Class-Imbalance-Corrected Multi-Label Classification of Thoracic Pathologies in Resource-Constrained Chest Radiograph Screening

Timothy Ayelagbe
Dept. of Electronic and Electrical Engineering
Obafemi Awolowo University
Ile-Ife, Nigeria
timothyayelagbe@gmail.com

Promise Osunyemi
Dept. of Electronic and Electrical Engineering
Obafemi Awolowo University
Ile-Ife, Nigeria
promiseosunyemi04@gmail.com

Glory Akanbi
Dept. of Electronic and Electrical Engineering
Obafemi Awolowo University
Ile-Ife, Nigeria
gloryinioluwa1@gmail.com

Abstract*—Chest radiograph screening in resource-constrained settings demands models robust to severe imbalance under limited compute. We train dense and convolutional neural networks (CNNs) on ChestMNIST—14 thoracic pathologies, 78,468/11,219/22,433 patient-disjoint train/validation/test images—and correct a $96.6\times$ imbalance ratio (Hernia 0.18% vs. Infiltration 17.73%) using a class-weighted focal loss with square-root-dampened inverse-frequency weights. Replacing the fixed 0.5 decision threshold with per-label, validation-set F1-optimal thresholds raises mean test F1 from 0.0057 to 0.1786 and mean recall from 0.0031 to 0.3600. A controlled ablation against plain binary cross-entropy (BCE) at matched optimized thresholds shows the focal-loss model achieves higher mean recall (0.3600 vs. 0.2911, $+24\%$) and mean F1 (0.1786 vs. 0.1757) at comparable mean area under the curve (AUC) (0.7241 vs. 0.7254), isolating loss from thresholding; the gain is sharpest for Hernia (rarest label), where BCE collapses to F1 $= 0.000$ while focal loss reaches F1 $= 0.049$. Three architectures are compared under an identical training protocol: a 28×28 dense network (AUC 0.7241, 578,894 params), a 64×64 dense network (AUC 0.7202, 2,274,638 params), and a 28×28 CNN (AUC 0.7582, 111,886 params). The CNN outperforms both dense variants with $5.2\times$ fewer parameters, confirming that spatial structure, not raw resolution, dominates at this scale. A BatchNorm recalibration check confirms the CNN gain is not a checkpoint artifact (AUC shift $+0.0006$). Bootstrap 95% confidence intervals (CIs) (1000 resamples) confirm discriminative signal above chance for all 14 pathologies; gradient saliency maps reveal diffuse activation for lowest-AUC small-lesion diseases. Results support a tiered, human-in-the-loop deployment model for low- and middle-income country (LMIC) screening.

***Index Terms*—chest radiograph classification, class imbalance, focal loss, multi-label classification, ChestMNIST, threshold optimization, convolutional neural network, LMIC screening**

I. Introduction

Automated chest radiograph triage is attractive for low- and middle-income country (LMIC) health systems facing severe radiologist shortages [1], where deep-learning-based screening has already demonstrated real-world diagnostic value against high-burden diseases in resource-limited clinics [9]. Yet two problems routinely inflate reported performance in this literature: severe multi-label imbalance and a uniform 0.5 threshold that is statistically meaningless for rare pathologies. Using the patient-disjoint ChestMNIST benchmark — a lightweight, standardized repackaging [3] of the U.S. National Institutes of Health (NIH) ChestX-ray14 corpus [2] — we quantify and correct both issues under a single reproducible, single-architecture-family protocol.

Most published multi-label classifiers report only aggregate AUC and one fixed-threshold accuracy figure [4], metrics that hide a key failure mode: under severe imbalance, a model that never flags a rare disease can still post high nominal accuracy. Our novelty lies not in a new loss or architecture, but in a fully controlled protocol backing four claims usually asserted, not tested, elsewhere: (1) a class-weighted focal loss [5] with square-root-dampened inverse-frequency weighting, validated against plain BCE via a matched-threshold ablation that isolates loss design from thresholding; (2) per-label F1-optimal thresholds selected on a held-out validation set and applied uniformly across models; (3) a same-protocol 28×28 vs. 64×64 dense/CNN comparison isolating input resolution from architectural bias, with a BatchNorm-recalibration diagnostic directly testing—rather than assuming—the cause of any train/test AUC gap; and (4) bootstrap confidence intervals flagging low-support, wide-interval per-disease AUC estimates, paired with gradient saliency maps, to ground a discussion of deployment feasibility in resource-constrained settings.

