Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 3, No. 3 (2026) 153–163.

Paper Type: Original Article

Image Covariates and Label Noise Robustness in Dermatoscopic Classification

Seyyed Ahmad Edalatpanah[1,*], Ismail A. Mageed[2], Lorenzo Cevallos-Torres[3]

[1] Department of Industrial Engineering, Ayandegan University, Tonekabon, Iran; s.a.edalatpanah@aihe.ac.ir.

[2] Faculty of Engineering and Digital Technologies, School of Computer Science, AI, and Electronics, University of Bradford, Bradford, UK; iamoham@bradford.ac.uk.

[3] Department of Mathematics, Brigham Young University, UT 84602 Provo, USA; lorenzo.cevallost@ug.edu.ec.

Citation:

Received: 19 November 2025 *Revised: 17 March 2026* *Accepted: 21 May 2026*	Edalatpanah, S. A., Mageed, I. A, & Cevallos-Torres, L. (2026). Image covariates and label noise robustness in dermatoscopic classification. *Soft Computing Fusion with Applications, 3*(3), 153-163.

Abstract

Medical image classification depends on labels that can contain diagnostic errors. This study compares six training strategies on DermaMNIST, with 7,007 training images in seven classes and synthetic symmetric, asymmetric, and pixel-difficulty-dependent label noise at nominal rates of 10–40%. The comparison comprises 36 noise-family, rate, and seed combinations, using three seeds and a seven-epoch budget per main network. The instance-dependent-minus-symmetric ERM balanced-accuracy gap changes from +0.063 at 10% to −0.061 at 40%; instance dependence is therefore not uniformly harder for this generator. Image-covariate-only noise scores have pooled AUC 0.502, whereas the combined covariate-and-loss estimator has AUC 0.301 compared with 0.832 for the loss-GMM score. The MIND-CT pipeline has higher reported mean balanced accuracy than its MIND and FC variants, but lower overall means than Co-teaching and ERM: 0.244 versus 0.284 and 0.276. At the designated INST-20% endpoint, Co-teaching reaches 0.343 versus 0.303 for MIND-CT. These findings support a scoped comparison of estimator quality and downstream classification, rather than a conclusion about genuine clinical metadata. Inferential results are exploratory because the grid reuses one archive and three seeds, and source records are needed to reconcile the rank diagram and verify implementation details.

Keywords: Label noise, Instance-dependent noise, Medical image classification, Dermatology, Co-teaching, Forward correction, DermaMNIST, Negative results.

1 | Introduction

Medical training labels may reflect observer disagreement, imperfect diagnoses, or automated extraction. Wei et al. [1] review label-noise detection, handling, and evaluation in medical prediction, while Hekler et al. [2] examine label noise in binary skin-cancer classification. CheXpert explicitly includes uncertainty labels derived from radiology reports [3]. These sources motivate evaluating label quality alongside predictive performance;

Corresponding Author: s.a.edalatpanah@aihe.ac.ir

https://doi.org/10.22105/scfa.v3i3.107

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

they do not establish that every instance-dependent mechanism is more damaging than class-conditional corruption.

Three questions organize the present comparison: How the distribution of corruptions affects performance at matched nominal rates; whether an explicit per-sample noise score identifies injected errors; and whether the resulting pipeline improves on simpler baselines under the stated training schedule. Matched-rate controls, injected-mask scoring, and full-pipeline comparisons address distinct questions. The hypotheses and endpoint are designated in the protocol described here; no external preregistration record is established.

The tested hypotheses concern severity, detectability, and downstream value. H1 predicts greater damage from the specified instance-dependent generator at matched nominal rates. H2 predicts an improvement from adding the available image covariates to loss-based estimation. H3 predicts that MIND-CT improves classification relative to its component variants and simple baselines. These hypotheses concern the implemented generators and features on this archive, rather than all clinical labeling processes.

DermaMNIST comprises 28×28 dermatoscopic images derived from HAM10000 and distributed through MedMNIST [4], [5]. Noise is injected into training labels only. The symmetric and asymmetric generators use class-conditional flip probabilities, whereas the instance-dependent generator weights corruption by pixel-space distance from a class centroid. Six strategies are compared: ERM, confidence filtering, the class-constant variant labeled FC, MIND soft-target training, Co-teaching [6], and MIND-CT with paired networks and agreement regularization. The grid contains 36 combinations (3 families × 4 rates × 3 seeds), plus clean-label references. Equal main-network epochs provide a common schedule, but warm-up networks, paired models, and filtering mean that total computational costs need not be equal.

