UNDER REVIEW

RESCIENCE C

Replication / Machine Learning

[Re] MedMNIST v2: Replicating the ResNet baselines and auditing what macro-AUC hides on DermaMNIST

Christopher Huang[1], **Ryan Ahn**[2], **Wenhao Lu**[3], **and Spursh Deshpande**[4]

[1]Heritage High School – [2]Yongsan International School of Seoul, Seoul, South Korea – [3]Millburn High School, Millburn, NJ, United States – [4]Crescenta Valley High School, La Crescenta, CA, United States

Edited by
(Editor)

Received
–

Published
–

DOI
–

1 Introduction

MedMNIST v2[1] is a widely used benchmark of standardized, lightweight biomedical image classification datasets. Its Table 3 reports ResNet-18 and ResNet-50[2] baselines at input sizes 28 and 224 across twelve 2D datasets, evaluated by macro one-vs-rest ROC-AUC (AUC) and accuracy (ACC). Because these baselines anchor a large body of follow-up work, verifying that they can be obtained independently, from the paper's description alone and without the authors' code, is a natural target for a ReScience C replication.

This article makes two contributions. First, we *replicate* the 28-pixel ResNet baselines of MedMNIST v2 with a from-scratch reimplementation in PyTorch[3]: 9 of 13 configurations land inside our pre-registered tolerance of $|\Delta\text{AUC}| \leq 0.02$ and $|\Delta\text{ACC}| \leq 0.03$, and all four misses are analyzed rather than discarded. A separate *reproduction* arm runs the authors' own scripts unmodified on DermaMNIST at 224 pixels and also lands inside tolerance. Second, we extend the study with an equity audit of DermaMNIST[4]: the benchmark's headline metric, macro-AUC, is essentially uncorrelated with class frequency (Pearson $r = -0.28$), while recall correlates strongly with it ($r = 0.81$). The practical consequence is stark: the baseline recalls only 8.7% of dermatofibroma test cases while that class's AUC reads a reassuring 0.917. We quantify this gap, its calibration counterpart, and the equity/accuracy tradeoffs of two standard mitigations, and we show that a $4\times$ parameter compression erases the rarest class entirely while leaving aggregate metrics almost untouched.

All code, run manifests, per-run records, prediction matrices, and figure-generation scripts are available in the accompanying repository; every number in this article regenerates from committed artifacts.

2 Scope of the replication

We test the following claims of the original paper:

1. The reported ResNet-18@28 test AUC/ACC values of Table 3 are attainable across the twelve 2D datasets using only the training protocol described in the paper (claim tested by our independent reimplementation).

Correspondence should be addressed to Wenhao Lu (wenhao@nxthorizon.org)
The authors have declared that no competing interests exist.
Code is available at https://github.com/AIscend-Research/medmnist-imaging-repro..
Data is available at https://medmnist.com – DOI 10.5281/zenodo.6496656.

2. The reported ResNet-50@28 value on DermaMNIST is likewise attainable, and deeper backbones do not help at 28×28 (ResNet-50 tracks or slightly trails ResNet-18).
3. The authors' released training script reproduces the paper's 224-pixel DermaMNIST rows (claim tested by the reproduction arm).

The extension then asks a question the original paper does not: *what do these aggregate metrics conceal on an imbalanced dataset?*

3 Methodology

3.1 Two arms with strictly separated provenance

We maintain two arms whose results are never pooled:

- **Replication arm** (the deliverable): a clean-room reimplementation. Nothing in it imports, copies, or adapts any code from the authors' experiments repository; every model, transform, training loop, and metric is our own. The official Python package supplies only (a) the standardized datasets via its loaders and (b) its evaluator class, used exclusively as an oracle to verify that our metric code agrees with the official one (agreement asserted to within 10^{-3} at the end of every run).
- **Reproduction arm**: the authors' released training script, run unmodified on DermaMNIST with ResNet-18 and ResNet-50 using its resize option (which upsamples the 28-pixel data to 224), one seed each. Its results are tabulated by a manifest-driven script and compared against the paper's 224-pixel column.