A. Related Work

Large public chest radiograph corpora — ChestX-ray14 [2] and CheXpert [4] — underpin most deep multi-label thoracic

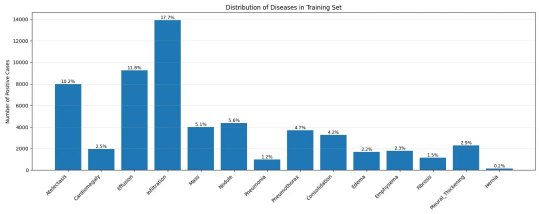

Fig. 1. Distribution of the 14 ChestMNIST pathology labels in the training set (78,468 images). Bars are annotated with prevalence (%); the imbalance ratio between Infiltration (17.7%) and Hernia (0.2%) is 96.6×.

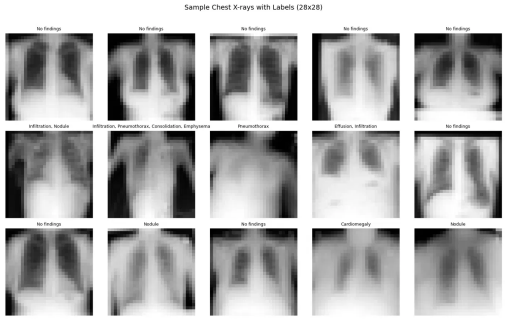

Fig. 2. Representative 28 × 28 ChestMNIST chest radiographs with their ground-truth multi-label annotations.

classifiers, including CheXNet [1], but are typically evaluated at full clinical resolution and fixed operating points, obscuring the threshold sensitivity quantified here. ChestMNIST [3] standardizes ChestX-ray14 into a low-resolution, MNIST-style benchmark, trading clinical fidelity for reproducibility and near-zero compute cost, making it a natural proxy for LMIC-grade hardware constraints. Class imbalance in convolutional networks is systematically studied by Buda et al. [6], motivating our inverse-frequency weighting scheme; the focal loss of Lin et al. [5], originally proposed for single-stage detectors facing extreme foreground–background imbalance, is the loss family we adapt to the multi-label medical setting. Calibration failures of modern deep networks, including the unreliability of the default 0.5 threshold, are documented by Guo et al. [7], supporting our use of validation-set F1-optimal thresholds. Deployment feasibility for LMIC screening is directly evidenced by Qin et al. [9], who evaluated deep-learning chest radiograph triage against tuberculosis in field clinics across five countries with radiologist-comparable sensitivity — the clinical precedent our tiered deployment model builds on. Oakden-Rayner [8] cautions that natural-language-processing (NLP)-extracted labels underlying ChestX-ray14/ChestMNIST carry systematic annotation noise, a limitation revisited in Section VI.

II. Dataset and Class Imbalance

ChestMNIST provides 78,468 training, 11,219 validation, and 22,433 test grayscale chest radiographs with 14 independent binary disease labels (multi-label, multi-class task), at a patient-disjoint split derived from the NIH ChestX-ray14 corpus [2] and repackaged at low resolution by the MedMNIST v2 benchmark [3]. Fig. 1 shows the per-label positive case distribution in the training set. Label prevalence ranges from 0.18% (Hernia, 144 cases) to 17.73% (Infiltration, 13,914 cases), an imbalance ratio of 96.6×. A model trained with unweighted BCE can trivially achieve high nominal accuracy by predicting the negative class for every label, since this is already correct for the overwhelming majority of (image, label) pairs. Fig. 2 shows representative training images with their ground-truth labels.

III. Methods

A. Architecture

The baseline dense network is a funnel-shaped fully connected network with four blocks: Dense(512), Dense(256), Dense(128), and Dense(64), each followed by BatchNormalization (BN) (omitted on the final block) and Dropout (rates 0.2, 0.2, 0.1, 0.1), then an output layer Dense(14, sigmoid). Independent sigmoid outputs (not a softmax) are required because each image may carry zero, one, or several disease labels simultaneously. At 28 × 28 input resolution (784 flattened features) this architecture has 578,894 trainable parameters.

The CNN comparator preserves the 2-D spatial structure of the input through three convolutional blocks — Conv2D(32)+BN+MaxPool, Conv2D(64)+BN+MaxPool, Conv2D(128)+BN — followed by GlobalAveragePooling2D, Dense(128, rectified linear unit (ReLU) activation), Dropout(0.3), and the same Dense(14, sigmoid) head. It has 111,886 trainable parameters, 5.2× fewer than the 28 × 28 dense model, and is trained with identical loss, class weights, and callback configuration so any performance difference reflects architecture rather than optimization protocol.

