Imaging and Radiation Research 2026, 9(1), 026410003.
https://doi.org/10.24294/irr026410003

Article

Transferability of cost-sensitive threshold policies in pneumoniamnist: Calibration, resampling, and image-integrity sensitivity

Arvy Noleal Osma[1,*], Mark Andre Balao[2]

[1] Mariners Polytechnic Colleges Foundation of Canaman Cam Sur, Canaman 4402, Philippines
[2] World Citi Colleges, Quezon City 1109, Philippines
*** Corresponding author:** Arvy Noleal Osma, aosma@marinerscanaman.edu.ph

CITATION

Osma A N, Balao M A. Transferability of cost-sensitive threshold policies in pneumoniamnist: Calibration, resampling, and image-integrity sensitivity. Imaging and Radiation Research. 2026; 9(1): 026410003. https://doi.org/10.24294/irr026410003

ARTICLE INFO

Received: 11 October 2026
Accepted: 29 October 2026
Available online: 30 October 2026

COPYRIGHT

Copyright © 2026 by author(s). *Imaging and Radiation Research* is published by EnPress Publisher, LLC. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: Background: Validation-selected thresholds may fail to retain their apparent error-cost advantage across imaging benchmark partitions, even with satisfactory discrimination. Objective: To examine transferability of fixed, analytical cost-based, and validation-selected classification policies in PneumoniaMNIST, with threshold-selection uncertainty and sensitivity analyses. Methods: We analyzed 5,856 28 × 28 pediatric chest radiographs (4,708 training; 524 source validation; 624 testing). Eight exact train-validation overlaps were excluded from validation procedures. A 100-component logistic baseline and compact convolutional neural network (CNN) were evaluated under four hypothetical false-negative/false-positive cost ratios. Revision-stage analyses included 500 validation-threshold resamples, 1,000 paired test resamples, exact-score threshold searches, class-composition standardization, calibration diagnostics, three CNN seeds, training-duplicate removal, and a development-role-separated sensitivity experiment. Results: Logistic and temperature-scaled CNN test AUCs were 0.921 and 0.925. At equal costs, the logistic tuned-versus-analytical loss difference was −0.0192 per image, whereas the CNN loss difference was +0.0144. CNN temperature scaling increased Brier score from 0.1303 to 0.1325 and test log loss from 0.5524 to 0.5951. At a 2:1 cost ratio, the logistic validation threshold was 0.13, with a bootstrap interval of 0.05–0.63. Conclusions: Threshold-policy gains were model- and cost-dependent, and validation selection was unstable. All new comparisons are retrospective exploratory benchmark analyses; patient-level independence and clinical utility remain unverified.

Keywords: medical image analysis; pediatric pneumonia; statistical decision theory; probability calibration; bootstrap uncertainty; benchmark transportability

1. Introduction

Public pediatric chest radiograph benchmarks make it possible to study that distinction reproducibly, but they also impose important constraints. PneumoniaMNIST comprises compressed 28 × 28 images derived from a historical pediatric image collection [1–3]. Benchmarks of this type are not a substitute for full-resolution radiology evaluation; image-level splits without patient identifiers cannot demonstrate patient-level separation. In the original experiment, eight exact training–validation image overlaps were detected and removed from validation procedures, but 18 repeated records remained within training. These features necessitate an explicit data-integrity audit rather than generic claims of independent observations.

A classifier can rank images effectively yet deliver different decisions under different operating thresholds. Discrimination reflects the ordering of scores, probability calibration concerns correspondence between scores and observed labels, and decision loss additionally depends on the cost assigned to false-negative and false-

positive classifications [4–8]. These properties are related but cannot be reduced to a single accuracy estimate. In imaging research, defaulting to a 0.50 cutoff can obscure the effect of explicit cost assumptions on operational decisions.

The methodological gap addressed here is the transferability of operating decisions selected using development observations: a cutoff with minimum observed validation loss need not retain that advantage on a held-out sample, even if both samples share a nominal dataset source. Selection variability, imperfect calibration, model fitting variability, class composition, and grid specification are potential contributors; this study is not designed to establish any one of them as the causal explanation. The question has direct relevance to transparent reporting principles in prediction-model research, although this paper does not purport to validate a clinical prediction system.

The objective is to compare three cost-sensitive decision policies on the same held-out PneumoniaMNIST images: the conventional cutoff of 0.50; the theoretical cost threshold $1/(1 + r)$; and a validation-selected threshold minimizing normalized loss. We then quantify conditional test uncertainty and the variability of threshold selection, examine probability scaling, compare discrete threshold grids with score-defined boundaries, standardize loss across hypothetical case fractions, and evaluate three CNN seeds, removal of training duplicates, and an exploratory role-separated development scheme. The third-policy comparison and additional sensitivity analyses were introduced during revision after the original test data had already been evaluated; all resulting inferences are explicitly exploratory.

