hoc approaches [13]. Calibration quality can degrade under dataset and covariate shift [7,9]. In pharmacy-adjacent visual identification, these distinctions matter because a classifier can remain reasonably accurate while its accept/reject operating point becomes poorly transported. Trustworthy health-AI guidance similarly emphasizes robustness, provenance, transparency, and context-aware evaluation [14].

2.4 Architecture and Adaptation Sensitivity under Shift

Additional robustness analyses examine whether the observed transportability pattern is specific to conventional CNNs or to a source-only deployment regime. Vision Transformer (ViT) [15] and Swin Transformer [16] provide non-convolutional architecture baselines. MixStyle perturbs feature statistics to increase style diversity for domain generalization [17], whereas DANN learns domain-invariant representations using labeled source data and unlabeled target data [18]. For test-time adaptation, Tent minimizes prediction entropy during inference [19], while EATA adds sample selection and anti-forgetting regularization [20]. These methods are treated as post hoc sensitivity analyses rather than replacements for the pre-specified source-only primary design.

3 Materials and Methods

3.1 Study Design and Research Questions

We adopted a hierarchical evaluation framework with four transportability levels: (i) closed-set performance transport, (ii) class-geometry transport, (iii) confidence operating-point transport, and (iv) specific open-set failure transport. Closed-set models were trained using NB-Web only and evaluated on NB-Web and NB-Pharmacy after model selection was frozen. A second component reused outputs from three previously frozen EfficientNet-B0 checkpoints associated with an internally curated 29-class legacy archive assembled

before the present NB transportability study. No legacy model was retrained or re-optimized for the present analysis, and prior-unknown NB outcomes were not used for legacy model or threshold selection.

Of the 20 NB classes, 16 had exact matches to the legacy archive, one was a semantic refinement, two corresponded only to previously studied probe classes, and one was new. Among the 16 exact matches, eight were legacy in-distribution classes and served as mapped-known controls; the other eight had previously been evaluated as unknown classes and formed the prior-unknown replication set. The full mapping is provided in Supplementary Table S1. The public repository reproduces the NB closed-set workflow and the new public NB-only sensitivity benchmark; it does not reproduce analyses whose estimand depends on unavailable legacy raw artifacts.

Fig. 1 summarizes the primary study design. The additional sensitivity analyses are described separately in Section 3.7 so that the original research questions and pre-specified primary endpoints remain distinguishable from post hoc robustness work.

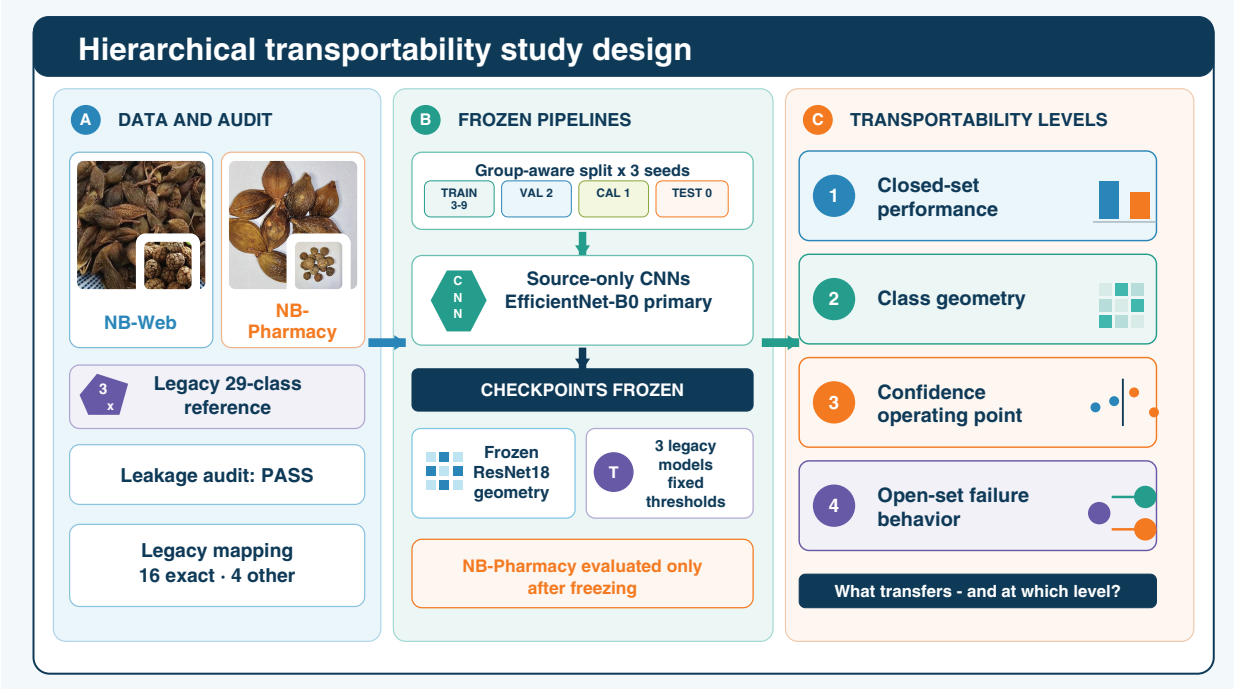


Figure 1: Conceptual overview of the hierarchical transportability study design. (**A**) NB-Web and NB-Pharmacy data, leakage audit, and cross-archive class mapping. (**B**) Group-aware closed-set training, checkpoint freezing, frozen ResNet18 geometry, and legacy-model threshold analyses; NB-Pharmacy was evaluated only after the primary analysis components were frozen. (**C**) Four evaluated levels of transportability: closed-set performance, class geometry, confidence operating point, and open-set failure behavior. Representative medicinal-material images are reproduced from NB-TCM-CHM [3] under CC BY 4.0, with layout adaptation by the authors.

3.2 Datasets, Acquisition Context, Label Mapping, and Audit Procedures

NB-TCM-CHM Version 3 contains 3384 web images (Dataset 1; NB-Web) and 400 pharmacy smartphone images (Dataset 2; NB-Pharmacy) covering 20 Chinese medicinal fruit/seed categories [3]. Dataset 1 contains 87–280 images per class; Dataset 2 contains exactly 20 images per class. According to the original dataset report, Dataset 2 was photographed against white A4 paper with a smartphone mounted on a support and LED fill lighting maintained at a constant intensity [3]. The public documentation does not report exact illuminance, color temperature, quantitative shadow measurements, or complete device metadata.

Inspection of the archived acquisition context indicates short within-class capture windows of approximately 3–5 min for most classes. We therefore treat NB-Pharmacy conservatively as 20 class-specific acquisition batches rather than 400 independent specimens. Class identity is consequently intertwined with batch-specific acquisition conditions, so all pharmacy-domain estimates in this paper are conditional on this archived acquisition. They should not be generalized to multiple pharmacies, devices, dates, or lighting environments.