B. Class-Weighted Focal Loss

Per-label positive frequency p_i is computed from the training set only. Raw inverse-frequency weights $w_i^{raw} = (1 - p_i)/p_i$ range up to 543.9× for Hernia, which empirically causes severe precision collapse if applied directly [6]. We instead use square-root-dampened weights:

$$w_i = \sqrt{\frac{1 - p_i}{p_i}} \tag{1}$$

which compresses the Hernia weight to 23.32× while still upweighting rare labels (Table I). These weights enter a class-weighted focal loss, adapted from the single-label formulation of Lin et al. [5] to the independent-sigmoid multi-label setting:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{14}\Big[w_i\,\alpha\,y_i(1-\hat{y}_i)^{\gamma}\log\hat{y}_i + (1-\alpha)(1-y_i)\hat{y}_i^{\gamma}\log(1-\hat{y}_i)\Big] \tag{2}$$

with $\gamma = 2.0$ and $\alpha = 0.25$, matching the defaults recommended in [5]. The focal modulation term $(1 - \hat{y}_i)^{\gamma}$ down-weights easy, confidently-correct examples (overwhelmingly true negatives under this imbalance), concentrating gradient

TABLE I
PER-LABEL PREVALENCE AND CLASS WEIGHTS (TRAINING SET)

Disease	Prev. (%)	Raw weight	Moderated w_i
Infiltration	17.73	4.64	2.15
Effusion	11.80	7.47	2.73
Atelectasis	10.19	8.81	2.97
Nodule	5.58	16.94	4.12
Mass	5.08	18.68	4.32
Pneumothorax	4.72	20.18	4.49
Consolidation	4.16	23.05	4.80
Pleural Thickening	2.90	33.43	5.78
Emphysema	2.29	42.62	6.53
Cardiomegaly	2.49	39.24	6.26
Edema	2.15	45.43	6.74
Fibrosis	1.48	66.76	8.17
Pneumonia	1.25	79.23	8.90
Hernia	0.18	543.92	23.32

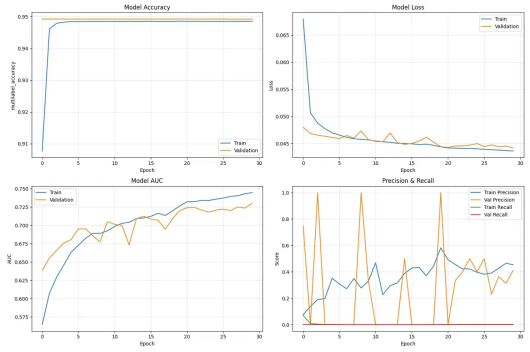

Fig. 3. Training history for the 28×28 dense network with class-weighted focal loss. Top-left: mean per-label accuracy saturates near 0.948 almost immediately, masking the much more gradual AUC improvement (bottom-left, train $0.566 \rightarrow 0.749$, validation $0.638 \rightarrow 0.730$). Validation precision/recall (bottom-right) are computed at the fixed 0.5 threshold and are consequently near zero for most epochs.

signal on hard examples; w_i separately penalizes missed rare positives more than missed common positives.

C. Training Protocol

All models use Adam (initial learning rate 5×10^{-4}), batch size 128, up to 30 epochs, `ReduceLROnPlateau` (factor 0.5, patience 4 on validation loss), and `EarlyStopping` on validation AUC (patience 15, best weights restored). For the reported 28×28 dense run, training reached its best validation AUC at epoch 30 (0.7301); Fig. 3 shows the full training trajectory, including a multilabel-accuracy curve plateauing near 0.948–0.949 from epoch 2 onward despite the much lower, more informative AUC and F1 trajectories — direct evidence accuracy alone poorly indicates model quality under this imbalance.

D. Per-Label Threshold Optimization

For each of the 14 labels independently, we search 200 candidate thresholds in $[0.01, 0.99]$ on validation-set predictions and select the threshold maximizing F1, then apply these frozen thresholds to the held-out test set. Optimal thresholds

TABLE II
SELECTED PER-LABEL OPTIMAL THRESHOLDS (28×28 DENSE MODEL)

Disease	Thr.	Disease	Thr.
Atelectasis	0.34	Pneumothorax	0.36
Cardiomegaly	0.34	Consolidation	0.32
Effusion	0.37	Edema	0.37
Infiltration	0.33	Emphysema	0.34
Mass	0.32	Fibrosis	0.31
Nodule	0.30	Pleural Thickening	0.32
Pneumonia	0.30	Hernia	0.34

TABLE III
FIXED (0.5) VS. OPTIMIZED THRESHOLD, TEST SET, DENSE 28×28

Metric	Fixed (0.5)	Optimized
Mean F1	0.0057	0.1786
Mean Precision	0.0737	0.1233
Mean Recall	0.0031	0.3600
Overall test accuracy	0.9474	
Overall test AUC	0.7239	

cluster between 0.30 and 0.37 (Table II), far below the conventional 0.5 default, because under this imbalance a well-trained model rarely assigns rare positive classes a probability above 0.5 [7].