The observed pattern is mixed. The ERM gap against symmetric corruption changes sign across rates, so H1 is not uniformly supported for this generator. The pooled AUC values do not support H2 for image covariates: Metadata-only scores are near chance and the combined score is below chance. MIND-CT has higher reported means than MIND and FC, but lower means than Co-teaching and ERM. This supports distinguishing component contrasts from full-pipeline value. It does not identify the cause of the combined estimator's poor ranking or establish the value of clinical age, sex, or site information, which is absent here.

2 | Related Work

2.1 | Label Noise in Medical Imaging

Clinical label-noise evaluation involves both incorrect labels and unresolved diagnostic uncertainty. The medical scoping review distinguishes detection, handling, and evaluation [1], and Hekler et al. [2] study sensitivity to introduced errors in skin-cancer classification [2]. CheXpert provides a separate setting with report-derived uncertainty labels [3]. Such natural uncertainty is not tested in the present synthetic-noise campaign.

2.2 | Robust-Learning Methods

Co-teaching exchanges small-loss samples between two networks [6]. DivideMix combines per-sample loss modeling with semi-supervised training [7]. Confident learning estimates a joint distribution of noisy and uncorrupted labels under class-conditional assumptions [8]. Patrini et al. [9] propose forward correction by transforming predicted class probabilities through a transition matrix before computing the loss. The FC variant described here instead uses transition-based target softening; without the executable loss expression, it should not be equated with canonical forward correction. The comparison therefore concerns the stated variants and training budget, rather than optimal performance of entire method families.

2.3 | Instance-Dependence and its Measurement

Class-conditional corruption depends on the clean class, whereas instance-dependent corruption may also vary with the individual input. In this experiment that dependence is introduced through pixel-space centroid distance. Matching the average flip probability separates nominal prevalence from the allocation of errors, provided every resulting probability is valid. A clinical interpretation requires additional evidence linking the chosen difficulty score to real diagnostic errors.

Loss-based scoring and sample selection use training behavior to rank suspected errors. Co-teaching uses small-loss exchange [6], while DivideMix models loss distributions [7]. Turning a ranking score into a noise probability adds a calibration requirement. The injected mask permits direct AUC and threshold-based evaluation here, but AUC alone neither calibrates a posterior probability nor distinguishes intrinsic sample difficulty from label error.

3 | Problem Setting and Noise Models

Let the training archive be D = {(x_i, y_hat_i)}, i = 1, ..., n = 7007, with x_i a real dermatoscopic patch and y_hat_i the observed (possibly corrupted) label in K = 7 classes; a clean validation and test set (1,003 and 2,005 images) is never corrupted. Noise enters only through the label corruptions:

$$\Pr(\hat{y}_i = y_i \mid x_i, y_i) = 1 - \rho_i, \quad \Pr(\hat{y}_i \neq y_i \mid x_i, y_i) = \rho_i. \tag{1}$$

Symmetric class-conditional noise (SYM) sets $\varrho_i = r$ and chooses destinations uniformly among other classes. Asymmetric Noise (ASYM) uses $\varrho_i = r$ with a fixed destination map; the complete map must be specified to reproduce the experiment. Describing this map as clinically plausible requires clinical evidence. Instance-dependent noise (INST) uses the difficulty-weighted expression below:

$$\rho_i = \frac{r\,\mathrm{sigmoid}(a z_i)}{\frac{1}{n}\sum_{j=1}^{n}\mathrm{sigmoid}\left(a z_j\right)}, \quad z_i = \mathrm{standardize}\big(\| x_i - \mu_{y_i} \|\big). \tag{2}$$

Here μc is the training-set pixel centroid of class c and a = 2.2. If all ϱ_i lie in [0,1], the normalization gives their mean as r. This validity condition is essential: Normalized sigmoid weights can otherwise produce probabilities exceeding one. Clipping changes the mean and requires a further calibration step; whether clipping was used in the reported runs must be established from the generator. The reported realized fraction 0.198 at an INST-20% setting does not verify this condition for all 36 runs. Balanced accuracy is the primary metric because of class imbalance (4,693 of 7,007 training images are nevus; the rarest class has 80).