Before modeling, we conducted archive-level audit procedures. File integrity and actual image formats were checked; exact and near duplicates were searched within NB-Web, within NB-Pharmacy, and across NB and the legacy archive; and confirmed duplicate groups were constrained to a single split partition. Exact duplicates were identified by hash matching. The original primary split registry used pHash/dHash screening, frozen ImageNet-feature similarity, and manual confirmation. In the public reproducibility package, an independently reconstructed duplicate registry is provided as a sensitivity audit; it is not used to retroactively redefine the original primary analysis. Cross-archive class mappings were categorized as exact, semantic refinement, previous-probe-only, or new-only.

A blinded two-rater processing-state reassessment was also performed for 107 images from Persicae Semen and Armeniacae Semen Amarum using the predefined categories peeled, intact coat, and mixed/fragmented. Because agreement was very low, these judgments were retained only as an annotation-quality audit and were not used to support mechanistic claims or model selection. Table 1 summarizes the dataset characteristics, acquisition context, and audit findings.

Table 1: Summary of the NB-TCM-CHM datasets and audit findings.

Item	NB-Web (Dataset 1)	NB-Pharmacy (Dataset 2)	Legacy Archive Relation
Number of classes	20	20	29 (frozen outputs only)
Number of images	3384	400	–
Images per class	87–280	20	–
Confirmed duplicate pairs/clusters	35 duplicate pairs/clusters; same-label	2 duplicate pairs	0 confirmed cross-archive near duplicates
Acquisition style	Web/Google images	Smartphone pharmacy photographs	Mixed historical archive
Exact class matches to legacy archive	16	16	Reference side
Semantic refinement matches	1	1	Reference side
Previous-probe-only matches	2	2	Reference side
New-only classes	1	1	Reference side

3.3 Closed-Set Training Protocol

Three convolutional backbones were evaluated in the pre-specified primary analysis: ResNet18 [21], EfficientNet-B0 [22], and MobileNetV3-Small [23]. EfficientNet-B0 was designated a priori as the primary

backbone before NB-Pharmacy results were examined, maintaining continuity with the frozen legacy pipeline. ResNet18 and MobileNetV3-Small served as robustness backbones and did not determine the primary-model designation.

Three pre-specified split/model seeds (20260814, 20260815, and 20260816) were used. For each seed, a 10-fold stratified group split was constructed from the leakage groups: folds 3–9 formed training, fold 2 validation, fold 1 a reserved calibration holdout, and fold 0 the internal web test partition. Checkpoint selection and early stopping used validation Macro-F1. The seed controlled Python, NumPy, PyTorch CPU/GPU randomization, data-loader order, model initialization, and the corresponding split file. Consequently, split and initialization effects vary together and are not separately identifiable.

All three CNNs used ImageNet-pretrained weights. Training used 224×224 inputs with RandomResizedCrop (scale 0.8–1.0), random horizontal flipping, mild ColorJitter, and ImageNet normalization; evaluation used Resize(256), CenterCrop(224), and the same normalization. Optimization used class-weighted cross-entropy with label smoothing 0.05, AdamW (learning rate 3×10^{-4}; weight decay 1×10^{-4}), cosine annealing, a maximum of 30 epochs, and early stopping after five validation epochs without improvement.

The executable training script and public reproduction package use batch size 64. An earlier frozen protocol copy recorded batch size 32. Rather than silently reconciling these records, we disclose the provenance discrepancy and report a post hoc paired sensitivity analysis of batch sizes 32 and 64 across the three primary seeds (Supplementary Methods S3). The split seeds, architecture set, primary-backbone designation, and sealed pharmacy evaluation were not redefined by this sensitivity analysis.

Closed-set performance was evaluated by Macro-F1 on both domains. For the primary EfficientNet-B0 analysis, the web-to-pharmacy gap was further summarized with a class- and split-stratified bootstrap interval. Class recall was computed separately in each domain to identify the largest domain-specific degradation patterns.

3.4 Frozen Feature Geometry Analyses

To test RQ2, we extracted features with an independently frozen ImageNet-pretrained ResNet18 representation and computed a centroid for each class in each domain. The class-specific centroid displacement was defined as:

$$D_c = 1 - \cos\left(\mu_c^{\text{web}}, \mu_c^{\text{pharmacy}}\right) \tag{1}$$

where the two μ terms denote normalized class centroids. Class-specific degradation was defined as:

$$\Delta R_c = R_c^{\text{web}} - R_c^{\text{pharmacy}}. \tag{2}$$

RQ2 was tested using the Spearman correlation between D_c and ΔR_c across the 20 classes, with a two-sided label-permutation test and bootstrap interval. Sensitivity analyses also examined partial rank correlations controlling for web recall or class sample size. Because NB-Pharmacy is batch-limited, a non-significant result is interpreted as limited identifiability rather than evidence that no association exists.

For cross-archive global geometry, the same frozen representation was used to build 16×16 pairwise class-similarity matrices for exact-match classes shared between the legacy archive and NB. Spearman correlations between archived pairwise-similarity summaries quantified global geometry transport, and nearest-neighbor retention quantified local neighbor stability. These legacy matrix correlations remain

descriptive fixed-reference summaries because the underlying legacy pairwise matrices are not part of the public archive.

3.5 Operating-Point Transportability and Legacy Open-Set Analyses

Three standard OOD scorers were used in the original legacy-reference analysis: MSP, energy, and full-covariance Mahalanobis distance, using the standardized detector definitions summarized by OpenOOD [10]; Ledoit-Wolf shrinkage was used for covariance estimation [24]. The Mahalanobis score was defined as:

$$S_M(x) = -\min_{c\in\mathcal{C}_{\mathrm{ID}}} \delta_c(x)^\top \widehat{\Sigma}_{\mathrm{LW}}^{-1} \delta_c(x), \qquad \delta_c(x) = f(x) - \mu_c. \tag{3}$$

For each scorer, acceptance followed a score-specific rule:

$$\mathrm{Accept}_s(x) = \mathbf{1}[S_s(x) \geq \tau_s]. \tag{4}$$

In the frozen-threshold setting, checkpoint-specific thresholds inherited from the legacy calibration partition were applied unchanged to NB. This tests direct operating-point transport. In a separate favorable sensitivity analysis, thresholds were recalibrated independently for each legacy checkpoint, NB domain, and scorer using all images from the eight mapped-known classes in that domain, targeting approximately 95% mapped-known acceptance. This full-known recalibration was descriptive and should not be interpreted as a deployment-realistic calibration/test split.

Mapped-known acceptance was summarized as IDAR = N_accepted, mapped-known/N_mapped-known. Prior-unknown acceptance was macro-averaged across the eight prior-unknown classes. Failure-order transport was measured by exact Spearman tests over all 8! label permutations. Threshold-free absorption analysis compared the dominant known-class target of each prior-unknown class with the frozen feature-centroid nearest legacy class.