2. Materials and methods

2.1. Study source, partitions, and integrity audit

The public PneumoniaMNIST image archive, released as part of MedMNIST v2 and archived on Zenodo (10.5281/zenodo.10519652), was obtained in NPZ format [1,9]. Its file MD5 checksum was 28209eda62fecd6e6a2d98b1501bb15f. The images are 28×28 grayscale arrays and the label encoding is normal = 0 and pneumonia = 1. The original creator-provided partitions were retained: 4,708 training records, 524 validation records, and 624 test records. A cryptographic BLAKE2b digest of the complete 784-byte pixel array was calculated for each image. Identical digests were grouped by partition; eight validation records whose digests matched training records were excluded, leaving 516 validation images. The train and test sets had no exact matching pixel arrays, but no patient identifiers were available to check for repeated patients or acquisition episodes.

Eighteen repeated training records were retained for the reconstructed original analysis, with a separate deduplicated-training sensitivity analysis was fitted to 4,690 unique pixel arrays. An exploratory difference-hash (dHash) nearest-neighbor screen used images downsampled to 9×8 pixels, 64 horizontal comparison bits, and a minimum cross-partition Hamming distance. Because these tiny radiographs produce many coincident, low-information hashes, dHash results indicate image-pattern similarity only and were not used as evidence of duplicate patients or for test-set exclusions.

Table 1. Audit and class composition of the authenticated PneumoniaMNIST archive.

Split	Normal	Pneumonia	Total	Quality-control disposition
Training	1,214	3,494	4,708	4,690 unique exact-pixel records
Original validation	135	389	524	8 exact overlaps with training
Audited validation	135	381	516	Used for selection, scaling and threshold analysis
Held-out test	234	390	624	No exact cross-split pixel matches

2.2. Models and training procedure

The PCA-logistic reference model used image intensities divided by 255, training-only standardization, 100 randomized principal components (seed 20261008), and logistic regression with inverse regularization C = 1.0 and a maximum of 1,500 optimizer iterations. The nonlinear comparator used two 3 × 3 convolutional blocks with 12 and 24 channels, rectified linear units, 2 × 2 max-pooling, and a 48-node dense layer with a single logit output. Optimization used Adam (learning rate 0.001, weight decay 0.0001), a minibatch size of 64, a maximum of 12 epochs, and binary cross-entropy with logits. In the original analysis, the CNN checkpoint minimizing audited-validation loss was retained, with training stopping after four consecutive epochs without at least 0.0001 improvement. Random seed 20261008 defined network initialization and batch ordering.

The original comparator used the 516-image audited validation partition for early stopping, temperature fitting, and threshold selection. That information reuse is retained as a historical reference and not mischaracterized as independent calibration. Two additional training seeds (20261009 and 20261010) were run for the CNN, each independently fitted but using the same validated procedure. A separate fitting sensitivity analysis excluded repeated exact-pixel training records from both logistic regression and the CNN (seed 20261008). These do not replace external testing.

2.3. Probability scaling, calibration and discrimination

Logistic regression probabilities were evaluated directly. CNN logits z were converted to $p = \mathrm{sigmoid}(z)$ before scaling, and to $p = \mathrm{sigmoid}(z/T)$ after validation-fitted temperature scaling. The positive temperature T was minimized using audited-validation binary log loss via bounded scalar minimization of log T. Binary discrimination was summarized by ROC AUC and average precision; probability performance by the Brier score and log loss. Because extreme 32-bit scores can round to exactly one, the reported log loss was recomputed consistently from exported predictions using a declared probability floor of 10^{-7} and 64-bit arithmetic; precision sensitivity is tabulated in the supplementary outputs. Calibration was also described by the intercept and slope obtained when regressing the binary test label against the logit of the predicted probability; ideal estimates are intercept 0 and slope 1. Two hundred class-stratified test bootstrap samples produced descriptive percentile intervals for these calibration coefficients. These are internal benchmark calibration measures, not transportability validation.