4 | Training Strategies and the Noise Estimator

4.1 | The Six Strategies

Preprocessing normalizes image values to [0,1] without augmentation. This provides a common input pipeline, but the effects of augmentation on the compared strategies are not tested. Confusion matrices and ordinary accuracy would support a per-class audit if the run records are supplied; the tables here report balanced accuracy.

All strategies are described as using a compact convolutional backbone with three convolution-and-batch-normalization blocks, a 96-channel trunk, and a dropout head (0.3), approximately 350K parameters. Main training uses Adam, learning rate 2e−3, weight decay 1e−4, batch size 256, and seven epochs. ERM uses cross-entropy on corrupted labels. CL-conf discards the lowest-confidence r-fraction after warm-up. The FC variant uses a class-constant transition estimate from out-of-fold warm-up predictions and noisy labels, mixed 0.3/0.7 with the identity. Transition-based target softening is distinct from the canonical forward-corrected loss $-\log[(T^{\mathrm{T}}q)\hat{y}]$ [9]. The exact implemented expression and transition orientation are necessary to interpret the FC observations.

Co-teaching deserves one design note: the keep-rate schedule follows the original paper's linear decay from 1 to 1-r across epochs, using the KNOWN nominal rate r. This is the standard evaluation convention and it advantages selection methods under synthetic noise (deployment would require estimating r); we keep the convention because our question is relative ordering under matched information about the noise budget, not deployment realism of any single method. The same convention governs CL-conf's filter fraction and the GMM threshold in the estimator audit, so every method that consumes a rate receives the true one.

MIND uses a per-sample score to construct the soft target:

$$t_i = (1 - \hat{p}_i)\text{onehot}(\hat{y}_i) + \frac{\hat{p}_i}{K-1}\left(\mathbf{1} - \text{onehot}(\hat{y}_i)\right), \quad (3)$$

where $\hat{p_i}$ is a score intended to represent label-error probability. *Eq. (3)* assigns the remaining mass uniformly across other classes; it is adaptive label smoothing, not an identified inverse of the corruption process. Co-teaching exchanges peer-selected small-loss samples [6]. MIND-CT combines these soft targets with paired training and the agreement term:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda\left[D_{KL}\left(q_1 \| \text{stopgrad}(q_2)\right) + D_{KL}\left(q_2 \| \text{stopgrad}(q_1)\right)\right], \quad \lambda = 0.1, \quad (4)$$

where detached predictions prevent gradients from propagating through the target of each directed KL term. The reported final classifier is selected from the pair using the clean validation split. This selection consumes validation labels; the estimator's claim of using no clean labels applies only to its fitting stage.

4.2 | The Per-Sample Noise Estimator (and Its Failure Mode)

The estimator uses two out-of-fold warm-up networks, with two epochs per fold, to obtain losses and confidences for training samples. A two-component Gaussian mixture on log-loss identifies the higher-mean component as the suspected-noise group. Four scores are evaluated: Its posterior, confidence-only, covariate-only, and a combined logistic-regression score trained on GMM pseudo-labels. Covariates are channel means and standard deviations, brightness, and gradient energy; clinical age, sex, and site are absent. Out-of-fold base predictions do not make the subsequent logistic-regression scores out of sample if that regression is fitted and scored on the same points. The reported AUC quantifies ranking against the injected mask, but it does not establish probability calibration or prove that pseudo-label overfitting caused the below-chance combined score. Score polarity, pseudo-label assignment, and fold-specific transformations require inspection.

5 | Experimental Protocol

Reproducibility requires the exact destination map, random seeds, generator validity checks, train/validation/test indices, per-run metrics, and loss implementations. The described protocol uses cell-specific seeds and records realized corruption fractions. The supplied manuscript alone does not establish an executable release or a timestamped preregistration. The empirical scope is one archive; the proposed three-archive and natural-uncertainty extensions were not executed.

The grid crosses {SYM, ASYM, INST}, nominal rates {10%,20%,30%,40%}, and seeds {11,12,13}, giving 36 combinations plus three clean-label references. Balanced accuracy is evaluated on the 2,005-image test split; INST-20% is the designated endpoint. Reported analyses use Friedman ranking, Nemenyi comparisons [10], and Wilcoxon signed-rank contrasts with Holm adjustment [11]. These are exploratory: Repeated rates and families on one archive and reused seeds do not constitute 36 independent datasets. Wilcoxon tests signed ranks under assumptions on paired differences, not the sign of their arithmetic mean. Mask-level AUC, precision, and recall assess error scoring separately. The stated 39 settings and roughly 430 network trainings require run logs for verification; equal epochs per network do not establish equal total compute.