3.6 Statistical Analysis and Reporting Units

Primary closed-set means and sample standard deviations are reported across the three pre-specified split/model runs (n = 3), and all occurrences of mean ± SD are labeled accordingly. These run-level SDs describe variation across fitted models and splits; they do not represent population-level sampling uncertainty for pharmacy acquisition. The class- and split-stratified bootstrap interval for the primary domain gap is reported separately. RQ2 uses permutation-based inference, whereas the original legacy global geometry correlations are descriptive fixed-reference summaries.

Because NB-Pharmacy represents class-specific acquisition batches rather than independent specimens, class/batch-aware uncertainty is preferred whenever class-level endpoints are resampled. Leave-one-class-out (LOCO) analyses are reported as sensitivity ranges rather than confidence intervals. The seven additional split/model seeds are explicitly post hoc and are not used to redefine the original primary analysis.

3.7 Additional Robustness and Fully Reproducible Public Sensitivity Analyses

To assess stronger baselines while preserving the original estimand, we added post hoc analyses whose information requirements are reported explicitly. Architecture robustness used ImageNet-pretrained ViT-B/16 [15] and Swin-T [16] on the same NB-Web-only train/validation design. Learning-rate pilots were selected from NB-Web validation only and then frozen before pharmacy evaluation. Seven additional split/model seeds (20260817–20260823) were evaluated for EfficientNet-B0 as a post hoc stability analysis.

Domain-generalization sensitivity used MixStyle [17] with physical and effective batch size matched to the source-only EfficientNet-B0 control. Unsupervised domain adaptation used DANN [18] with 10 pharmacy images per class as an unlabeled adaptation partition and the remaining 10 per class as held-out evaluation, repeated across five balanced target partitions for each of three source checkpoints. Pharmacy labels were used only offline to construct balanced partitions and compute final metrics; they were never provided to the adaptation loss.

Test-time adaptation used Tent [19] and EATA [20] over five randomized pharmacy stream orders per source checkpoint. Each run started from the same frozen source checkpoint. A zero-update configuration control measured the effect of switching the model/normalization configuration used by TTA without performing optimizer steps, allowing configuration effects to be separated from parameter-update effects.

Probability calibration was evaluated independently of OOD threshold recalibration. Temperature scaling, vector scaling, and Dirichlet calibration [13] were fitted on the reserved NB-Web calibration partition only and transported unchanged to NB-Web test and NB-Pharmacy. Metrics were negative log-likelihood (NLL), Brier score, expected calibration error with 15 equal-width bins (ECE15), and Macro-F1.

A new, fully reproducible public NB-only OOD sensitivity benchmark was added because the original legacy images, checkpoints, features, and score archive are not distributable. The eight publicly mapped-known classes defined the ID task, and the eight public prior-unknown classes were held out from training and detector fitting. Six detectors were evaluated within a consistent framework informed by OpenOOD [10]: MSP, energy, Mahalanobis-LW [24], ViM [11], ReAct-Energy, and NNGuide [12]. NB-Web mapped-known calibration data selected fixed operating points targeting 95% acceptance; domain-matched NB-Web and NB-Pharmacy subsets provided ID/OOD tests. Standard OOD metrics included AUROC and FPR95, with detailed AUPR and per-class acceptance outputs provided in Supplementary Data S1.

For deployment-oriented threshold sensitivity, each mapped-known pharmacy class was partitioned into 5/15, 10/10, or 15/5 calibration/test images, repeated over 30 fixed stratified partitions per checkpoint and detector. Thresholds targeted 95% mapped-known acceptance using calibration images only; prior-unknown pharmacy images were never used to select thresholds. Public class geometry was additionally reanalyzed with 10,000 simultaneous class-label row/column permutations and LOCO analysis, while prior-unknown acceptance used 10,000 class/batch-bootstrap replicates and LOCO ranges rather than image-level resampling. Full protocols and run-level outputs are provided in Supplementary Methods S1–S9, Tables S2–S13, Figures S1–S7, and Supplementary Data S1.

4 Results

4.1 Audit Results and Data Quality Findings

The initial archive audit supported continued analysis but with downgraded external inference. NB-Web contained 3384 images across 20 classes, whereas NB-Pharmacy contained 400 images with exactly 20 images per class. Internal duplicate inspection identified 35 confirmed duplicate pairs/clusters in NB-Web and 2 duplicate pairs in NB-Pharmacy, all within the same class. Split-registry verification found no confirmed exact or near-duplicate group crossing the training, validation, reserved-calibration, or test partitions in any pre-specified split. No confirmed near duplicates were found between the legacy archive and either NB domain.

The mapping to the legacy archive yielded 16 exact matches, 1 semantic refinement, 2 previous-probe-only matches, and 1 new-only class. The blinded processing-state reassessment produced 57 agreements among 107 reviewed images (raw agreement 0.5327; Cohen's κ = 0.0699). Because agreement was poor,

image-level processing-state proportions were treated as unresolved and were not used to support causal or mechanistic interpretation.

4.2 Closed-Set Web-to-Pharmacy Degradation and Architecture Robustness

Representative cross-domain examples are shown in Fig. 2. All three pre-specified CNN backbones showed consistent performance loss from NB-Web to NB-Pharmacy (Table 2; Fig. 3A). ResNet18 had the highest mean NB-Pharmacy Macro-F1 (0.8573), while EfficientNet-B0 remained the primary backbone because that designation was fixed before pharmacy results were examined. EfficientNet-B0 achieved 0.9400 ± 0.0193 on NB-Web and 0.8304 ± 0.0133 on NB-Pharmacy (mean ± SD, n = 3), a mean decline of 0.1096. A class- and split-stratified bootstrap yielded a 95% confidence interval of 0.0564–0.1707, and all three split-specific domain gaps had the same direction.

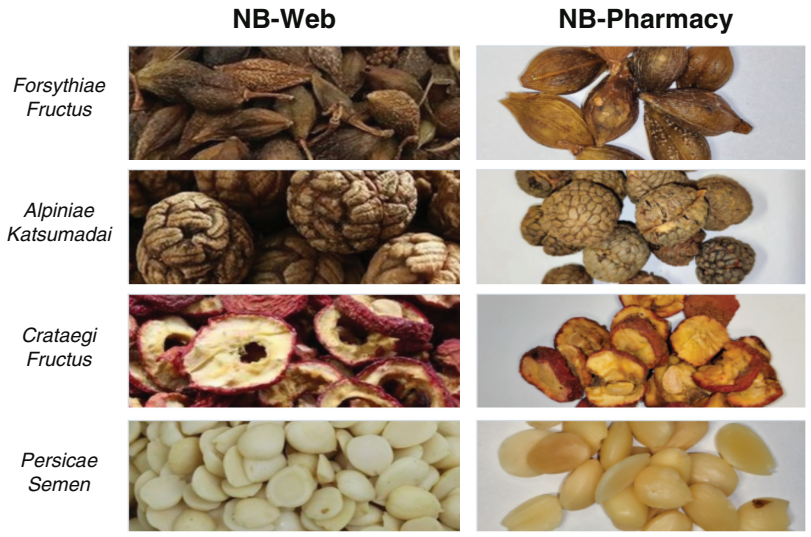


Figure 2: Representative NB-Web and NB-Pharmacy images for four classes with pronounced domain-specific recall degradation: Forsythiae Fructus, Alpiniae Katsumadai Semen, Crataegi Fructus, and Persicae Semen. Images are reproduced from NB-TCM-CHM [3] under CC BY 4.0, with layout adaptation by the authors. The processing-state reassessment is no longer used as a principal visual result because its inter-rater reliability was very low.