2.4. Comparison of three decision policies

Let y = 1 denote a pneumonia-labeled benchmark image, *p* its model-estimated probability, and r = C_FN / C_FP a hypothetical relative cost of false negatives to false positives. With C_FP normalized to 1, the empirical loss per image is L(t;r) = [r FN(t) + FP(t)] / *n*, with a positive prediction defined unambiguously as $p \geq t$. The Bayes-optimal decision threshold for target-population-calibrated probabilities, fixed class-conditional costs, and the stated loss rule is t* = 1/(1 + r) [7]. Analytical optimality is not assumed to hold for the fitted benchmark scores.

$$L(t; r) = \{r\ FN(t) + FP(t)\} / n, \qquad \text{decision} = 1 \text{ if } p \geq t. \tag{1}$$

$$t^* = C_FP / (C_FP + C_FN) = 1 / (1 + r). \tag{2}$$

For r = 1, 2, 5 and 10, six reference operating policies were evaluated: conventional t = 0.50, analytical t*, validation-grid argmin, validation exact-score-boundary argmin, always positive (t = 0), and always negative (t > 1). The original grid contains t = 0.01, 0.02, …, 0.99. Equal-loss ties are resolved by selecting the candidate nearest t*, deterministically. For the exact search, all distinct validation predicted scores and both constant-policy boundaries are considered. The minimum-loss tie count and range are retained. The conventional/analytical/empirical three-policy contrasts were added during revision; all r values are hypothetical and not clinical estimates.

2.5. Threshold-selection variability and conditional held-out uncertainty

To evaluate threshold-selection variability apart from fitting variability, 500 class-stratified bootstrap samples of the audited validation records were generated. Model fits and predicted probabilities were held fixed, and the grid search was repeated within each sample. The median, 2.5th–97.5th percentile interval, the number of distinct selected thresholds and the count of tied loss-minimizing grid candidates were summarized. These intervals characterize a conditional threshold-selection distribution under fixed class composition, not uncertainty arising from retraining the model or changing the clinical case mix.

Held-out loss contrasts used the same 624 images under two policies with 1,000 paired class-stratified bootstrap test resamples, fixing both fitted predictions and preselected cutoffs. We report point differences (validation-grid minus analytical, validation-grid minus fixed, and analytical minus fixed) and 95% percentile intervals [10], without multiplicity adjustment. These are descriptive conditional intervals, not confirmatory repeated-sample statistical tests. No primary cost ratio was selected retrospectively for formal significance testing. The primary descriptive estimand is the test loss contrast between validation-grid and analytical threshold across each of the eight model–cost combinations; all other diagnostics are exploratory.

2.6. Class composition, alternative development roles and reporting scope

A hypothetical class-composition sensitivity analysis held test-set sensitivity and specificity fixed while varying the positive-label fraction π at 0.50, 0.625, 381/516, and 0.80. Reweighted normalized loss was defined as $L(t; \pi, r) = r\,\pi[1-\text{Sensitivity}(t)] + (1-\pi)[1-\text{Specificity}(t)]$. This arithmetic sensitivity calculation does not prove

prevalence shift or isolate any mechanism of threshold deterioration. Benchmark class proportions cannot be interpreted as clinical disease prevalence.

$$L(t; \pi, r) = r\,\pi\{1 - \text{Sensitivity}(t)\} + (1 - \pi)\{1 - \text{Specificity}(t)\}. \quad (3)$$

An additional revision-stage analysis separated roles within development observations. A stratified 12% of the published training partition was provisionally reserved for CNN stopping; exact training-to-stopping pixel matches were purged, leaving 4,143 fitting records and 562 stopping records. The audited 516-image validation set was divided into 258 calibration and 258 threshold-selection records using a fixed stratified split. Temperature was fitted using only the calibration subset, and the threshold search used only the threshold-selection subset. One CNN fit (seed 20261011) evaluated this design. This role separation reduces information reuse but sacrifices development sample size and does not create a newly uninspected test set.

The analyses were conducted in Python using NumPy, pandas, SciPy, scikit-learn, PyTorch, and matplotlib. Exact software versions, source checksums, random seeds, predictions, and raw results are included in the reproducibility package. We consulted TRIPOD + AI [11] and CLAIM 2024 [12] for reporting guidance, while noting the absence of patient identifiers, the lack of clinician-defined cost functions, the absence of an external clinical cohort, and the exploratory, previously examined test partition. No individual patient treatment decisions were evaluated.

3. Results

3.1. Model discrimination and probability calibration

The original test partition contained 390 pneumonia-labeled images (62.5%) and 234 normal images, while the audited validation partition contained 381 pneumonia-labeled images (73.8%) and 135 normal images. Logistic regression and temperature-scaled CNN achieved ROC AUCs of 0.9212 and 0.9251, respectively. **Table 2** shows that temperature scaling did not improve the test Brier score or log loss; in this run, both increased. The CNN temperature was 0.895, based on validation optimization. Calibration slopes near 0.5 and nonzero negative intercepts indicate departures from the ideal calibration line; these coefficients are model- and sample-specific descriptive diagnostics. The calibrated CNN did not demonstrate clinically transportable predicted risks.