3.2 Training specification (replication arm)

The paper's prose leaves the 28-pixel backbone underspecified; a standard torchvision ResNet-18 (7×7 stride-2 stem plus max-pool) collapses a 28×28 input to 1×1 before the last stage. The specification that makes the numbers line up is a CIFAR-style ResNet: 3×3 stride-1 stem, no max-pool, stage strides $[1, 2, 2, 2]$, widths $[64, 128, 256, 512]$ (ResNet-18: BasicBlock $[2, 2, 2, 2]$; ResNet-50: Bottleneck $[3, 4, 6, 3]$). Data are the standardized 28-pixel archives loaded as RGB, normalized to $[-1, 1]$, no augmentation. Training uses Adam (lr 10^{-3}), batch size 128, 100 epochs, `MultiStepLR` with $\gamma = 0.1$ at epochs 50 and 75, cross-entropy loss (BCE-with-logits for multi-label ChestMNIST). We select the checkpoint with the best validation macro-AUC and report its test AUC/ACC, matching the paper's protocol. AUC is macro one-vs-rest ROC-AUC on softmax probabilities; ACC is argmax accuracy.

3.3 Run matrix, tolerances, and compute

A compute-minimal slice of Table 3 was fixed before running: all twelve 2D datasets at ResNet-18@28, plus ResNet-50@28 on DermaMNIST, for 13 configurations in total. The seven small/medium datasets and both DermaMNIST configurations use 3 seeds $\{0, 1, 2\}$ (mean ± std reported); the four large datasets (Tissue, OCT, Path, Chest) use a single seed for compute reasons, disclosed here and stable at that data scale. This yields a 31-run replication matrix (40 runs total including extension variants). Tolerances were set in advance: a configuration passes if $|\Delta\text{AUC}| \leq 0.02$ *and* $|\Delta\text{ACC}| \leq 0.03$ (Δ = ours − paper); anything outside is flagged `OUT_OF_TOL`, not silently accepted. Each run writes a structured record with the full configuration, seed, best epoch, determinism and mixed-precision flags, and pinned library versions.

Dataset	Model	Seeds	Our AUC	Paper	ΔAUC	Our ACC	Paper	ΔACC
BloodMNIST	R18	3	0.997 ± 0.000	0.998	-0.001	0.954 ± 0.001	0.958	-0.004
BreastMNIST†	R18	3	0.874 ± 0.012	0.901	-0.027	0.806 ± 0.008	0.863	-0.057
ChestMNIST	R18	1	0.767	0.768	-0.001	0.948	0.947	$+0.001$
DermaMNIST	R18	3	0.914 ± 0.002	0.917	-0.003	0.729 ± 0.006	0.735	-0.006
DermaMNIST	R50	3	0.906 ± 0.002	0.913	-0.007	0.724 ± 0.002	0.735	-0.011
OCTMNIST†	R18	1	0.936	0.943	-0.007	0.695	0.743	-0.048
OrganAMNIST	R18	3	0.997 ± 0.000	0.997	-0.000	0.932 ± 0.005	0.935	-0.003
OrganCMNIST	R18	3	0.993 ± 0.000	0.992	$+0.001$	0.913 ± 0.005	0.900	$+0.013$
OrganSMNIST	R18	3	0.970 ± 0.002	0.972	-0.002	0.754 ± 0.010	0.782	-0.028
PathMNIST	R18	1	0.987	0.983	$+0.004$	0.915	0.907	$+0.008$
PneumoniaMNIST†	R18	3	0.942 ± 0.014	0.944	-0.002	0.812 ± 0.054	0.854	-0.042
RetinaMNIST†	R18	3	0.738 ± 0.001	0.717	$+0.021$	0.511 ± 0.025	0.524	-0.013
TissueMNIST	R18	1	0.928	0.930	-0.002	0.673	0.676	-0.003

Table 1. Replication arm at 28×28 vs. MedMNIST v2 Table 3. Δ = ours − paper; tolerance $|\Delta\text{AUC}| \leq 0.02$, $|\Delta\text{ACC}| \leq 0.03$. †out of tolerance (discussed in the text).

4 Results: replication

Table 1 and Figures 1 and 2 summarize the replication arm. Nine of thirteen configurations fall inside tolerance. DermaMNIST, the focus of the extension, replicates to 0.003 AUC, and PathMNIST slightly *exceeds* the paper on both metrics. ResNet-50@28 on DermaMNIST tracks just below ResNet-18, consistent with the paper's finding that depth does not pay at this resolution.