Table 2: Closed-set recognition performance on the web and pharmacy domains. Values are mean ± sample SD across the three pre-specified split/model runs (n = 3).

Backbone	NB-Web Macro-F1	NB-Pharmacy Macro-F1	Web-Pharmacy Gap
EfficientNet-B0	0.9400 ± 0.0193	0.8304 ± 0.0133	0.1096 ± 0.0289
MobileNetV3-Small	0.9360 ± 0.0045	0.8158 ± 0.0283	0.1202 ± 0.0307
ResNet18	0.9390 ± 0.0103	0.8573 ± 0.0100	0.0818 ± 0.0184

The largest EfficientNet-B0 recall losses occurred in Forsythiae Fructus (0.952 to 0.533), Alpiniae Katsumadai Semen (0.967 to 0.650), Crataegi Fructus (0.972 to 0.667), Persicae Semen (0.333 to 0.100), Psoraleae Fructus (1.000 to 0.783), and Gardeniae Fructus (0.982 to 0.800) (Fig. 3B). Persicae Semen requires a different interpretation because its NB-Web recall was already only 0.333; its pharmacy decline therefore aggravates a pre-existing class-boundary problem rather than representing a purely acquisition-induced failure.

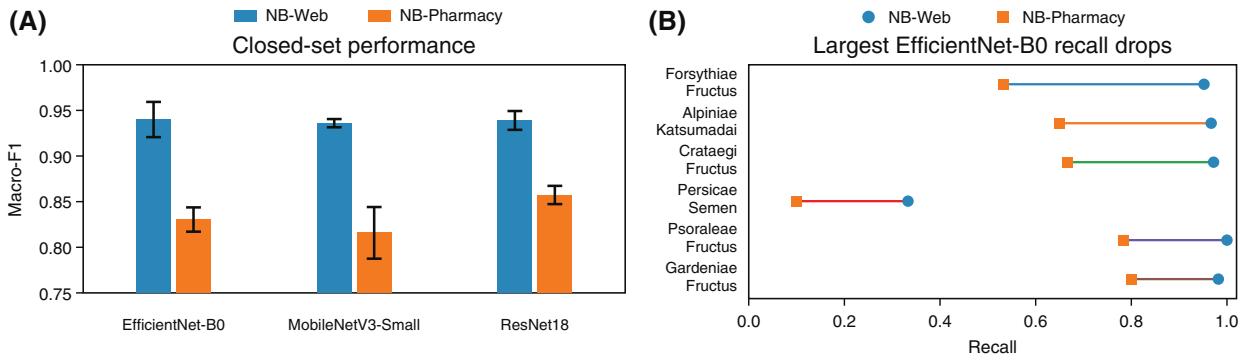


Figure 3: Closed-set web-vs-pharmacy performance. (**A**) Macro-F1 for the three pre-specified CNN backbones; error bars show sample SD across n = 3 split/model runs. (**B**) The six largest class-level recall drops for EfficientNet-B0. Legends are placed outside the plotting regions to avoid obscuring data.

Additional robustness analyses yielded the same directional pattern. In an independently rerun public sensitivity pipeline, the EfficientNet-B0 web-to-pharmacy gap was 0.098 ± 0.026 at batch size 32 and 0.093 ± 0.009 at batch size 64 (n = 3 each); the paired difference in the gap was 0.005 (95% CI −0.072 to 0.082), so the qualitative conclusion was insensitive to the recorded batch-size discrepancy without establishing formal equivalence. Across all ten batch-64 split/model seeds, including seven post hoc seeds, the mean gap was 0.097 (95% CI 0.084–0.110) and all 10/10 gaps were positive. ViT-B/16 yielded a mean gap of 0.096 and Swin-T 0.070 (n = 3 each), indicating that stronger Transformer backbones did not remove the acquisition-specific loss, although Swin-T uncertainty was wide (Supplementary Table S5 and Fig. S1).

4.3 Processing-State Reassessment

The blinded processing-state reassessment was retained only as an annotation-quality result. The visual distinction between peeled kernels and light-colored retained seed coats could not be reconstructed reproducibly from the archived images, and the Persicae Semen/Armeniacae Semen Amarum confusion could therefore not be attributed confirmatorily to image-visible processing state. For future medicinal-material datasets, variables that may shape difficult class boundaries should be recorded at acquisition time rather than inferred retrospectively.

4.4 Class Geometry and the RQ2 Association Analysis

RQ2 asked whether class-specific recall degradation could be predicted from frozen feature centroid displacement. No association was identifiable with precision under the present design: Spearman $\rho = 0.2123$, two-sided permutation $p = 0.3696$, with bootstrap 95% CI −0.2007 to 0.5633. Partial rank correlations controlling for web-domain recall or class sample size were similarly small. Because each pharmacy class is represented by a class-specific acquisition batch and the confidence interval is broad, this result does not distinguish a true absence of association from limited power or identifiability.

Global visual geometry nevertheless showed positive descriptive concordance across archives. For the 16 exact-match classes, legacy-vs.-NB-Web pairwise similarity summaries had $\rho = 0.6294$, legacy-vs.-NB-Pharmacy $\rho = 0.5895$, and NB-Web-vs.-NB-Pharmacy $\rho = 0.7908$. Nearest-neighbor preservation was 0.625 for legacy-to-web and 0.438 for legacy-to-pharmacy comparisons. Table 3 and Fig. 4 summarize these results. The differing RQ2 and global-geometry outcomes reflect different geometric summaries and inferential questions rather than a change of feature extractor.

Table 3: Summary of class-geometry transportability and the RQ2 association analysis.