Table 2. Test discrimination, probability scores and calibration diagnostics; intercept/slope intervals are conditional class-stratified bootstrap estimates.

Model	AUC	Brier	Log loss	Intercept [95% CI]	Slope [95% CI]
PCA logistic	0.9212	0.1243	0.5967	−1.29 [−1.59, −1.01]	0.51 [0.44, 0.60]
CNN unscaled	0.9251	0.1303	0.5524	−1.33 [−1.67, −1.06]	0.58 [0.49, 0.71]
CNN temperature-scaled	0.9251	0.1325	0.5951	−1.33 [−1.76, −1.10]	0.52 [0.44, 0.64]

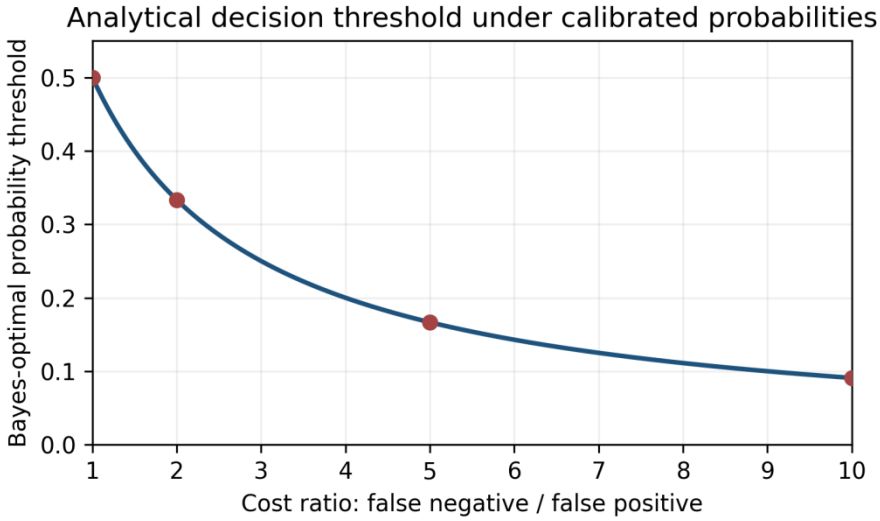


Figure 1. ROC discrimination on the 624-image test partition; two model probabilities evaluated on identical observations.

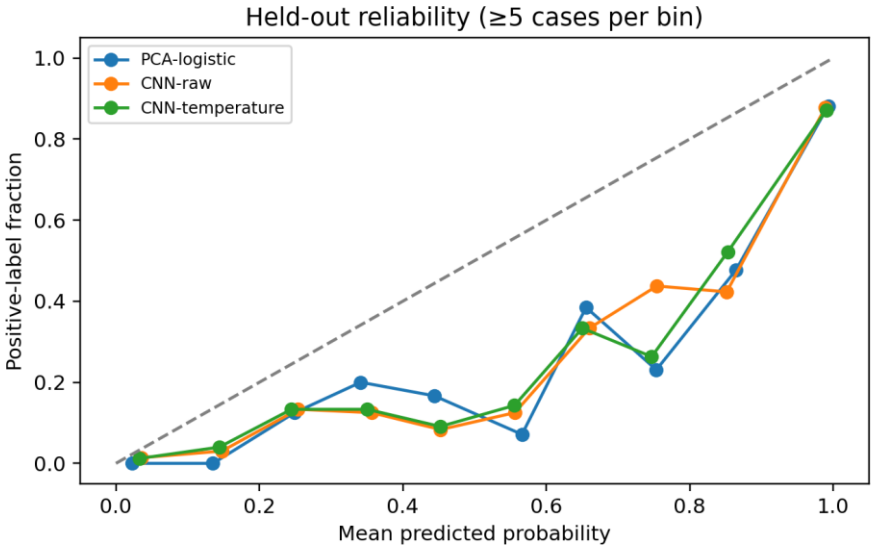


Figure 2. Held-out calibration diagram comparing unscaled and scaled CNN scores and PCA-logistic probabilities; bins with fewer than five observations are excluded.