The four `OUT_OF_TOL` flags follow a pattern. OCTMNIST and PneumoniaMNIST are accuracy-only misses with AUC essentially on the paper ($|\Delta\text{AUC}| \leq 0.007$): our threshold-free ranking quality matches, but the argmax operating point differs, which is exactly the failure mode expected when checkpoint selection is by validation AUC and the datasets are small or label-noisy. RetinaMNIST is flagged for the opposite reason: its accuracy is within tolerance and its AUC is 0.021 *above* the paper (a miss under a symmetric tolerance, but not a deficiency). BreastMNIST is the only configuration genuinely below the paper on both metrics; with $n = 546$ training images it shows the highest seed sensitivity in the matrix (Figure 3), and single-checkpoint results on it should be treated as draws from a wide distribution. We deliberately report these misses rather than re-tuning until they pass: the point of a fixed tolerance is to make disagreement informative.

4.1 Reproduction arm: the authors' own code

Independently of the replication, two of us ran the authors' released script unmodified on DermaMNIST with its resize option (28-pixel data upsampled to 224), one seed each. Both land inside tolerance against the paper's 224-pixel column: ResNet-18 reaches AUC 0.9212 / ACC 0.7401 (paper: 0.920 / 0.754) and ResNet-50 reaches AUC 0.9119 / ACC 0.7377 (paper: 0.912 / 0.731). One provenance correction is recorded in the repository: an early summary compared the ResNet-18 run against the paper's 28-pixel row; since the resize option implies 224, the tabulation uses the 224 reference and preserves the original summary untouched as a record of what was first computed. The full 2005×7 test softmax matrix of the ResNet-50 run is retained, so prediction-level analyses can be recomputed without a GPU.

Taken together: the paper's numbers reproduce with the authors' code and replicate from their description alone. MedMNIST v2 passes the replication test it implicitly sets.

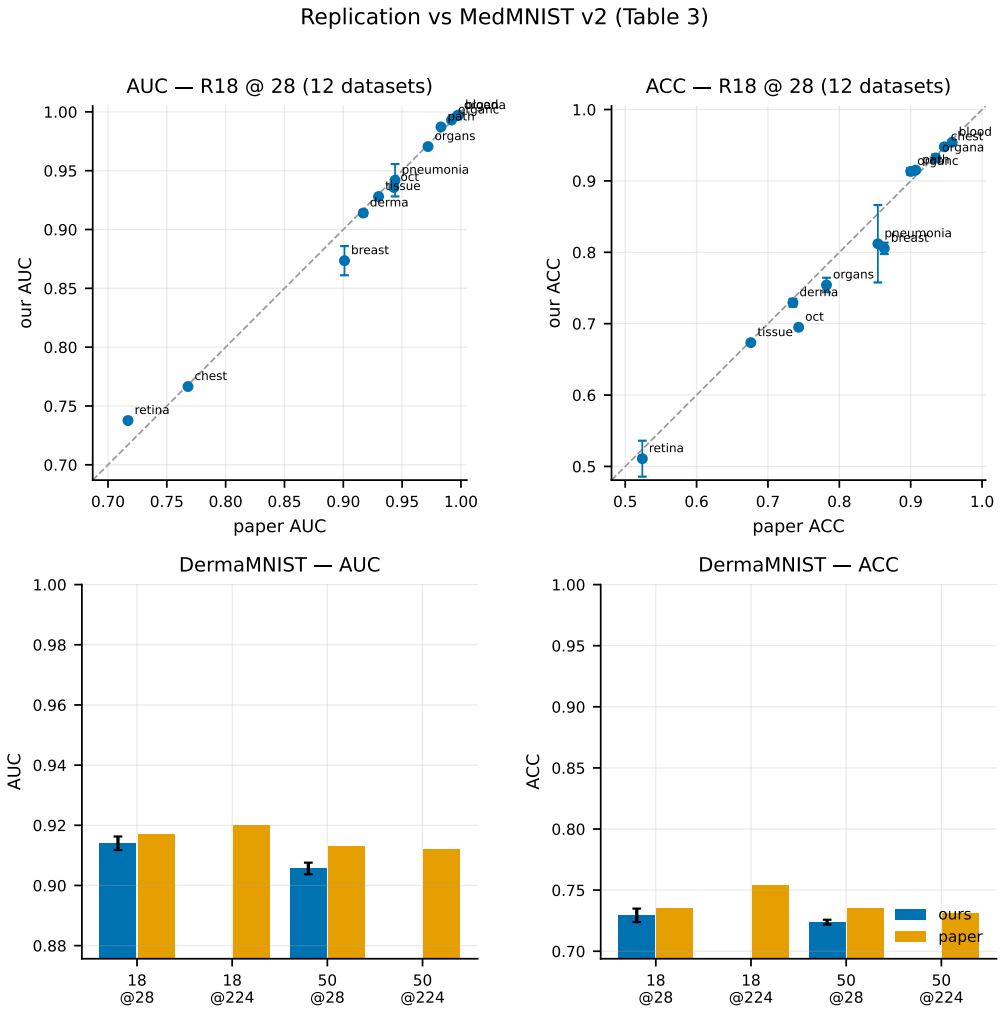

Figure 1. Our results vs. MedMNIST v2 Table 3 across the twelve 2D datasets at ResNet-18@28 (scatter; the diagonal is perfect agreement), plus the DermaMNIST bars for both backbones.

