Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing | https://www.frpubs.org
Vol. 4, No. 1, 2025

Compact CNN and CNN-Transformer Models for PneumoniaMNIST+ Chest Radiograph Classification

Elric Donnelly
University of Wisconsin-La Crosse, La Crosse, USA
edonnelly@uwlax.edu

Original publication: 1 January 2025 | Revised: 14 November 2026

Abstract

This revised study compares compact convolutional and CNN-Transformer classifiers on public pediatric chest radiographs. Experiments were rerun in November 2026 using the 64 by 64 PneumoniaMNIST+ release, with pneumonia as the positive class. Exact-pixel duplicate removal, performed before model fitting, retains 4,690 training, 516 validation, and 621 test images. Both models use a three-stage convolutional stem and are trained from random initialization for 30 epochs using the same optimizer settings. Three fixed seeds are evaluated per architecture. Checkpoints and classification thresholds are selected using validation data; all six models are frozen before test evaluation. Mean test ROC AUC is 0.9704 for the CNN and 0.9537 for the CNN-Transformer. Mean test accuracy is 92.22% and 88.84%, respectively, at validation-selected thresholds. Per-seed metrics, confusion matrices, learning curves, and image-level bootstrap intervals are reported together with reproducibility files. The CNN has the higher mean test AUC in this comparison; adding attention does not provide a consistent advantage. The findings describe retrospective benchmark classification and require separate patient-level and external validation before clinical interpretation.

Keywords: PneumoniaMNIST+; chest radiographs; convolutional neural network; Transformer; duplicate removal; reproducibility

1. Introduction

A reproducible medical-image classification experiment requires an identified dataset, explicit split membership, a fully specified training procedure, and predictions that support the reported metrics. MedMNIST provides standardized biomedical image classification tasks [1]. Its PneumoniaMNIST subset derives from pediatric chest radiographs associated with the study by Kermany et al. [2]. The available labels support a normal-versus-pneumonia classification task; they do not provide disease-onset times, longitudinal outcomes, or a test of early diagnosis.

The present comparison asks whether a small Transformer context head improves held-out discrimination over a similarly sized convolutional context head when preprocessing, training budget, and checkpoint selection are held fixed. The contributions are a documented duplicate audit, a controlled six-run comparison, and traceable test predictions. The study evaluates one imaging modality and one public classification task.

2. Dataset and image integrity

The official MedMNIST+ release at Zenodo record 10519652 supplies 64 by 64 grayscale PneumoniaMNIST images under CC BY 4.0 [3]. Labels are 0 for normal and 1 for pneumonia. The distributed training, validation, and test arrays contain 4,708, 524, and 624 images. The directly downloaded NPZ matches the published MD5 digest.

Before fitting, every image was hashed over its 4,096 uint8 pixel values using SHA256. Splits were processed in the order training, validation, test; only the first occurrence of each hash was retained. No image was transferred between splits. This removes 18 repeated training images, 8 validation images matching training images, and 3 repeated test images. All duplicate labels are consistent. Retained indices and removal matches are recorded in the supplementary materials.

Split	Official total	Retained normal	Retained pneumonia	Retained total
Training	4,708	1,213	3,477	4,690
Validation	524	135	381	516
Test	624	234	387	621
All splits	5,856	1,582	4,245	5,827

Table 1. Official and retained image counts. Counts represent images, not independently verified patients.

Dataset file: pneumoniamnist_64.npz
MD5: 8f4eceb4ccffa70c672198ea285246c6
SHA256: 97c5a52c65a74f6c5b77a3d8b3bb8aba58276d5a1aaa3b0340d079e54f8e9756