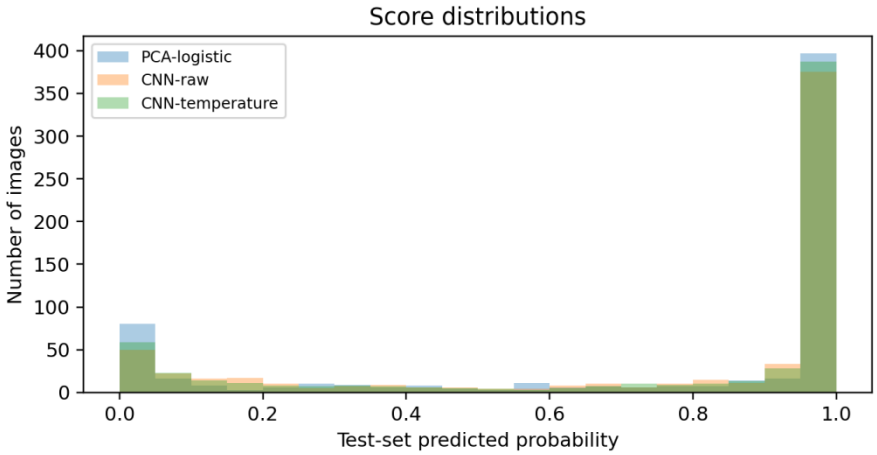


Figure 3. Held-out predicted-score histograms; plotted numbers of images do not represent clinical disease prevalence.

3.2. Three-policy held-out decision costs

The novel revision-stage comparison shows that model ranking and relative policy performance are not interchangeable. At r = 1, PCA-logistic grid tuning reduced normalized test loss from 0.1571 (fixed and analytical) to 0.1378, whereas CNN tuning increased loss from 0.1603 to 0.1747. At r = 2, both models selected thresholds with higher loss than the conventional cutoff, and the direction of the grid versus analytical comparison differed by model. Full three-policy values are presented in **Table 3**, and conditional paired differences in **Table 4**. These comparisons were added after an initial inspection of the test set and cannot support confirmatory policy claims.

Table 3. Three-policy normalized test error cost at four hypothetical FN/FP cost ratios; lower values indicate less specified loss.

Model	r	Analytical t	Grid t	Fixed loss	Analytical loss	Grid loss
Logistic	1	0.500	0.61	0.157	0.157	0.138
Logistic	2	0.333	0.13	0.168	0.188	0.212
Logistic	5	0.167	0.12	0.202	0.208	0.212
Logistic	10	0.091	0.02	0.258	0.226	0.279
Scaled CNN	1	0.500	0.41	0.160	0.160	0.175
Scaled CNN	2	0.333	0.34	0.171	0.196	0.192
Scaled CNN	5	0.167	0.17	0.205	0.231	0.229
Scaled CNN	10	0.091	0.17	0.261	0.269	0.245

3.3. Empirical threshold stability and boundary search

Point-optimal validation thresholds concealed substantial selection instability (**Table 5**). For logistic regression at r = 2, the originally selected threshold of 0.13 was accompanied by a 500-resample 2.5th–97.5th percentile selection range of 0.05–0.63; 24 distinct grid thresholds were selected across the resamples. At r = 1, the scaled CNN had a threshold range of 0.28–0.69. These are resampling intervals for selection conditional on fixed predictions, not uncertainty intervals for an optimal clinical

threshold. The number and span of equal-loss thresholds were also recorded. An exact-score search sometimes achieved lower validation loss than the coarse grid, notably for PCA logistic r = 2, 5 and 10, but this difference in development loss does not establish superior test performance.

Table 4. Paired held-out normalized test loss differences, 1,000 class-stratified bootstrap samples; conditional on fixed fitted models and cutoffs.

Model	r	Grid minus analytical [95% interval]	Grid minus fixed [95% interval]
Logistic	1	−0.019 [−0.030, −0.008]	−0.019 [−0.030, −0.008]
Logistic	2	+0.024 [+0.005, +0.043]	+0.043 [+0.019, +0.067]
Logistic	5	+0.003 [+0.000, +0.008]	+0.010 [−0.037, +0.051]
Logistic	10	+0.053 [+0.037, +0.071]	+0.021 [−0.069, +0.096]
Scaled CNN	1	+0.014 [+0.005, +0.024]	+0.014 [+0.005, +0.024]
Scaled CNN	2	−0.003 [−0.008, +0.000]	+0.021 [+0.006, +0.035]
Scaled CNN	5	−0.002 [−0.005, +0.000]	+0.024 [−0.018, +0.058]
Scaled CNN	10	−0.024 [−0.053, +0.014]	−0.016 [−0.098, +0.046]