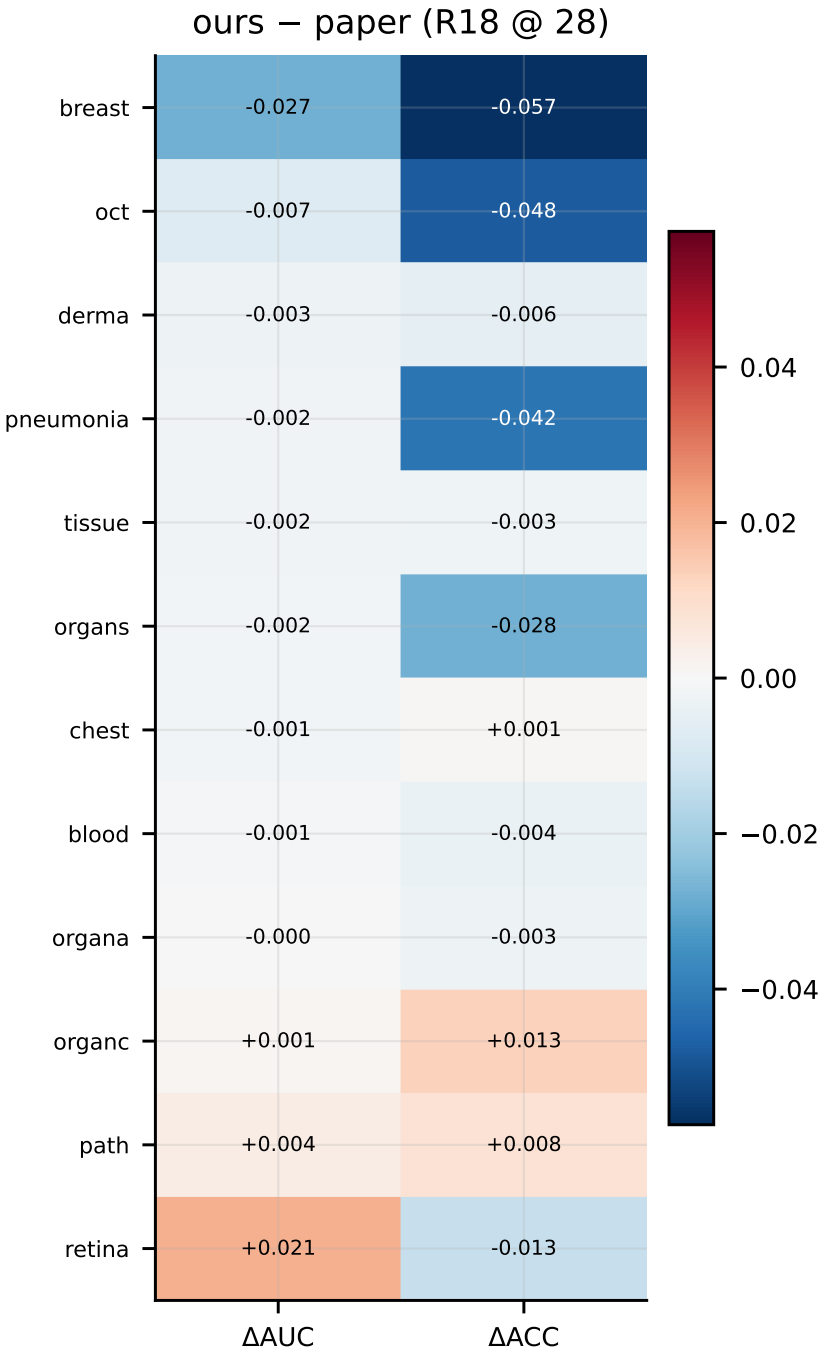

Figure 2. Signed deltas (ours − paper) per dataset and metric. Deviations are small, predominantly negative in ACC, and concentrated in small or noisy datasets.