Analysis	Result	Interpretation
Centroid displacement vs. recall degradation	ρ = 0.2123; p = 0.3696; 95% CI −0.2007 to 0.5633	No association identifiable with precision under the batch-limited design
Legacy vs. NB-Web geometry	ρ = 0.6294	Positive descriptive cross-archive concordance
Legacy vs. NB-Pharmacy geometry	ρ = 0.5895	Positive descriptive cross-archive concordance
NB-Web vs. NB-Pharmacy geometry	ρ = 0.7908	Strongest descriptive pairwise concordance
Nearest-neighbor retention, legacy to web	0.625	Moderate local neighbor stability
Nearest-neighbor retention, legacy to pharmacy	0.438	Weaker local neighbor stability

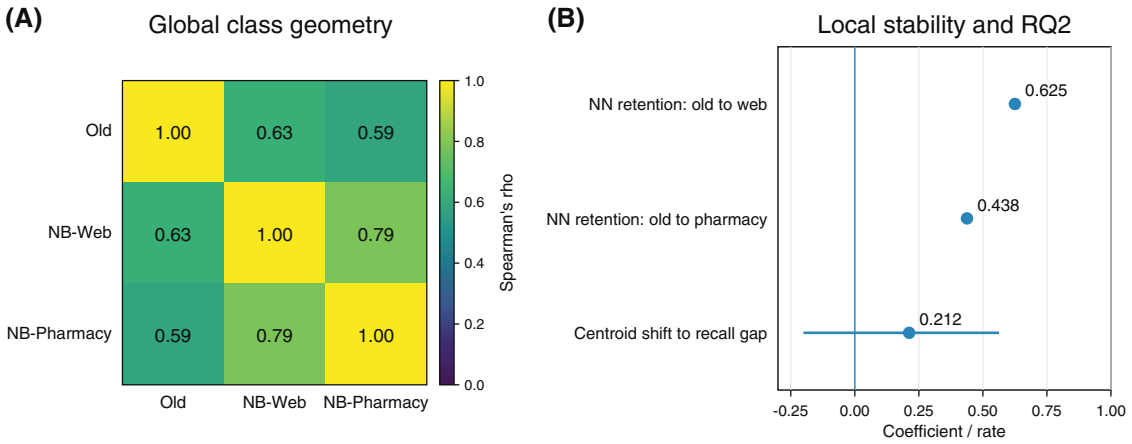


Figure 4: Geometry transportability results. (**A**) Pairwise class-geometry correlations across the legacy archive, NB-Web, and NB-Pharmacy. (**B**) Summary of nearest-neighbor retention and the RQ2 association between centroid displacement and recall degradation; the horizontal interval denotes the bootstrap 95% confidence interval.

The new fully reproducible public 8-class sensitivity analysis provided a complementary class-aware check on NB-Web/NB-Pharmacy geometry. Across the three checkpoints, class-similarity correlations were 0.767–0.833; simultaneous row/column class-label permutation gave p = 0.0003–0.0004, and the combined LOCO range was 0.605–0.927 (Supplementary Table S11). This public-task result is not a reproduction of the unavailable legacy cross-archive geometry, but it strengthens the within-NB conclusion that broad pairwise structure can remain coherent despite acquisition-specific performance loss.

4.5 Operating-Point Transportability and Held-Out Recalibration

Fixed confidence thresholds transported differently to the two NB domains. With legacy thresholds applied unchanged, mapped-known acceptance in NB-Web was 0.9785 ± 0.0044 for MSP, 0.9881 ± 0.0024

for energy, and 0.9925 ± 0.0014 for Mahalanobis-LW. In NB-Pharmacy, the corresponding values were 0.7667 ± 0.0641, 0.8708 ± 0.0237, and 0.8479 ± 0.0452. Here and below, ± denotes sample SD across n = 3 frozen legacy checkpoints, not image-level sampling uncertainty. Table 4 and Fig. 5 summarize the frozen-threshold and full-known recalibration results.

Table 4: Legacy-reference operating-point transportability. Values are mean ± sample SD across $n = 3$ frozen legacy checkpoints. Frozen mapped-known acceptance uses thresholds inherited from the legacy archive. Recalibrated unknown acceptance uses thresholds re-estimated on mapped-known NB controls only.

Scorer	Frozen Known NB-Web	Frozen Known NB-Pharmacy	Recal. Unknown Legacy	Recal. Unknown NB-Web	Recal. Unknown NB-Pharmacy
MSP	0.9785 ± 0.0044	0.7667 ± 0.0641	0.5226 ± 0.0090	0.3636 ± 0.0987	0.7354 ± 0.1028
Energy	0.9881 ± 0.0024	0.8708 ± 0.0237	0.5690 ± 0.0254	0.2086 ± 0.0523	0.5604 ± 0.0532
Mahalanobis-LW	0.9925 ± 0.0014	0.8479 ± 0.0452	0.6208 ± 0.0210	0.1161 ± 0.0122	0.6229 ± 0.0957

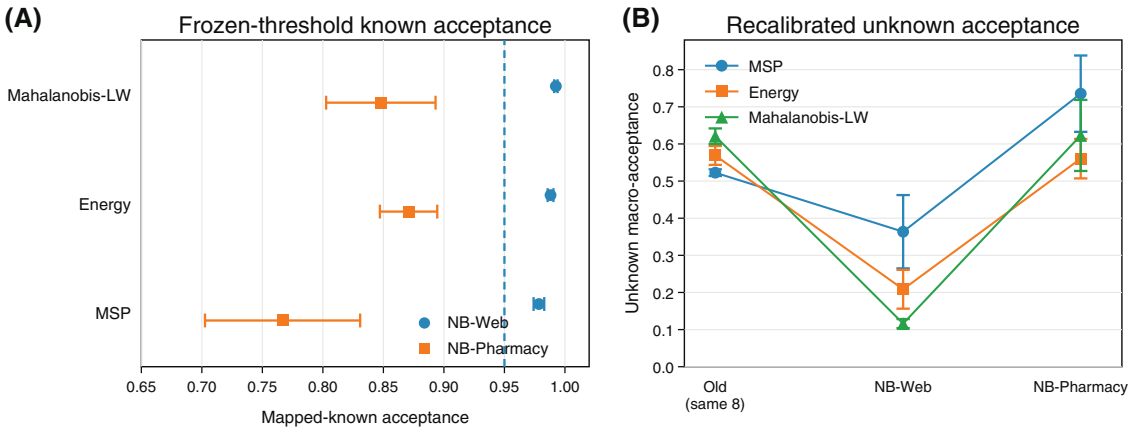


Figure 5: Legacy-reference operating-point transportability. (**A**) Mapped-known acceptance under frozen legacy thresholds; the dashed line denotes the nominal 0.95 acceptance target. (**B**) Unknown macro-acceptance for the same eight prior-unknown classes after full-known domain-specific recalibration. Error bars show sample SD across n = 3 frozen legacy checkpoints.

In the favorable full-known recalibration sensitivity analysis, thresholds were re-estimated within each NB domain using all eight mapped-known classes. Realized mapped-known acceptance was approximately 0.95 by construction. For the same eight prior-unknown classes, macro-acceptance in NB-Pharmacy was 0.7354 ± 0.1028 for MSP, 0.5604 ± 0.0532 for energy, and 0.6229 ± 0.0957 for Mahalanobis-LW. Thus, restoring mapped-known coverage did not restore the prior-unknown failure pattern.