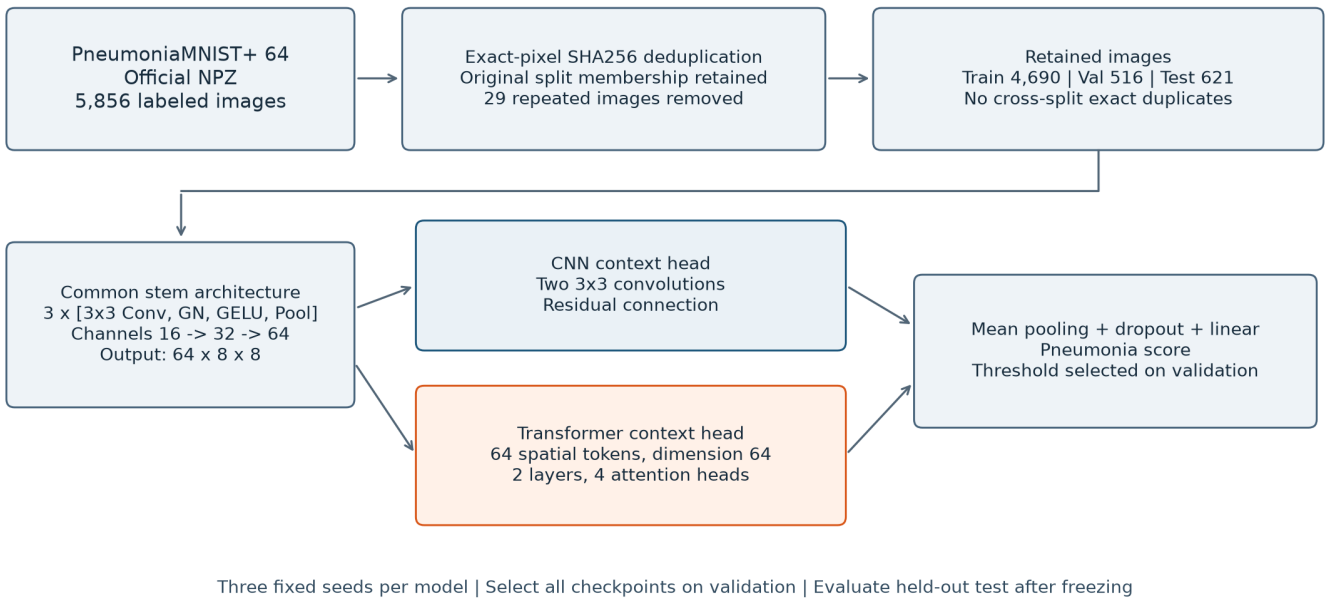

Figure 1. Implemented data preparation and model comparison. Both heads receive the same convolutional feature dimensions; validation data determine checkpoints and thresholds.

Retained pixel hashes do not overlap across splits. Because patient identifiers are unavailable, exact-pixel deduplication cannot establish patient independence or exclude near-duplicate acquisitions.

3. Models and training

3.1 Common stem architecture and context heads

Each image is converted to float32 and scaled by x/127.5 - 1. Both models use the same stem architecture, trained independently. Three convolutional stages use 16, 32, and 64 channels. Each stage contains a bias-free 3 by 3 convolution with padding 1, four-group Group Normalization [4], GELU, and 2 by 2 max pooling. The resulting feature tensor has shape 64 by 8 by 8. No augmentation, pretrained weights, segmentation masks, or additional modalities are used.

Component	CNN	CNN-Transformer
Stem	Three Conv-GN-GELU-Pool stages	Identical stem dimensions
Context	Two bias-free 3x3, 64-channel convolutions; GN after each; GELU after the first; residual addition then GELU	64 tokens of dimension 64; learned positional embeddings; two pre-norm Transformer layers; 4 heads; feed-forward width 128
Pooling and output	Spatial mean; dropout 0.1; linear 64-to-1	Final LayerNorm; token mean; dropout 0.1; linear 64-to-1
Trainable parameters	97,457	94,641

Table 2. Implemented architecture. GN denotes Group Normalization. The CNN has approximately 3.0% more parameters than the CNN-Transformer.

The hybrid head uses self-attention [5] over the 64 spatial tokens, with four attention heads and two encoder layers. Each layer uses pre-normalization, GELU, a feed-forward width of 128, and dropout 0.1. A final Layer Normalization [6] precedes token averaging. The CNN head supplies a spatial convolutional control with a comparable parameter count. Both heads produce a single logit, transformed to a pneumonia score by the sigmoid function.

3.2 Fixed optimization procedure

Setting	Value used for both architectures
Random seeds	17, 29, 43
Training budget	30 epochs per seed, batch size 128
Optimizer	AdamW [7]; betas (0.9, 0.999); epsilon 1e-8
Learning rate	0.001 initially; cosine schedule toward 0.00001 over 30 epochs, no restarts [8]
Weight decay / gradient clip	0.0001 / global gradient norm 1.0
Loss	Binary cross-entropy with positive-class weight 1,213/3,477 = 0.348864
Precision and hardware	float32; Apple M2, 8 GB unified memory; PyTorch MPS

Table 3. Training configuration. Loss weighting depends only on the retained training labels.