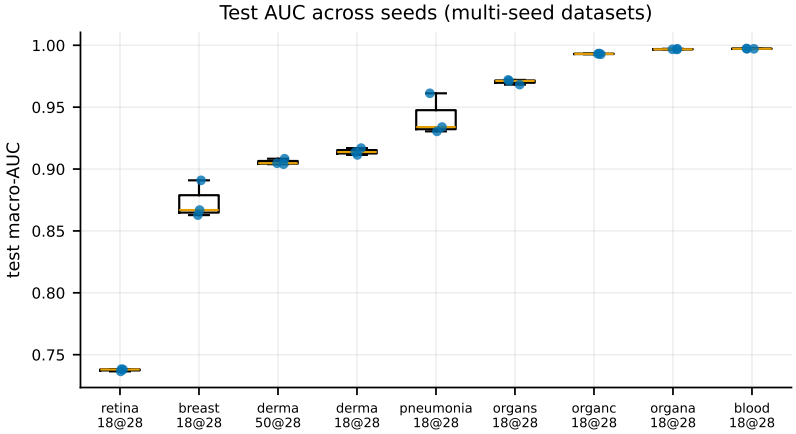

Figure 3. Test-AUC spread across seeds $\{0, 1, 2\}$. BreastMNIST and PneumoniaMNIST, the two smallest datasets, dominate the variance, explaining their tolerance misses.

Class	Support	AUC	Recall	F1
Actinic keratoses / IEC	66	0.936 ± 0.005	0.323 ± 0.125	0.329 ± 0.065
Basal cell carcinoma	103	0.938 ± 0.006	0.463 ± 0.061	0.456 ± 0.013
Benign keratosis-like	220	0.873 ± 0.016	0.341 ± 0.058	0.419 ± 0.038
Dermatofibroma	23	0.917 ± 0.015	0.087 ± 0.071	0.149 ± 0.117
Melanoma	223	0.856 ± 0.007	0.459 ± 0.011	0.429 ± 0.014
Melanocytic nevi	1341	0.902 ± 0.002	0.893 ± 0.016	0.868 ± 0.001
Vascular lesions	29	0.976 ± 0.006	0.563 ± 0.183	0.556 ± 0.116

Table 2. Per-class test metrics of the baseline (mean ± std over 3 seeds). Dermatofibroma pairs the study's headline dissociation: AUC 0.917, recall 0.087.

5 Extension: what macro-AUC hides on DermaMNIST

DermaMNIST is a 28×28 rendering of HAM10000[4,5]: 7,007 training images over seven lesion classes, of which melanocytic nevi alone account for ∼67% while dermatofibroma has 73 training and 23 test images (Figure 4). All extension analyses use the replication-arm ResNet-18@28 baseline (3 seeds).

5.1 Per-class performance and the frequency versus metric dissociation

Table 2 and Figure 5 contain the core result. Per-class recall correlates strongly with training frequency ($r = 0.805$), the canonical imbalance signature[6]. Per-class ROC-AUC is essentially uncorrelated with frequency ($r = -0.282$): being threshold-free and prevalence-insensitive, it certifies that the model *ranks* rare-class positives well while the classifier almost never *predicts* them. Dermatofibroma is the extreme case (AUC 0.917, recall 0.087): two correct out of 23 test cases at argmax. A benchmark scored by macro-AUC and overall accuracy reports this model as healthy.

The errors are also directional, not random. In the misclassified-case gallery (Figure 6), every rare-class error collapses into the 67%-majority nevi class or into melanoma; the headline case is a dermatofibroma predicted as melanocytic nevi at softmax confidence 1.00.

Figure 4. Eight training samples per class at native 28×28. Little visual signal separates the rare classes at this resolution, which is the raw material of the equity problem.