Table 1. Six training strategies with a common backbone and main-network epoch schedule.

Strategy	Family	Uses Per-Sample Noise Estimate	Co-Training
ERM	Baseline	No	no
CL-conf	Selection	No (confidence rank)	no
FC	Loss correction	No (class-constant T)	no
MIND	Instance-dep. correction	Yes (metadata+loss)	no
Co-teaching	Co-training selection	No (small-loss)	yes
MIND-CT	Proposed pipeline	Yes (metadata+loss)	yes + agreement

6 | Results

6.1 | Overall Comparison and the Primary Endpoint

Table 2 reports the pooled mean ordering as Co-teaching (0.284±0.086), ERM (0.276±0.065), MIND-CT (0.244±0.058), confidence filtering (0.227), MIND (0.220), and FC (0.187). The clean-label reference mean is 0.371 (SD 0.065; seed values 0.411, 0.407, and 0.296). The reported Friedman result is $\chi^2(5)=54.4$, $p=1.7e{-}10$, and *Table 3* retains the reported Holm-adjusted signed-rank p-values. These inferential summaries require the per-run records and an appropriate dependence analysis. *Fig. 3's* rank orientation and stated critical difference are not reconciled with the source claim that MIND-CT ranks best; no rank-based winner is inferred here. At INST-20%, the reported means are 0.343 for Co-teaching, 0.303 for MIND-CT, 0.291 for MIND, and 0.281 for ERM.

The family-specific columns of *Table 2* show different orderings: Co-teaching has the highest displayed mean under SYM and INST, whereas ERM leads under ASYM. The FC and MIND observations concern their stated implementations. Differences between families do not identify why an estimator or strategy succeeds. Pooled clean-minus-noisy degradation ranges from the reported +0.087 for Co-teaching to +0.184 for FC.

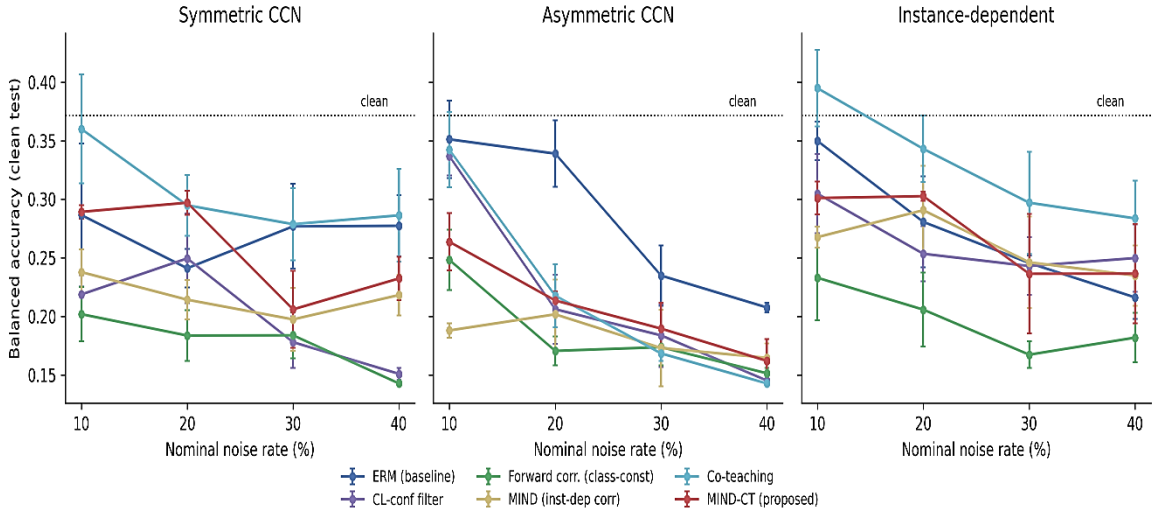

Fig. 1. Balanced accuracy versus nominal noise rate under the three noise families (error bars: SE over 3 seeds; dotted line: Clean-label mean).