For a given seed, both architectures receive the same epoch-specific training permutations. The cosine schedule is advanced after each epoch. All 30 epochs are completed, and the checkpoint with maximum validation ROC AUC is retained; exact numerical ties keep the earlier epoch. No architecture-specific hyperparameter search is performed.

4. Evaluation and reproducibility

4.1 Separation of fitting, model selection, and testing

The training program reads only the retained training and validation arrays. For each selected checkpoint, the classification threshold maximizes validation Youden J, defined as sensitivity plus specificity minus one. Numerical ties within 1e-12 select the threshold closest to 0.5, then the higher threshold. All six checkpoints and thresholds are saved in a frozen manifest before the evaluation program reads the test images. No seed, threshold, or checkpoint is selected using a test metric.

Pneumonia is the positive class throughout. A score greater than or equal to the selected threshold predicts pneumonia. ROC AUC is the primary endpoint. Threshold-dependent metrics use each run's validation-selected threshold; a fixed-0.5 sensitivity analysis is retained in the supplementary results. Metrics are computed independently for the three runs of each architecture. Their arithmetic mean and sample standard deviation describe variation across initialization seeds on the same test images.

Metric	Definition or computation
ROC AUC	Area under the ROC curve from continuous pneumonia scores
Average precision (AP)	Stepwise precision-recall summary from continuous scores
Accuracy / balanced accuracy	(TP+TN)/N; (sensitivity+specificity)/2
Sensitivity / specificity	TP/(TP+FN); TN/(TN+FP)
Precision / F1	TP/(TP+FP); 2TP/(2TP+FP+FN)
Brier score	Mean squared difference between sigmoid score and binary dataset label

Table 4. Test metric definitions. Scores are not calibrated clinical risks.

4.2 Image-level uncertainty and diagnostic displays

Seed 29 was fixed in the protocol as the representative run for ROC, precision-recall, confusion matrices, and error displays. For that seed, 2,000 stratified image-bootstrap replicates preserve the test class counts of 234 normal and 387 pneumonia images. The same resampled indices are used for both models. Percentiles 2.5 and 97.5 give descriptive intervals for each AUC and for the paired hybrid-minus-CNN AUC difference. The bootstrap random seed is 20261114. These intervals condition on the fitted models and sampled images; unknown patient clustering is not modeled.

4.3 Execution records

The rerun used Python 3.12.14, PyTorch 2.15.0, NumPy 2.3.5, SciPy 1.18.1, and scikit-learn 1.9.1 on macOS 15.1.1. Protocol and training-source hashes were recorded before fitting and verified after all runs. Supplementary files include source code, the protocol, dataset digests, retained indices, epoch logs, checkpoint-selection records, and per-image predictions. Fixed seeds document the run configuration; bit-identical retraining across devices or library versions is not assumed.

5. Results

5.1 Aggregate held-out performance

The CNN obtains mean test ROC AUC 0.9704, compared with 0.9537 for the CNN-Transformer. The hybrid-minus-CNN difference in seed-averaged AUC is -0.0167. Mean balanced accuracy is 90.21% and 86.14%, respectively. Table 5 reports all prespecified summary metrics, including sensitivity and specificity, so that accuracy alone does not determine the comparison.

Test metric	CNN: mean +/- SD	CNN-Transformer: mean +/- SD
ROC AUC	0.970 +/- 0.001	0.954 +/- 0.007
Average precision	0.970 +/- 0.002	0.962 +/- 0.007
Accuracy (%)	92.22 +/- 0.73	88.84 +/- 1.19
Balanced accuracy (%)	90.21 +/- 1.00	86.14 +/- 1.52
Sensitivity (%)	98.36 +/- 0.15	97.07 +/- 0.91
Specificity (%)	82.05 +/- 2.14	75.21 +/- 3.08
Precision (%)	90.07 +/- 1.05	86.64 +/- 1.43
F1	0.940 +/- 0.005	0.916 +/- 0.009
Brier score	0.096 +/- 0.012	0.123 +/- 0.024

Table 5. Held-out results on 621 deduplicated test images, three seeds per model. SD is the sample standard deviation across runs, not a population confidence interval. Threshold-dependent metrics use validation-selected thresholds.