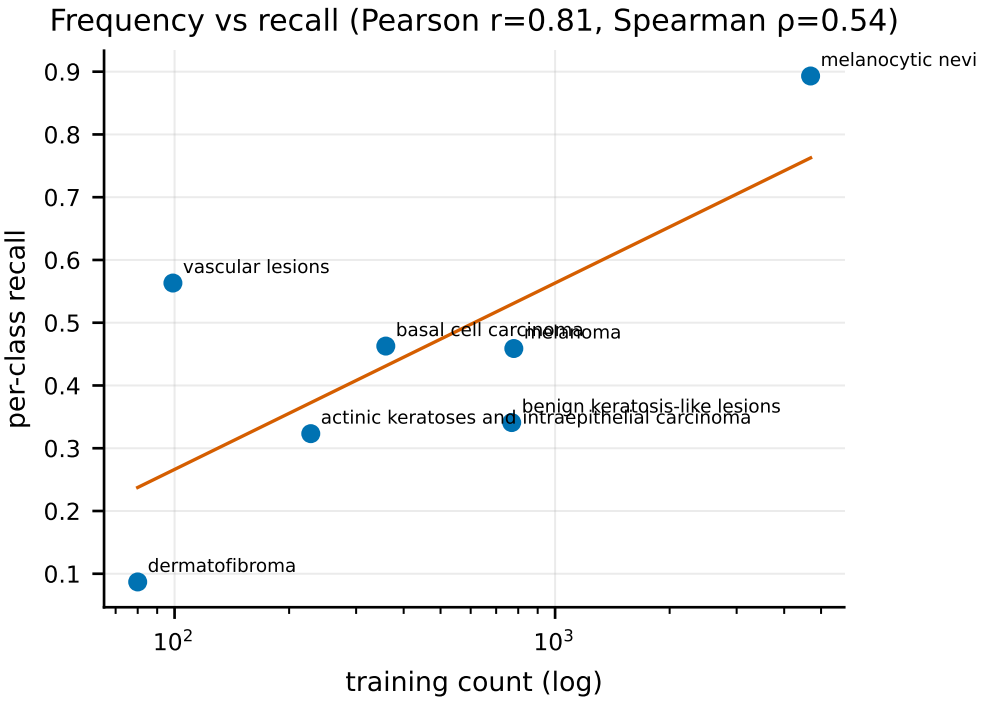

Figure 5. Training count vs. per-class recall (left) and per-class AUC (right). Recall tracks frequency (Pearson $r = 0.805$, Spearman $\rho = 0.536$); AUC does not ($r = -0.282$). The benchmark's headline metric is blind to the equity failure.

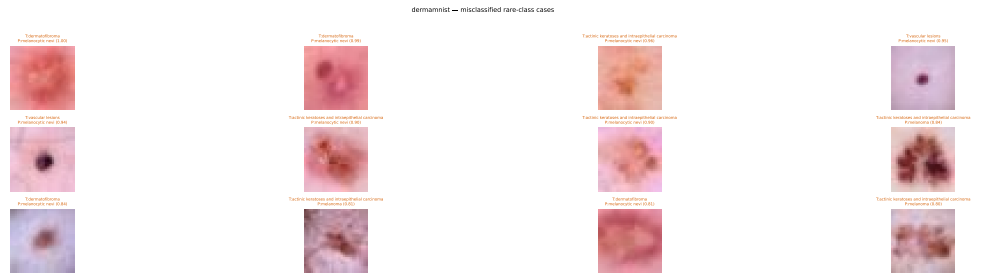

Figure 6. Rare-class test images the baseline misclassifies, with true label, prediction, and confidence. All shown errors collapse into the majority (nevi) class or melanoma.

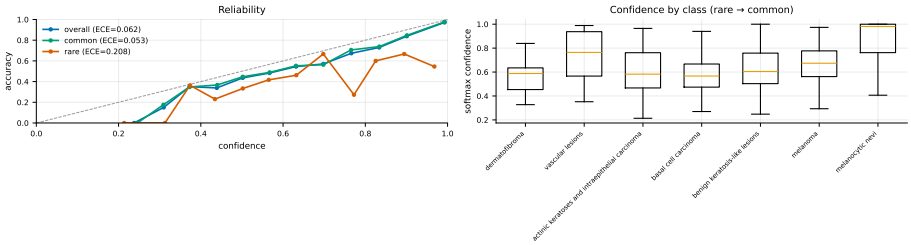

Figure 7. Reliability diagram and per-class confidence. ECE: 0.062 overall, 0.053 common classes, 0.208 rare classes.

Variant	AUC	ACC	Macro-F1	Min-class F1	Min-class recall
Baseline	0.914	0.730	0.465	0.160	0.087
Weighted sampler	0.897	0.735	0.485	0.214	0.130
Weighted loss	0.894	0.633	0.449	0.139	0.435

Table 3. Bias mitigation on DermaMNIST (3 seeds each; worst class is dermatofibroma throughout).

5.2 Calibration inherits the imbalance

Expected calibration error[7] is 0.062 overall, but this too is an average dominated by the majority. Splitting by class rarity, ECE is 0.053 on common classes and 0.208 on rare ones (Figure 7): the model is nearly four times worse-calibrated exactly where its errors are most consequential, and its confidence on rare-class inputs cannot be taken at face value.