Table 2. Balanced accuracy on the clean test set (mean±sd over blocks; degradation is the clean mean minus the overall mean).

Strategy	Overall (36 Blocks)	SYM	ASYM	INST	Primary: INST 20%	Degradation vs. Clean
ERM (plain CE)	0.276±0.065	0.270	0.283	0.273	0.281±0.067	+0.096
CL-conf filter	0.227±0.067	0.199	0.218	0.263	0.254±0.041	+0.145
FC (class-constant)	0.187±0.043	0.178	0.186	0.197	0.206±0.055	+0.184
MIND (instance-dep.)	0.220±0.051	0.217	0.182	0.260	0.291±0.066	+0.152
Co-teaching	0.284±0.086	0.305	0.218	0.330	0.343±0.049	+0.087
MIND-CT (proposed)	0.244±0.058	0.256	0.207	0.269	0.303±0.007	+0.127

Table 3. Paired wilcoxon signed-rank tests over the 36 blocks (positive Δ = MIND-CT better).

Comparison (MIND-CT - Alternative)	Mean Δ Balanced Acc.	Wilcoxon P (Holm)
MIND-CT vs ERM (plain CE)	-0.031	3.51e-02
MIND-CT vs CL-conf filter	+0.018	9.22e-02
MIND-CT vs FC (class-constant)	+0.057	4.02e-06
MIND-CT vs MIND (instance-dep.)	+0.025	7.02e-03
MIND-CT vs Co-teaching	-0.040	6.65e-03

6.2|RQ1: Is Instance-Dependent Noise more Damaging?

Fig. 2 and *Table 4* compare ERM across matched nominal rates. The INST-minus-SYM differences are +0.063, +0.040, −0.031, and −0.061 as the rate increases from 10% to 40%. Against ASYM, the differences are −0.001, −0.058, +0.011, and +0.009. Thus this pixel-difficulty generator is not uniformly more damaging than either class-conditional control. A mechanism in which easy examples remain intact is plausible, but these aggregate results do not isolate that mechanism or establish a general ordering of noise families.

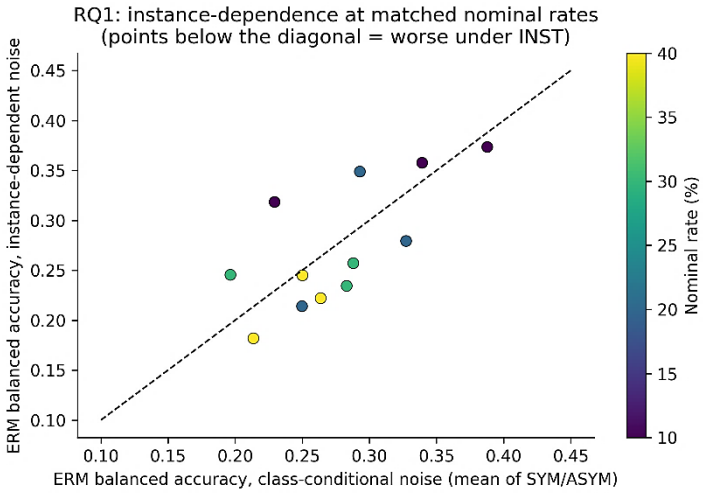

Fig. 2. ERM balanced accuracy under instance-dependent noise versus class-conditional noise at matched nominal rates (each point: One seed x rate; below diagonal = INST harder).

Table 4. RQ1: ERM balanced accuracy under matched-rate noise models (mean over 3 seeds).

Nominal Rate	ERM @ INST	ERM @ SYM	ERM @ ASYM	INST-SYM	INST-ASYM
10%	0.350	0.287	0.351	+0.063	-0.001
20%	0.281	0.241	0.339	+0.040	-0.058
30%	0.246	0.277	0.235	-0.031	+0.011
40%	0.216	0.277	0.208	-0.061	+0.009

6.3 | RQ2: Does the Metadata Channel Help Noise Estimation?