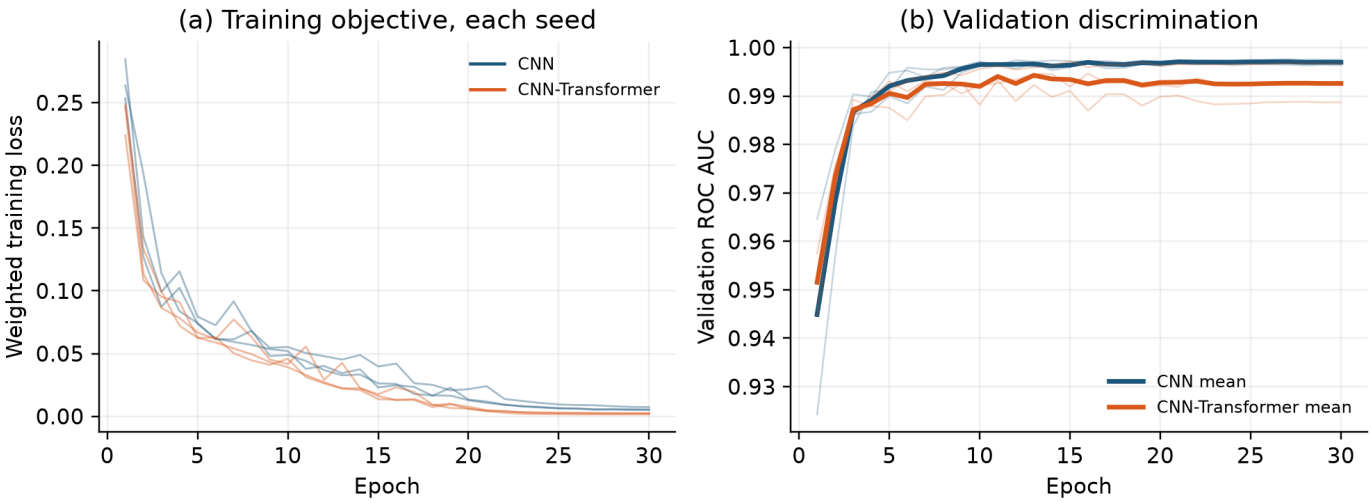

Figure 2. Recorded training loss and validation AUC for all six runs. Thin lines identify separate seeds; thick validation curves are architecture means. Checkpoint selection uses individual validation curves.

All six runs complete 30 epochs. Mean recorded training-and-validation duration is 94.4 seconds per CNN run and 172.3 seconds per hybrid run on the stated M2 system, including checkpoint writes. These wall-clock measurements describe this execution rather than deployment throughput. The CNN has the higher mean test AUC in this comparison; adding attention does not provide a consistent advantage.

5.2 Per-seed behavior and classification errors

Model	Seed	Epoch	Val cut	Test AUC	Acc. (%)	Sens. (%)	Spec. (%)
CNN	17	23	0.909	0.9713	92.91	98.19	84.19
CNN	29	27	0.903	0.9695	92.27	98.45	82.05
CNN	43	21	0.908	0.9704	91.47	98.45	79.91
Hybrid	17	11	0.486	0.9572	87.92	96.12	74.36
Hybrid	29	16	0.996	0.9579	90.18	97.16	78.63
Hybrid	43	15	0.903	0.9461	88.41	97.93	72.65

Table 6. All six selected checkpoints and their held-out results. Epoch and threshold are chosen using validation data; no test-based run exclusion is applied.