5.3 Mitigation: an explicit equity/accuracy tradeoff

We test two standard mitigations[6], each over 3 seeds (Table 3, Figure 8). A `WeightedRandomSampler` is the Pareto-friendly option: min-class F1 rises $0.160 \rightarrow 0.214$ and accuracy *improves* slightly ($0.730 \rightarrow 0.735$), at a cost of 1.7 AUC points. Inverse-frequency loss weighting buys the largest recall gain on the worst class ($0.087 \rightarrow 0.435$, a $5\times$ improvement) but pays with a nine-point accuracy drop ($0.730 \rightarrow 0.633$) as the decision boundary moves aggressively toward the rare classes. Neither dominates: the choice is a deployment decision about the relative cost of missing a rare lesion versus over-calling it, and it should be made on per-class numbers, not on the aggregate metrics that concealed the problem.

5.4 Compression erases the rarest class

A ResNet-18 at $0.5\times$ width ($[32, 64, 128, 256]$) cuts parameters $11.17\text{M} \rightarrow 2.80\text{M}$ and weights $42.7 \rightarrow 10.7$ MB while aggregate performance barely moves (AUC $0.914 \rightarrow 0.905$, ACC $0.730 \rightarrow 0.735$); per-image latency is unchanged ($2.02 \rightarrow 2.04$ ms; inference at 28×28 is memory-bound, so the win is model size rather than speed). The aggregate view calls this compression free. The per-class view does not: dermatofibroma F1 goes $0.16 \rightarrow 0.00$ (the compressed model never predicts the rarest class at all), and the F1 drop correlates with class rarity (Figure 9). Capacity reduction is paid disproportionately by the classes the aggregate metrics weight least.

5.5 Robustness

Under common corruptions the baseline degrades monotonically with severity: macro-AUC falls from 0.914 to 0.836 under Gaussian noise, 0.874 under JPEG compression, and

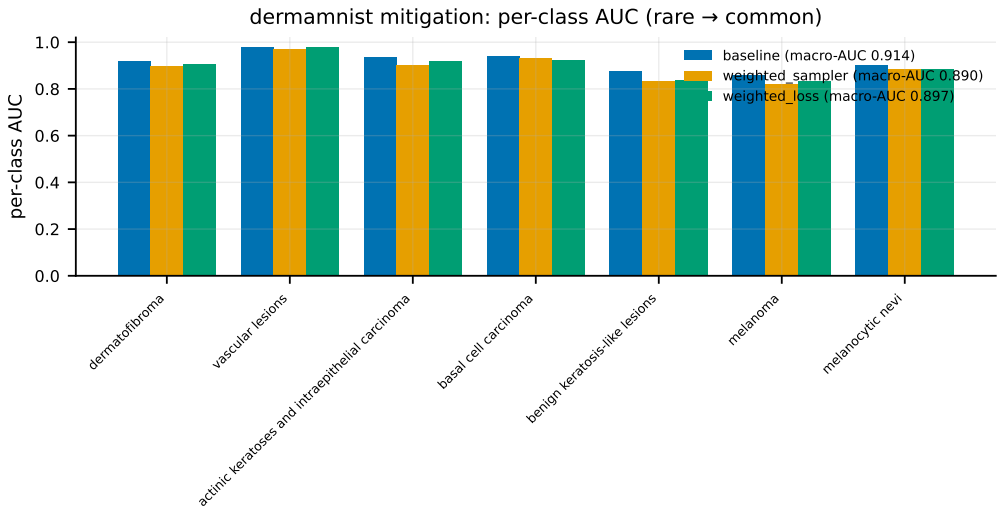

Figure 8. Baseline vs. weighted sampler vs. weighted loss: aggregate and per-class effects of the two mitigations.

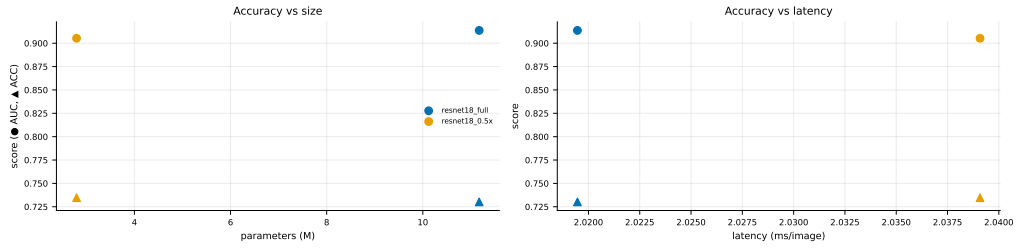

Figure 9. $0.5\times$-width ResNet-18: parameters, size, and latency vs. accuracy, and per-class degradation. Aggregate cost ≈ 0; rarest-class F1 $\to 0$.