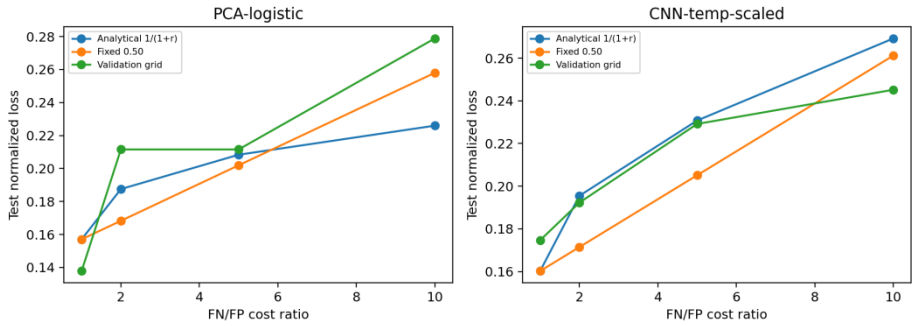


Figure 4. Fixed, analytical and validation-selected threshold policies evaluated on the same held-out test images. The analyses were added during revision.

Table 5. Conditional threshold-selection variability under 500 class-stratified validation resamples.

Model	r	Median cutoff	95% percentile range	Unique	Median ties	Mean tie width
Logistic	1	0.61	[0.13, 0.74]	20	3	0.055
Logistic	2	0.14	[0.05, 0.63]	24	2	0.046
Logistic	5	0.07	[0.02, 0.17]	12	2	0.013
Logistic	10	0.05	[0.02, 0.13]	9	1	0.007
Scaled CNN	1	0.41	[0.28, 0.69]	10	5	0.074
Scaled CNN	2	0.34	[0.17, 0.40]	9	5	0.062
Scaled CNN	5	0.22	[0.04, 0.40]	10	3	0.035
Scaled CNN	10	0.17	[0.04, 0.40]	12	1	0.019

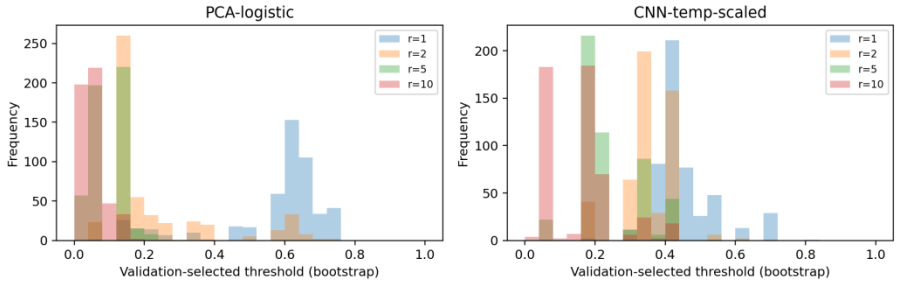


Figure 5. Distribution of selected validation grid cutoffs across 500 class-stratified resamples for each model and hypothetical error-cost ratio.

Table 6. Coarse-grid versus all observed score-boundary validation optimization; exact search includes both constant decisions.

Model	r	Grid t	Exact t	Grid val loss	Exact val loss	Grid tied
Logistic	1	0.61	0.619	0.0465	0.0465	5
Logistic	2	0.13	0.135	0.0640	0.0620	3
Logistic	5	0.12	0.058	0.0814	0.0775	4
Logistic	10	0.02	0.028	0.0891	0.0833	1
Scaled CNN	1	0.41	0.411	0.0329	0.0329	2
Scaled CNN	2	0.34	0.364	0.0446	0.0446	5
Scaled CNN	5	0.17	0.172	0.0698	0.0698	1
Scaled CNN	10	0.17	0.172	0.0988	0.0988	1

3.4. Standardized class-composition sensitivity

With the test-set class-conditional sensitivities and specificities held fixed, changing π alters cost weighting across operating policies. **Figure 6** illustrates this arithmetic effect at r = 5 using hypothetical positive-label fractions of 0.50, 0.625, 381/516 and 0.80. A policy preference can change under different weighting, even without retraining. No causal claim is made that the audited validation–test class difference caused observed cost deterioration, because class-conditional score distributions, reference-label errors, population changes, and selection effects were not separately identified.

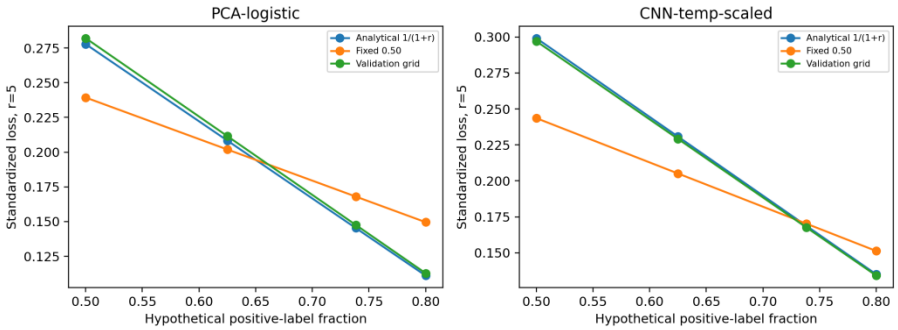


Figure 6. Standardized loss at r = 5 under four illustrative positive-label fractions; class-conditional error rates are held fixed.