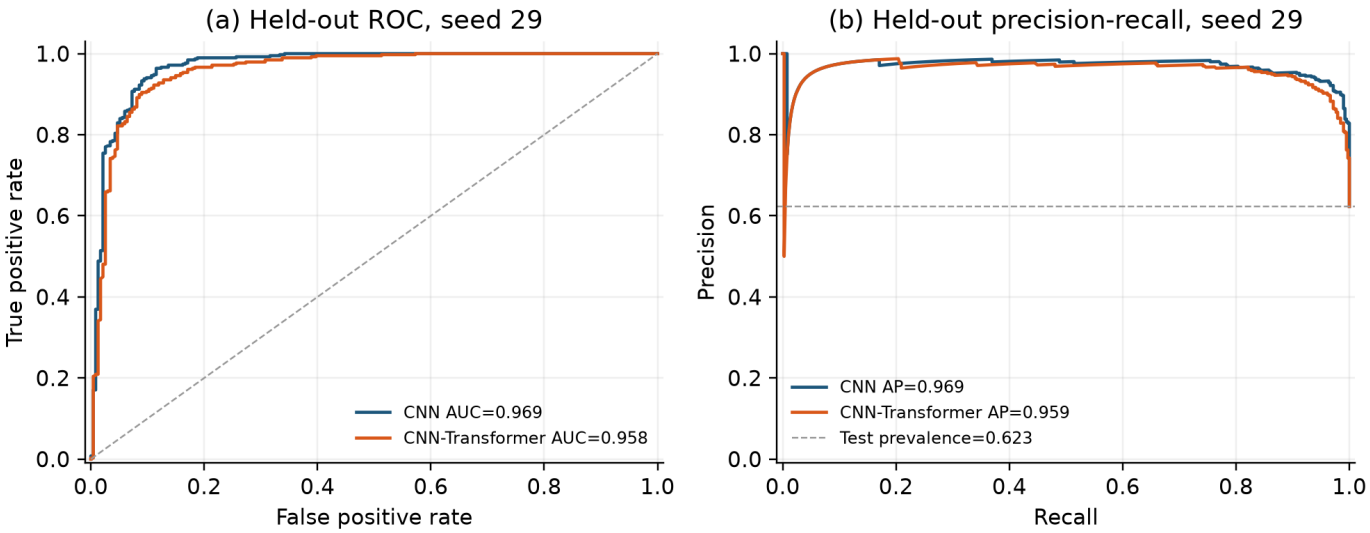

Figure 3. Actual test ROC and precision-recall curves for protocol-fixed seed 29. The dashed precision baseline equals the retained test pneumonia prevalence, 387/621.

5.3 Image-level uncertainty

Seed-29 AUC quantity	Estimate	Image-bootstrap 95% interval
CNN	0.9695	[0.9539, 0.9833]
CNN-Transformer	0.9579	[0.9394, 0.9744]
Hybrid minus CNN	-0.0116	[-0.0239, -0.0008]

Table 7. Paired stratified image-bootstrap analysis for prespecified seed 29 (2,000 replicates). Intervals do not incorporate patient clustering or the full range of training uncertainty.

The paired AUC-difference interval excludes zero for this representative run. Its interpretation remains conditional on one fitted pair and image-level resampling, rather than new patients or institutions.

5.4 Classification errors

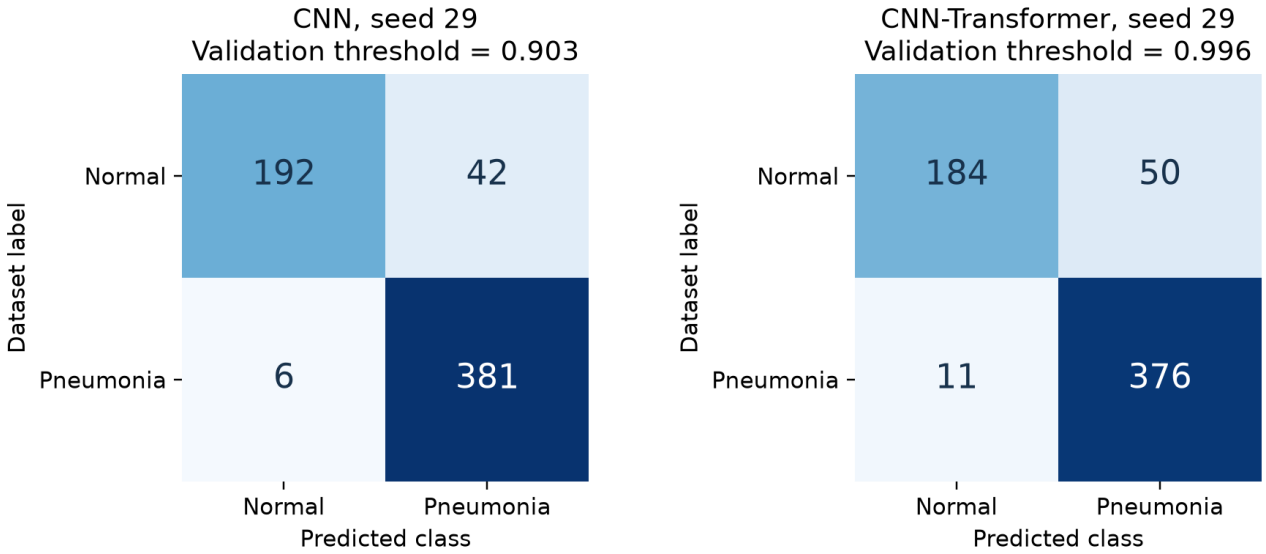

Figure 4. Test confusion matrices for seed 29 using each model's validation-selected threshold. Rows are dataset labels and columns are predictions; each matrix sums to 621 images.