The public held-out recalibration analysis removed the calibration/test overlap. For each checkpoint and detector, 5, 10, or 15 mapped-known pharmacy images per class were used for threshold calibration, with the remaining known images held out for testing over 30 stratified partitions. At 10 calibration images per class, held-out mapped-known acceptance remained 0.939–0.955 across the six detectors, but prior-unknown macro-acceptance remained 0.591–0.768. Results were similar across the 5/10/15 calibration-size settings (Supplementary Table S10 and Fig. S7). More samples from the same constrained acquisition batch therefore did not by themselves solve unknown rejection.

4.6 Failure-Order Transport and Threshold-Free Open-Set Behavior

After full-known recalibration, the ordering of the eight legacy prior-unknown classes was reproduced in NB-Web but not in NB-Pharmacy. Legacy-to-NB-Web Spearman correlations were 0.8333 for MSP (exact two-sided p = 0.0154), 0.9524 for energy (p = 0.0011), and 0.8571 for Mahalanobis-LW (p = 0.0107). In NB-Pharmacy, the corresponding correlations were −0.4791 (p = 0.2327), −0.1667 (p = 0.7033), and −0.0964 (p = 0.8300). The symmetric LOCO analysis supported the same contrast: NB-Web correlations remained strongly positive, whereas NB-Pharmacy correlations remained weak or negative overall. Fig. 6A summarizes the full-set and LOCO failure-order results.

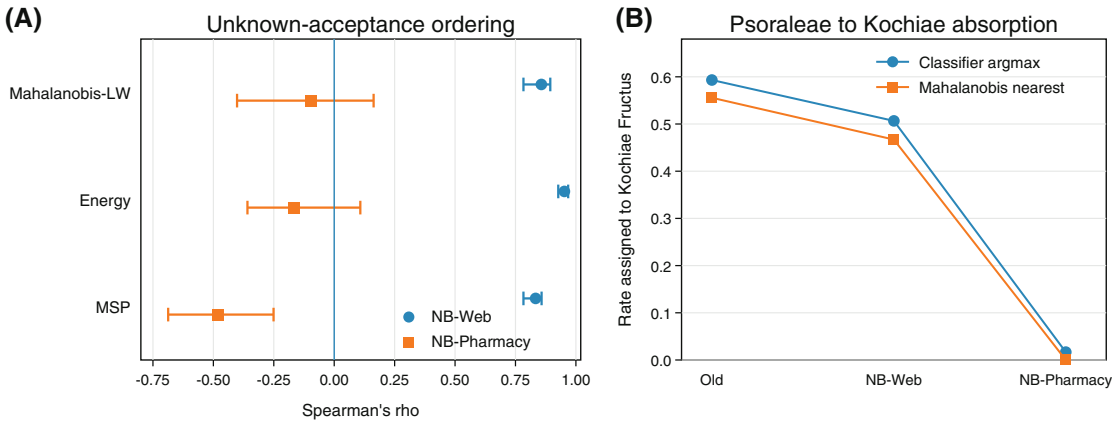


Figure 6: Open-set behavior transport. (**A**) Full-set Spearman correlations of legacy-to-NB unknown-acceptance ordering after mapped-known recalibration; horizontal bars show descriptive leave-one-class-out ranges rather than confidence intervals. (**B**) Threshold-free absorption of Psoraleae Fructus into Kochiae Fructus across the legacy archive, NB-Web, and NB-Pharmacy.

Threshold-free absorption reinforced this interpretation. Psoraleae Fructus was predominantly absorbed into Kochiae Fructus in the frozen legacy reference and again in NB-Web, but the relation nearly disappeared in NB-Pharmacy. Under classifier argmax, the proportion of Psoraleae Fructus images classified as Kochiae Fructus fell from 0.5932 in the legacy archive and 0.5066 in NB-Web to 0.0167 in NB-Pharmacy; under Mahalanobis nearest-class assignment, the corresponding proportions were 0.5556, 0.4670, and 0.0000. Fig. 6B shows this example. The pharmacy shift therefore affected not only thresholds but also the detailed local topology of specific open-set failures.

4.7 Additional Method Robustness and Public OOD Sensitivity

The expanded method comparison did not identify a general remedy for the acquisition-specific gap. Under the matched physical/effective batch-size control, MixStyle reduced NB-Pharmacy Macro-F1 by 0.029 relative to the paired source-only baseline (95% CI −0.046 to −0.012; n = 3). DANN showed no identifiable checkpoint-level pharmacy improvement (mean change −0.008; 95% CI −0.128 to 0.112; n = 3) and reduced retained NB-Web performance by approximately 0.030. Tent and EATA likewise failed to improve pharmacy performance reliably. Importantly, zero-update controls showed that the large retained-Web decrease observed under the TTA implementation was dominated by the BatchNorm/model-configuration change (about −0.411) rather than by optimizer updates, which contributed only about +0.001. We therefore do not interpret that decrease as adaptation-induced catastrophic forgetting (Supplementary Tables S6 and S7 and Figs. S2–S4).

Web-fitted probability calibration improved selected confidence metrics without removing classification shift. Temperature scaling reduced mean NB-Pharmacy NLL from 0.448 to 0.414 and ECE15 from 0.064 to 0.042 while leaving Macro-F1 unchanged at 0.850. Vector scaling produced the highest mean pharmacy Macro-F1 (0.862) but had poorer and more variable NLL, and Dirichlet calibration was not uniformly beneficial (Supplementary Table S8 and Fig. S5).

The fully reproducible public NB-only OOD benchmark showed deterioration for all six detectors. ViM achieved the strongest mean discrimination, with AUROC decreasing from 0.971 on NB-Web to 0.855 on NB-Pharmacy and Pharmacy FPR95 = 0.571. Mahalanobis-LW decreased from AUROC 0.966 to 0.845 with Pharmacy FPR95 = 0.588. MSP, energy, ReAct-Energy, and NNGuide showed the same directional degradation (Supplementary Table S9 and Fig. S6). Class-aware uncertainty remained wide because only eight prior-unknown class/batch units were available; for example, Pharmacy prior-unknown macro-acceptance under Web-calibrated thresholds averaged 0.102 for ViM and 0.133 for Mahalanobis-LW, with wider class-bootstrap and LOCO ranges reported in Supplementary Table S12. These public results complement, but do not reproduce or replace, the unavailable legacy-reference analyses.

5 Discussion and Limitations

5.1 Hierarchical Transportability under Acquisition Shift

The main contribution of this study is conceptual rather than architectural. Across the pre-specified primary analyses, reliability properties did not move in parallel after the shift from web images to the archived pharmacy acquisition. Closed-set Macro-F1 declined across all three CNN backbones; broad class geometry remained comparatively coherent; fixed confidence operating points transported poorly to pharmacy images; and detailed open-set failure ordering was less stable.