Table 5 reports loss-GMM AUC 0.832±0.103 overall and 0.893 on INST runs. Confidence-only and covariate-only pooled AUC values are 0.501 and 0.502. The combined logistic-regression score has AUC 0.301±0.277, below the 0.5 chance reference, with a reported paired p-value 7.4e−09 against loss-GMM. This is a poor ranking in the stated scoring direction, not proof that the covariates contain no information. Reversing a score would change AUC to 1−AUC, but doing so after inspecting masks would be an oracle intervention, not a validated deployment procedure. Score polarity, pseudo-label assignment, calibration, and in-sample fitting are alternative issues to examine. The downstream contrasts do not identify which of these explains performance.

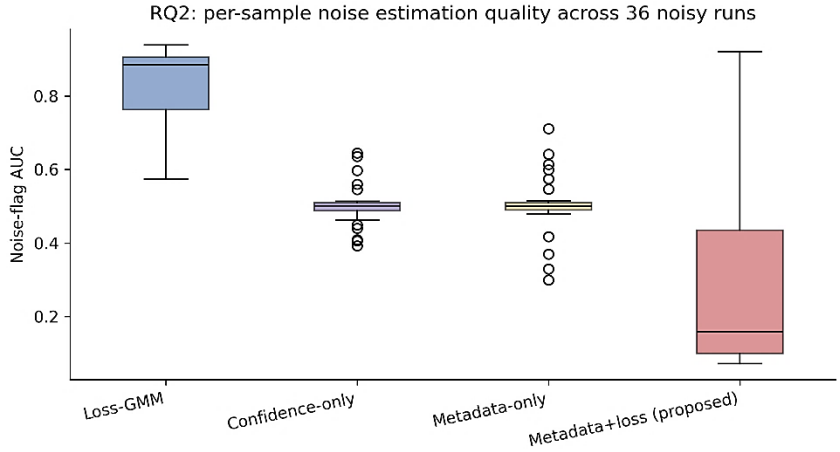

Fig. 4. Noise-flag AUC of the four estimators across the 36 noisy runs (box = quartiles; whiskers = range).

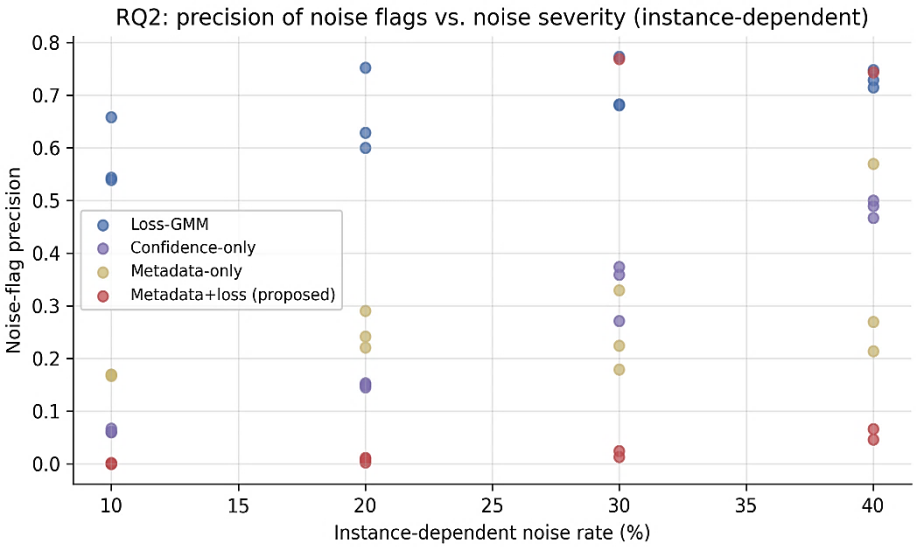

Fig. 5. Noise-flag precision versus nominal rate on instance-dependent runs.

Table 5. RQ2: Noise-flag quality against the true injected mask (thresholds set at the true noise prevalence).

Estimator	AUC (36 Runs)	AUC (INST Runs)	Precision	Recall
Loss-GMM (baseline)	0.832±0.103	0.893	0.572	0.572
Confidence-only	0.501±0.051	0.502	0.252	0.252
Metadata-only	0.502±0.074	0.511	0.250	0.250
Metadata+loss (proposed)	0.301±0.277	0.238	0.151	0.151

6.4 | Component and Pipeline Contrasts

Table 6 contains four system contrasts. MIND-CT versus MIND combines paired training and agreement regularization, so +0.025 does not isolate an agreement effect. MIND versus FC compares different target constructions and estimator use (+0.033). Co-teaching versus confidence filtering compares dynamic peer exchange with static filtering (+0.057). MIND-CT versus Co-teaching has a reported mean difference−0.040. The p-values are retained as exploratory summaries of these contrasts; they do not establish that every added component independently earns its place.