At the validation-selected threshold, the seed-29 CNN makes 42 false-positive and 6 false-negative predictions. The corresponding hybrid counts are 50 and 11. Figure 5 displays the first retained NPZ test index for each model and error type. This deterministic rule avoids selecting only visually persuasive examples. The images and scores are linked to per-image prediction files.

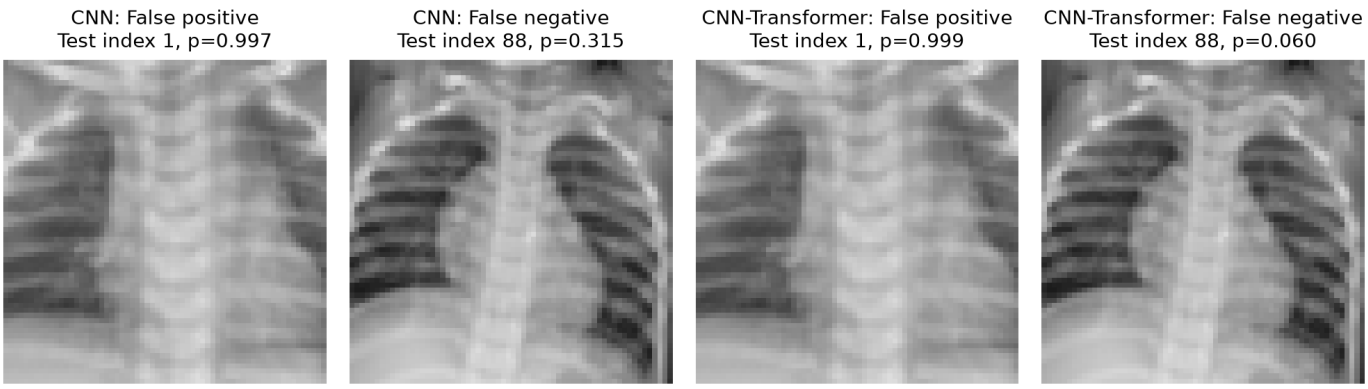

Figure 5. Misclassified public test radiographs for seed 29. Each panel shows the first original test index meeting its error category and the recorded pneumonia score. Source images: PneumoniaMNIST+ [2,3], CC BY 4.0. No diagnostic annotations or lesion boundaries are inferred.

The error displays support traceability of the numerical evaluation. They do not establish which anatomical structures caused a prediction. No localization accuracy is calculated because lesion masks and reader-reviewed localization targets are absent from the experiment.

6. Discussion and conclusion

Under a common training budget and closely matched parameter counts, CNN has the higher seed-averaged test ROC AUC. This comparison changes the context head while retaining a common stem, input resolution, loss, and selection procedure. It provides an empirical comparison of these implementations; it does not isolate every architectural difference or establish a general ranking of convolution and attention.

The benchmark has several limits. Exact-pixel hashing does not detect all near duplicates, and unavailable patient identifiers prevent a patient-level independence audit. Test images come from one public pediatric source, with a different class balance from training. There is no external-center evaluation, prospective workflow, onset-time target, reader study, or clinical outcome endpoint. Weighted training and validation-selected thresholds also mean that sigmoid scores should not be interpreted as calibrated disease risks. Three seeds describe limited optimization variation; larger model and training searches could change the comparison.

The CNN has the higher mean test AUC in this comparison; adding attention does not provide a consistent advantage. The main outcome is a reproducible chest-radiograph classification experiment with specified data provenance, retained indices, fixed model-selection rules, and predictions supporting the reported tables and figures. Further work should assess patient-separated data, external institutions, probability calibration, and higher-resolution inputs before considering clinical use.

Data and code availability

The public dataset is available at https://zenodo.org/records/10519652 [3]. Source code, protocol, data digests, retained indices, logs, and test predictions accompany this revision as a supplementary ZIP on the article page. Model checkpoints are retained in the revision's local experiment archive. Only public benchmark images and labels were used; no participants were recruited and no additional clinical records were accessed.

References

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

[2] Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

[3] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2024). *MedMNIST+* (Version 3.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10519652

[4] Wu, Y., & He, K. (2018). Group normalization. *arXiv*. https://doi.org/10.48550/arXiv.1803.08494

[5] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

[6] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv*. https://doi.org/10.48550/arXiv.1607.06450

[7] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*. https://arxiv.org/abs/1711.05101

[8] Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*. https://arxiv.org/abs/1608.03983