The geometry results require a similarly qualified interpretation. The original cross-archive pairwise similarity matrices showed positive concordance, whereas RQ2 produced a weak and imprecise class-level association between centroid displacement and recall degradation. These are not contradictory findings. One concerns preservation of broad inter-class relations; the other asks whether a single class-specific displacement statistic predicts performance loss across only 20 batch-linked classes. The new public 8-class permutation/LOCO analysis supports broad within-NB geometry stability, but it does not increase the power of the original RQ2 test or reproduce the unavailable legacy matrices.

5.2 Robustness, Adaptation, Calibration, and OOD Detection

The additional sensitivity analyses strengthened this conclusion by showing that the same acquisition-specific problem persisted across ViT-B/16 and Swin-T and was not consistently removed by MixStyle, DANN, Tent, EATA, or post-hoc probability calibration. The expanded OOD analyses make the distinction between recognition and rejection especially clear. On the fully reproducible public 8-known/8-prior-unknown benchmark, ViM and Mahalanobis-LW outperformed simpler detectors, yet both still showed a marked AUROC decline and high FPR95 on NB-Pharmacy. Held-out target-domain recalibration restored mapped-known coverage close to the 95% target, but unknown acceptance remained substantial. Thus, access to more known images from the same acquisition batch can repair the operating point for known classes without restoring reliable unknown rejection. This is the practical meaning of hierarchical transportability: a repaired threshold is not equivalent to a repaired failure structure.

The present source-domain scores should also be interpreted in relation to prior NB-TCM-CHM reports. The original dataset paper and a later ShuffleCANet study reported higher within-Dataset-1 F1 values than the approximately 0.94 primary NB-Web Macro-F1 reported here [3,4]. Those studies used

different split, optimization, and evaluation procedures, and the later study evaluated Dataset 1 only [4]. The lower source-domain score in the present study is therefore not evidence that the benchmark has become intrinsically harder; it reflects a protocol designed around group-aware leakage control, pre-specified splits, a reserved calibration fold, and subsequent cross-domain evaluation rather than maximization of within-archive accuracy.

5.3 Limitations, Reproducibility, and Future Directions

Several limitations constrain external inference. First, NB-Pharmacy is a small, tightly constrained public acquisition represented here as 20 class-specific batches. Class identity and batch-specific acquisition conditions are intertwined, so none of the pharmacy results estimate performance across pharmacies, devices, sessions, lighting systems, or specimen sources. Second, n = 3 primary split/model runs is small; the additional seven seeds are post hoc sensitivity runs and do not convert the archive into an independent population sample. Third, the public OOD benchmark contains only eight prior-unknown class/batch units, which produces wide class-aware uncertainty. Fourth, the processing-state audit had near-zero kappa and therefore cannot support a mechanistic explanation of the Persicae Semen confusion boundary.

A further limitation is the legacy reproducibility boundary. The original cross-archive geometry, operating-point, and failure-order analyses depend on a frozen internal 29-class archive whose raw images, checkpoints, complete logits/features, and score archive are not publicly distributable. We therefore treat those analyses as fixed-reference evidence and no longer present them as independently reproducible. To improve auditability, we provide a fully public NB-only OOD benchmark, held-out calibration sensitivity, class-aware statistics, detailed derived outputs, and scripts in the supplementary/reproducibility package. These additions address the same methodological concerns without claiming to reconstruct the legacy archive.

From a broader sustainable-health perspective, this emphasis on reliable, context-aware decision support is consistent with Sustainable Development Goal 3 (Good Health and Well-Being). Recent multimodal medical-imaging work has likewise shown that discrimination, calibration, and operating-point performance should be evaluated separately when considering deployment-oriented clinical AI [25].

Recent literature further reinforces the need for explicit shift-aware evaluation. Fine-grained benchmarking has shown that method rankings can vary across different distribution-shift mechanisms [26]; stable test-time adaptation remains challenging under mixed shifts and small batches [27]; and clinical model transport can fail when data-generating processes differ across sites [28]. Post-market monitoring has therefore been proposed to detect distribution shift after deployment [29], while recent mammography evidence shows that mismatch between calibration and target populations can materially alter clinically relevant operating points [30]. Recent INTI-affiliated studies also illustrate the expanding use of deep learning across HER2 pathology, oral clinical images, and breast-cancer histopathology [31–33].

Future work should prioritize data design over further model proliferation. Repeated pharmacy acquisition should include multiple devices, dates, lighting conditions, sites, and specimen batches, with the same classes represented across multiple sessions so that class identity is not confounded with acquisition batch. Prospective calibration studies should use explicit calibration/test separation and should evaluate classification, calibration, OOD rejection, and failure neighborhoods jointly. If adaptation is pursued, it should be judged not only by target-domain accuracy but also by whether it preserves source performance and improves reliable abstention under genuinely independent acquisition conditions.

6 Conclusion

This study evaluated the transportability of deep learning-based Chinese medicinal material recognition from web images to a batch-limited pharmacy acquisition using a hierarchical reliability framework. Closed-set performance declined consistently across the pre-specified CNN backbones and remained directionally degraded in Transformer sensitivity analyses. Broad class geometry was comparatively reproducible, whereas the class-level relation between centroid displacement and recall degradation could not be identified with precision under the present design. Confidence operating points and detailed open-set failure behavior were less stable, and stronger OOD detectors or same-batch recalibration did not eliminate the pharmacy-domain rejection problem. These findings support reporting classification performance, geometry, calibration, operating points, and open-set behavior as distinct reliability properties rather than inferring deployment robustness from closed-set accuracy alone. Because the target data come from a tightly constrained, class-confounded acquisition and part of the original cross-archive analysis relies on a non-public legacy reference, the conclusions should be interpreted as evidence from the archived evaluation setting rather than as population-level estimates of pharmacy deployment performance.

Acknowledgement: During manuscript preparation and revision, the authors used ChatGPT (OpenAI) to assist with language editing, structural revision, and LaTeX/Word formatting. All scientific analyses, numerical results, interpretations, citations, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jifei Gao and Yuchen Zhang; methodology, Jifei Gao and Yuchen Zhang; software, Jifei Gao; validation, Jifei Gao and Yuchen Zhang; formal analysis, Jifei Gao; investigation, Jifei Gao; data curation, Jifei Gao; visualization, Jifei Gao; interpretation of results in the context of traditional Chinese medicine, Heng Aik Teng, Fan Yue and Fan Ying; writing—original draft preparation, Jifei Gao; writing—review and editing, Jifei Gao, Yuchen Zhang, Heng Aik Teng, Fan Yue and Fan Ying; project administration, Yuchen Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: NB-TCM-CHM Version 3 is publicly available from Mendeley Data under CC BY 4.0 (DOI: 10.17632/2kjmzjyrmd.3) [3]. The public code repository is https://github.com/litlelight/chm-transport. The revised public package provides the NB closed-set workflow, stronger architecture and adaptation baselines, calibration analyses, and the fully reproducible public 8-known/8-prior-unknown OOD sensitivity benchmark, together with split registries, checkpoints, per-sample scores, thresholds, class-level acceptance rates, features/logits, centroids, similarity matrices, calibration partitions, and analysis scripts. The original internal 29-class legacy archive, its raw images, and complete frozen artifact archive cannot be redistributed; results depending on that archive therefore remain fixed-reference analyses. Supplementary Materials and Supplementary Data S1 provide the detailed derived outputs used in the public sensitivity analyses.