An agreement regularizer can encourage similar predictions, but no isolated λ=0 versus λ=0.1 comparison is reported. Likewise, *Eq. (3)* defines a soft target without a bound on its error relative to the unknown clean-label distribution. Its mass on alternative classes does not imply that its error is at most $\hat{p_i}$. Mechanistic claims about variance reduction or superiority of instance-dependent correction require targeted controls and estimator-calibration evidence.

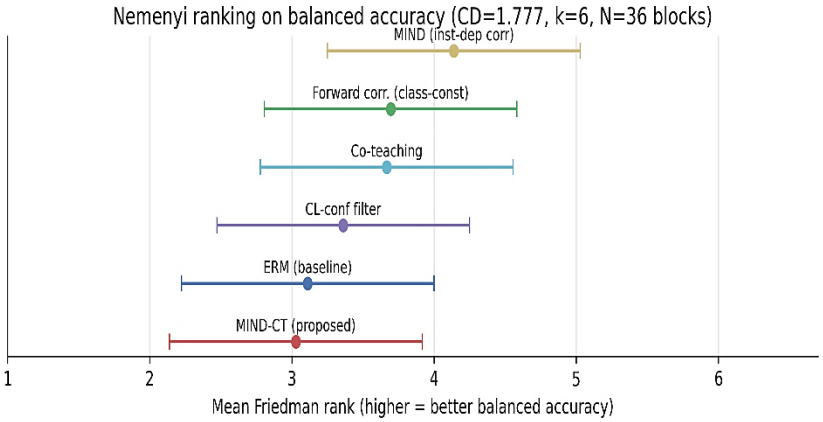

Fig. 3. Source rank display, retained descriptively.

Its axis states that higher ranks are better, while the source prose used the opposite orientation. The displayed CD=1.777 also requires verification; under the conventional normalized Nemenyi constant q=2.850, k=6 and N=36 give a conditional CD≈1.257. Neither value resolves the dependence of the grid cells (*Fig. 3*).

Table 6. Reported system contrasts over 36 grid combinations (positive Δ = A better).

Ablation (A - B)	Mean Δ Balanced Acc.	Wilcoxon P
MINDCT vs MIND	+0.025	2.34e-03
MIND vs FC	+0.033	5.10e-04
CT vs CLCONF	+0.057	2.44e-04
MINDCT vs CT	-0.040	1.66e-03

6.5 | Variance and What a Fixed Budget Hides

Class imbalance motivates balanced accuracy. Under independent 40% symmetric flips, the expected number of uncorrupted labels among 80 rare-class training images is 48. This is an expectation, not an observed class count or a verified recall result. A class-specific analysis requires the confusion matrices and realized masks.

Clean-label balanced accuracy spans 0.296–0.411 across the three seeds. With only three seeds per cell, *Fig. 1's* standard errors are imprecise. Pooling the grid does not remove dependence created by reusing one archive. Seven epochs define a short-budget comparison, and longer training may change the ordering. The results therefore describe this schedule, rather than asymptotic performance or a compute-matched deployment recommendation.

7 | Discussion

The comparison shows that higher component-level means need not imply a better complete pipeline. Co-teaching and ERM have higher pooled means than MIND-CT, while the explicit combined noise score ranks poorly against the mask. This motivates auditing scores before relying on their downstream value. It does not establish that explicit probability estimation generally fails or that image covariates cannot help another archive.

A held-out fit-and-score design for the combined estimator would reduce direct reuse of fitting points. It provides no guarantee of matching or exceeding loss-GMM AUC. Such a variant, and a score-orientation audit, must be run and reported before assigning the observed failure to overfitting.

The rate-dependent sign changes suggest that evaluations should vary both prevalence and how errors are allocated. Real diagnostic ambiguity, site conventions, and annotator effects may differ from pixel-centroid difficulty. Testing those mechanisms requires corresponding data and generators.

Matched nominal rates provide useful controls only when generator probabilities and realized rates are documented. A family-by-rate grid exposes changes hidden by a single-cell comparison, but conclusions remain conditional on the chosen generators and archive.