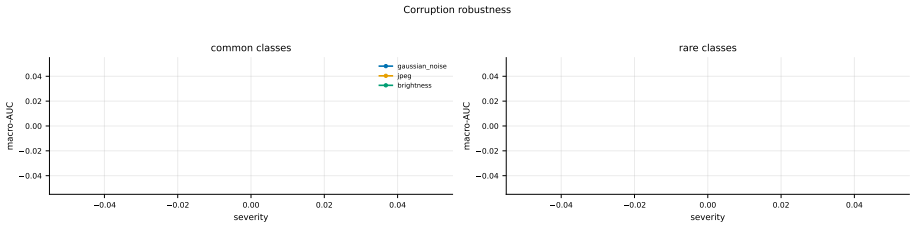

Figure 10. Macro-AUC under Gaussian noise, JPEG compression, and brightness shift at increasing severity.

0.789 under brightness shift at the highest severity (Figure 10). Brightness, a plausible clinical acquisition artifact, is the most damaging, which is unsurprising for dermatoscopy, where color and luminance carry diagnostic signal.

6 Discussion and conclusions

The replication succeeds. MedMNIST v2's 28-pixel ResNet baselines are attainable from the paper's description alone, to within 0.02 AUC in 12 of 13 configurations and within joint tolerance in 9 of 13, using an independent reimplementation whose metrics are verified against the official evaluator. The residual misses are explainable (accuracy-only operating-point differences on small/noisy datasets; high seed variance on BreastMNIST; an AUC *surplus* on RetinaMNIST), and the authors' own script reproduces the 224-pixel DermaMNIST rows within tolerance. One reproducibility caveat is worth recording for future users: the 28-pixel backbone is underspecified in the paper (a stock torchvision ResNet stem does not work at this resolution), and the CIFAR-style stem given in Section 3.2 is the specification that makes the numbers line up.

But the metrics the benchmark reports are the wrong ones for its most imbalanced dataset. On DermaMNIST, macro-AUC and overall accuracy are near-blind to a model that misses 91% of dermatofibromas, is 4× worse-calibrated on rare classes, fails them directionally into the majority class at full confidence, and loses the rarest class *entirely* under a compression that the same aggregate metrics score as free. None of this contradicts the original paper, whose numbers our replication confirms, but it sharpens what those numbers mean. We suggest that leaderboards built on MedMNIST-style benchmarks report per-class recall/F1 and rarity-stratified calibration alongside macro-AUC, and that any efficiency claim be accompanied by per-class degradation. The cost of doing so is one table; the cost of not doing so is mistaking a ranking score for a diagnosis rate.

Acknowledgements

We thank the MedMNIST authors for releasing standardized data, code, and an evaluator that made both arms of this study possible, and the ReScience C community for the venue and template[8].

References

1. J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification." In: **Scientific Data** 10.41 (2023). DOI: 10.1038/s41597-022-01721-8.
2. K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

3. A. Paszke et al. “PyTorch: An imperative style, high-performance deep learning library.” In: **Advances in Neural Information Processing Systems 32 (NeurIPS)**. 2019, pp. 8024–8035.
4. P. Tschandl, C. Rosendahl, and H. Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.” In: **Scientific Data** 5 (2018), p. 180161. DOI: 10.1038/sdata.2018.161.
5. N. Codella et al. “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC).” In: **arXiv preprint arXiv:1902.03368** (2019).
6. M. Buda, A. Maki, and M. A. Mazurowski. “A systematic study of the class imbalance problem in convolutional neural networks.” In: **Neural Networks** 106 (2018), pp. 249–259. DOI: 10.1016/j.neunet.2018.07.011.
7. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. “On calibration of modern neural networks.” In: **Proceedings of the 34th International Conference on Machine Learning (ICML)**. 2017, pp. 1321–1330.
8. N. P. Rougier, K. Hinsen, et al. “Sustainable computational science: the ReScience initiative.” In: **PeerJ Computer Science** 3 (2017), e142. DOI: 10.7717/peerj-cs.142.