Ethics Approval: Not applicable. This study used publicly available images of medicinal materials and did not involve human participants or animals.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/iasc.2026.091018/s1. Supplementary Data S1 provides detailed tabulated class mappings, audit/split summaries, run-level and detector-level outputs, thresholds, per-class acceptance summaries, held-out recalibration results, and class-aware geometry/acceptance statistics.

References

1. Xu Y, Wen G, Hu Y, Luo M, Dai D, Zhuang Y, et al. Multiple attentional pyramid networks for Chinese herbal recognition. Pattern Recognit. 2021;110(1):107558. doi:10.1016/j.patcog.2020.107558.
2. Sun X, Qian H, Xiong Y, Zhu Y, Huang Z, Yang F. Deep learning-enabled mobile application for efficient and robust herb image recognition. Sci Rep. 2022;12(1):6579. doi:10.1038/s41598-022-10449-9.
3. Tian D, Zhou C, Wang Y, Zhang R, Yao Y. NB-TCM-CHM: image dataset of the Chinese herbal medicine fruits and its application in classification through deep learning. Data Brief. 2024;54:110405. doi:10.1016/j.dib.2024.110405.
4. Hou Q, Yang W, Liu G. Chinese herbal medicine recognition network based on knowledge distillation and cross-attention. Sci Rep. 2025;15(1):1687. doi:10.1038/s41598-025-85697-6.
5. Tang Z, Fu Y, Li M, Liang H, Tang Y, Wen J. MPR-net: medicinal plant recognition network with dual-branch attention fusion. Pattern Recognit. 2026;171(1):112185. doi:10.1016/j.patcog.2025.112185.
6. Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 5389–400.
7. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst. 2019;32(2):13991–4002. doi:10.1007/978-3-030-59710-8_80.
8. Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, et al. WILDS: a benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 5637–64.
9. Gong Y, Lin X, Yao Y, Dietterich TG, Divakaran A, Gervasio M. Confidence calibration for domain generalization under covariate shift. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 8958–67. doi:10.1109/ICCV48922.2021.00883.
10. Yang J, Wang P, Zou D, Zhou Z, Ding K, Peng W, et al. OpenOOD: benchmarking generalized out-of-distribution detection. Adv Neural Inf Process Syst. 2022;35:32598–611. doi:10.52202/068431-2362.
11. Wang H, Li Z, Feng L, ViM ZW. Out-of-distribution with virtual-logit matching. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 4921–30. doi:10.1109/CVPR52688.2022.00487.
12. Park J, Jung YG, Teoh ABJ. Nearest neighbor guidance for out-of-distribution detection. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 1686–95. doi:10.1109/ICCV51070.2023.00162.
13. Kull M, Perello-Nieto M, Kängsepp M, Silva Filho T, Song H, Flach P. Beyond temperature scaling: obtaining well-calibrated multiclass probabilities with Dirichlet calibration. Adv Neural Inf Process Syst. 2019;32:12316–26.
14. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva, Switzerland: World Health Organization; 2021 [cited 2026 Sep 17]. Available from: https://www.who.int/publications/i/item/9789240029200.
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: Proceedings of the Ninth International Conference on Learning Representations; 2021 May 3–7; Vienna, Austria.
16. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: hierarchical vision Transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22. doi:10.1109/ICCV48922.2021.00986.
17. Zhou K, Yang Y, Qiao Y, Xiang T. Domain generalization with MixStyle. In: Proceedings of the 2021 International Conference on Learning Representations; 2021 May 4; Vienna, Austria.
18. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016;17(59):1–35. doi:10.1007/978-3-319-58347-1_10.
19. Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T. Tent: fully test-time adaptation by entropy minimization. In: Proceedings of the Ninth International Conference on Learning Representations; 2021 May 3–7; Vienna, Austria.

20. Niu S, Wu J, Zhang Y, Chen Y, Zheng S, Zhao P, et al. Efficient test-time model adaptation without forgetting. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 16888–905.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
22. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 6105–14.
23. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1314–24. doi:10.1109/ICCV.2019.00140.
24. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal. 2004;88(2):365–411. doi:10.1016/S0047-259X(03)00096-4.
25. Kalejahi BK, Rajabi MJ. Multimodal deep learning for papillary thyroid carcinoma diagnosis using ultrasound and cytology. BMC Med Imaging. 2026;26(1):449. doi:10.1186/s12880-026-02526-6.
26. Wiles O, Gowal S, Stimberg F, Rebuffi SA, Ktena I, Dvijotham K, et al. A fine-grained analysis on distribution shift. In: Proceedings of the Tenth International Conference on Learning Representations; 2022 Apr 25–29; Online.
27. Niu S, Wu J, Zhang Y, Wen Z, Chen Y, Zhao P, et al. Towards stable test-time adaptation in dynamic wild world. In: Proceedings of the Eleventh International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
28. Lasko TA, Strobl EV, Stead WW. Why do probabilistic clinical models fail to transport between sites. npj Digit Med. 2024;7(1):53. doi:10.1038/s41746-024-01037-4.
29. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. npj Digit Med. 2024;7(1):120. doi:10.1038/s41746-024-01085-w.
30. Schurz H, Solander K, Åström D, Cossío F, Choi T, Dustler M, et al. Simulating mismatch between calibration and target population in AI for mammography: the retrospective VAIB study. npj Digit Med. 2025;8(1):259. doi:10.1038/s41746-025-01623-0.
31. Rashid TA, Majidpour J, Thinakaran R, Batumalay M, Dewi DA, Hassan BA, et al. NSGA-II-DL: metaheuristic optimal feature selection with deep learning framework for HER2 classification in breast cancer. IEEE Access. 2024;12:38885–98. doi:10.1109/ACCESS.2024.3374890.
32. Subramanian RR, Sudharsan RR, Vairamuthu B, Dewi DA. Neural network models for diagnosing recurrent aphthous ulcerations from clinical oral images. Sci Rep. 2025;15(1):29519. doi:10.1038/s41598-025-06951-5.
33. Ponraj A, Nagaraj P, Balakrishnan D, Srinivasu PN, Shafi J, Kim W, et al. A multi-patch-based deep learning model with VGG19 for breast cancer classifications in the pathology images. Digit Health. 2025;11:20552076241313161. doi:10.1177/20552076241313161.