Component contrasts and simple baselines answer complementary questions. Here the reported gains against MIND and FC coexist with lower pooled performance than Co-teaching and ERM. This supports reporting both comparisons before interpreting added complexity.

8 | Limitations and Threats to Validity

The empirical scope is one low-resolution archive with synthetic training-label corruption. Clinical metadata and naturally uncertain labels are absent. MedMNIST is intended as a lightweight benchmark rather than a clinical deployment evaluation [4]. The ISIC archive [12] and CheXpert [3] are possible extensions, not executed validations.

The common main-network epoch budget does not match total model fits or warm-up compute. The FC description differs from canonical forward correction [9], so its observations cannot be called a lower bound for that entire family. Three seeds and overlapping grid conditions limit inference. The designated endpoint can be reported transparently, but its pre-execution status requires a dated protocol record.

Thresholds use the known injected prevalence, so threshold-based results are oracle-assisted. Mean precision and recall 0.572 mean that 57.2% of flags are correct and 57.2% of corruptions are recovered on average over the reported runs; they do not establish a three-quarter detection rate at 20% prevalence. Equal precision and recall are expected when the number flagged equals the number corrupted. Confident learning provides a class-conditional route to estimating label-error distributions [8], but its assumptions need examination under instance dependence.

The source tables and figures require traceability to run records, injected masks, checkpoints, and the analysis code. Programmatic document generation can reduce transcription errors, but it cannot guarantee validity or prevent incorrect statistics, score polarity, or interpretation. Data/code availability and author-approved declarations must be established before publication.

Pooled means, signed-rank tests, and mean ranks summarize different features of the paired results. *Fig. 3's* orientation and critical difference remain unreconciled, so it is descriptive only. Dependence across cells, seed variability, and comparisons across multiple research questions require an explicit inferential plan. Reported p-values are exploratory and do not by themselves establish mean-direction effects or clinical generalization.

9 | Conclusion

On DermaMNIST under the stated synthetic-noise generators and short training budget, instance-dependent corruption is not uniformly more damaging than class-conditional corruption. Image covariates do not improve the reported pooled noise-score AUC, and the combined score is below chance in its stated direction. MIND-CT has higher reported means than MIND and FC, but lower pooled means than Co-teaching and ERM. The useful contribution is the separation of mask-level score evaluation from downstream classification. Clinical metadata, natural label uncertainty, isolated agreement effects, compute-matched baselines, and independent validation remain future tests. The present conclusions also require generator validity checks, exact implementation records, and reconciliation of the rank display before publication.

Funding

This research received no external funding.

Data Availability

Sufficient data is available upon request.

Conflict of Interest

No potential conflict of interest was reported by the authors.

Consent for Publication

All authors have provided their consent for the publication of this manuscript.

References

[1] Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., & Peng, Y. (2024). Deep learning with noisy labels in medical prediction problems: A scoping review. *Journal of the American medical informatics association*, *31*(7), 1596–1607. https://doi.org/10.1093/jamia/ocae108

[2] Hekler, A., Kather, J. N., Krieghoff-Henning, E., Utikal, J. S., Meier, F., Gellrich, F. F., ... & Brinker, T. J. (2020). Effects of label noise on deep learning-based skin cancer classification. *Frontiers in medicine*, *7*, 177. https://doi.org/10.3389/fmed.2020.00177

[3] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597). https://doi.org/10.1609/aaai.v33i01.3301590

[4] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., ... & Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, *10*(1), 41. https://doi.org/10.1038/s41597-022-01721-8

[5] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, *5*(1), 180161. https://doi.org/10.1038/sdata.2018.161

[6] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., … & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, *31*. https://doi.org/10.48550/arXiv.1804.06872

[7] Li, J., Socher, R., & Hoi, S. C. H. (2020). *Dividemix: Learning with noisy labels as semi-supervised learning*.. https://doi.org/10.48550/arXiv.2002.07394

[8] Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of artificial intelligence research*, *70*, 1373–1411. https://doi.org/10.1613/jair.1.12125

[9] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1944-1952). IEEE. https://doi.org/10.48550/arXiv.1609.03683

[10] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*(1), 1–30. https://dl.acm.org/doi/epdf/10.5555/1248547.1248548

[11] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

[12] International Skin Imaging Collaboration. (2026). *ISIC challenge data*. https://challenge.isic-archive.com/data/