3.5. Repeated seeds, deduplicated training and development-role separation

Three CNN seeds produced AUCs of 0.9251, 0.9277 and 0.9236, with differing Brier scores (**Table 7**). A single exact-deduplicated training refit at seed 20261008 reduced CNN AUC from 0.9251 to 0.9206, while deduplicating logistic training changed AUC only slightly from 0.92117 to 0.92123. These results demonstrate some model-fitting and training-data sensitivity, but three CNN seeds do not characterize the full distribution of model-training variability. The auxiliary role-separated CNN produced different test losses across three policies (**Table 8**), illustrating a practical way to reduce shared-validation functions. It is because of a different development procedure assessed retrospectively, its results should not be treated as an unbiased comparison with the original model.

Table 7. CNN repeat-seed and exact-training-duplicate removal sensitivity analyses; each row is a separate fit.

Model	Seed	Training	*n*	Test AUC	Brier
CNN	20261008	Original training	4708	0.9251	0.1325
CNN	20261009	Original training	4708	0.9277	0.1109
CNN	20261010	Original training	4708	0.9236	0.1285
Logistic	20261008	Deduplicated training	4690	0.9212	0.1256
Logistic	20261008	Original training	4708	0.9212	0.1243
CNN	20261008	Deduplicated training	4690	0.9206	0.1469

Table 8. Additional exploratory CNN with distinct stopping, calibration and threshold-selection development subsets; test was already inspected.

r	Analytical t	Selected t	Fixed loss	Analytical loss	Selected loss
1	0.500	0.61	0.168	0.168	0.160
2	0.333	0.37	0.176	0.194	0.192
5	0.167	0.37	0.200	0.234	0.207
10	0.091	0.11	0.240	0.269	0.258

4. Discussion

The revised contribution is an empirical account of threshold-policy transferability rather than a contest to select the best classification architecture. The held-out study shows that a cutoff that minimizes loss on a development partition can yield no advantage, or even higher loss, when transferred to a second benchmark partition. Introducing the analytical threshold as a separate policy demonstrates that the conventional 0.50 rule, theoretically cost-based policy, and empirically tuned cutoff should not be used interchangeably. The analytical decision threshold is optimal under ideal calibration for a target distribution and the chosen loss, not automatically for observed benchmark outputs [7].

Several analyses help interpret the instability without identifying its cause. First, calibration slopes far below 1 and nonzero negative intercepts suggest probability miscalibration. Temperature scaling did not consistently improve the two examined proper scoring rules on test, despite its validation objective [4,5]. Second, threshold

resampling shows large uncertainty in which empirical cutoff is selected. Third, the validation-to-test change in label fraction motivates, but does not explain, the class-composition reweighting exercise. Fourth, the exact-score search illustrates an additional source of discretion: two algorithms seeking the same nominal minimum may select distinct boundaries or yield tied regions. Fifth, deduplication and training-seed sensitivities show that fitting choices also matter. These components are observationally entangled, and their causal contributions remain unresolved.

The sensitivity results should be interpreted with particular caution. Original CNN fitting, temperature tuning, and cutoff selection used the same audited validation set; the role-separated revision alleviates some reuse at the expense of smaller subsamples and a different fitted model. Bootstrap intervals for test loss hold all models and cutoffs fixed, whereas threshold-resampling intervals hold scores fixed but repeat selection. Neither is a full-pipeline, external-site uncertainty interval. Model-comparison tests were not assigned to a confirmatory family, no multiple-testing correction was performed, and no cost ratio was selected as primary based on the realized test estimates. Once prior test performance is known, the expanded revision-stage analyses remain exploratory and require a truly independent future test cohort for prospective confirmation.

The data limitations are substantial. The published benchmark has reduced-resolution 28 × 28 pixel arrays, unknown clinically meaningful decision utilities, labels with provenance inherited from the source dataset, and no original patient identifiers. Although pixel hashing excluded eight exact train–validation collisions and found no exact test overlaps, neither near-duplicate screens nor hash-negative pairs can establish patient-level independence. Low-resolution dHash produced hundreds of close-pattern pairs across nominal partitions, an illustration of how perceptual similarity criteria can generate uninformative flags rather than proof of leakage. No external institution, scanner domain or clinically adjudicated full-resolution outcomes were evaluated. CLAIM 2024 urges reporting the partition disjointness level; we report only what the available data support [12].

Future work should reserve a genuinely independent institution- and patient-disjoint evaluation cohort before policy development, obtain cost ratios from radiologists and patient-centered clinical workflows, separately fit calibration and operating point functions or cross-fit development scores with proper nesting, repeat entire pipelines across diverse seeds, assess near-duplicates using medically defensible image and metadata criteria, and evaluate calibration across imaging sources and observed case mix. Such work would be needed to assess external generalization or clinical benefit. Present results are deliberately confined to the methodological benchmark.

5. Conclusion

Using authenticated PneumoniaMNIST data, the revised experiment found that validation-selected decision thresholds do not consistently retain their apparent cost advantage over fixed or analytically derived cutoffs on a held-out benchmark partition. Additional revision-stage analyses showed substantial selection variability, imperfect probability calibration, occasional grid dependence, and sensitivity to development

choices. Similar AUCs did not imply similar decision loss under stipulated hypothetical error costs. Because the same test set was inspected during earlier stages, these results are exploratory, not confirmatory; patient-level partition independence and clinical utility are not established. Explicit policy definitions, partition auditing, conditional uncertainty estimands, and independent future validation should accompany medical-imaging threshold optimization.

Author contributions: Arvy Noleal Osma (A.N.O.), the first and corresponding author, contributed to the conceptualization and methodological design of the study, critical revision of the manuscript, and refinement of its academic content and presentation. Mark Andre Balao (M.A.B.), the second author, contributed to the critical review of the computational analyses, examination of the analytical procedures, and review of the manuscript's scientific accuracy. M.A.B. has approved the final manuscript for submission. Both authors are accountable for their respective contributions to the research.

Funding: The authors received no funding for this research.

Acknowledgments: The authors acknowledge the creators of the PneumoniaMNIST dataset and the original pediatric chest radiograph collection for making the imaging data available for scientific research. The original dataset sources are appropriately credited in References [1–3,9].

Institutional review board statement: This study involved a retrospective computational analysis of a publicly available, openly licensed medical imaging dataset. No new human participants were recruited, no clinical interventions were performed, and no identifiable patient records were accessed. No institutional review board approval or exemption reference number is available or claimed for this secondary-data analysis.

Informed consent: No new participants were enrolled or contacted. Consent for the original image acquisition is governed by the source data custodians and is not independently verifiable from the 28×28 benchmark archive.

Data and code availability: The published dataset is available at doi:10.5281/zenodo.10519652. This revision includes a reproducible local analysis archive with the exact code, dataset checksum, model configurations, de-identified image-level test scores, figures, and CSV summaries. A public permanent code-archive URL and model checkpoints have not yet been deposited; these must be addressed or explicitly disclosed prior to journal submission.

Conflict of interest: The authors declare no conflict of interest.

AI assistance disclosure: Generative AI was used for text drafting, code development, editing and document preparation. The computational analyses reported here were executed on the supplied dataset. Both human authors independently reproduced and verified all data processing, numerical results, interpretations and citations; they retain full responsibility for the manuscript.

References

1. Yang J, Shi R, Wei D, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 2023; 10: 41. doi: 10.1038/s41597-022-01721-8
2. Yang J, Shi R, Ni B. MedMNIST Classification Decathlon: A lightweight AutoML benchmark for medical image analysis. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI). 2021: 191–195. doi: 10.1109/ISBI48211.2021.9434062
3. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018; 172(5): 1122–1131.e9. doi: 10.1016/j.cell.2018.02.010
4. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Medicine. 2019; 17: 230. doi: 10.1186/s12916-019-1466-7
5. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017; 70: 1321–1330. https://proceedings.mlr.press/v70/guo17a.html
6. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010; 21(1): 128–138. doi: 10.1097/EDE.0b013e3181c30fb2
7. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making. 2006; 26(6): 565–574. doi: 10.1177/0272989X06295361
8. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE. 2015; 10(3): e0118432. doi: 10.1371/journal.pone.0118432
9. Yang J, et al. MedMNIST+: 18× standardized datasets for 2D and 3D biomedical image classification. Zenodo, version 3.0. 2024. doi: 10.5281/zenodo.10519652
10. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York, NY: Chapman & Hall; 1993. doi: 10.1007/978-1-4899-4541-9
11. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024; 385: e078378. doi: 10.1136/bmj-2023-078378
12. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology: Artificial Intelligence. 2024; 6(4): e240300. doi: 10.1148/ryai.240300