IV. RESULTS

A. Threshold Optimization Impact

Table III contrasts fixed 0.5-threshold and optimized-threshold evaluation on the same predicted probabilities. At the fixed threshold, 11 of 14 labels receive zero positive predictions on the test set; mean F1 is 0.0057 and mean recall is 0.0031, despite an overall `model.evaluate` accuracy of 0.9474. Switching to per-label optimal thresholds raises mean F1 to 0.1786 ($31\times$) and mean recall to 0.3600, at the cost of mean precision falling only marginally in the opposite direction (precision actually improves slightly because the fixed-threshold precision figure is dominated by near-undefined zero-positive-prediction labels). AUC is identical between the two columns by construction, since AUC depends only on predicted-probability ranking, not on where the decision boundary is drawn. Fig. 4 shows qualitative test-set predictions under the optimized per-label thresholds.

B. Ablation: Focal Loss vs. Plain BCE

To isolate the loss design's contribution from thresholding, we train an architecturally identical model with plain unweighted BCE and apply the same per-label threshold search. Table IV shows the two models reach near-identical mean AUC (BCE 0.7254 vs. focal 0.7241), confirming ranking quality is not meaningfully different. However, the focal-loss model achieves substantially higher mean recall (0.3600 vs. 0.2911, a 24% relative increase) at a modest precision cost (0.1233 vs. 0.1323), and higher mean F1 (0.1786 vs. 0.1757). The gain is most pronounced for the rarest label, Hernia,

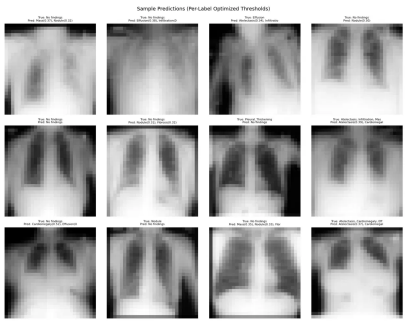

Fig. 4. Qualitative test-set predictions under per-label optimized thresholds (dense 28×28 model), comparing predicted disease labels (with confidence scores) against ground truth.

TABLE IV
ABLATION: PLAIN BCE VS. CLASS-WEIGHTED FOCAL LOSS, BOTH WITH OPTIMIZED PER-LABEL THRESHOLDS (TEST SET, MEAN ACROSS 14 LABELS)

Metric	**Plain BCE**	**Focal loss**
Mean AUC	0.7254	0.7241
Mean F1	0.1757	0.1786
Mean Precision	0.1323	0.1233
Mean Recall	0.2911	0.3600
Hernia F1 (rarest label)	0.000	0.049

where BCE collapses to F1 $= 0.000$ (zero recall) while focal loss achieves F1 $= 0.049$. This supports describing the class-weighted focal loss as a genuine contribution beyond threshold tuning, specifically for rarest pathologies.

C. Bootstrap Confidence Intervals

We compute 95% bootstrap confidence intervals for per-label AUC (1000 resamples of the test set; resamples with only one class present are skipped). Fig. 5 shows that all 14 labels' intervals lie entirely above the chance level (AUC$= 0.5$), confirming genuine discriminative signal for every pathology. Edema attains the highest AUC (0.839, 95% CI [0.822, 0.857]); Nodule the lowest (0.599, CI [0.583, 0.613]). Hernia, with only 42 positive test cases, has the widest interval (CI width 0.143), reflecting high sampling variability rather than any anatomical property of the disease; our pipeline flags this automatically whenever a CI width exceeds 0.10.

D. Confusion Matrices

Fig. 6 shows optimized-threshold confusion matrices for the four most prevalent labels. Infiltration shows the largest false-positive count (6,883), reflecting its high base rate (17.7%) and recall-favoring threshold (0.33).

E. Resolution and Architecture Comparison

Table V compares three model variants trained under an identical protocol: the 28×28 dense network (578,894 parameters), a 64×64 dense network (2,274,638 parameters, 4096 input features), and a 28×28 CNN (111,886 parameters)

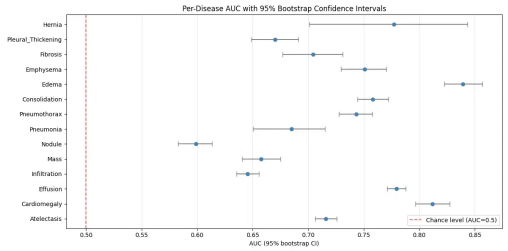

Fig. 5. Per-disease test-set AUC with 95% bootstrap confidence intervals (1000 resamples). All 14 intervals exclude the chance level of 0.5.

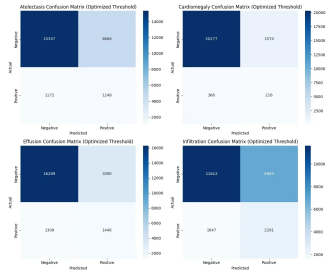

Fig. 6. Confusion matrices at optimized per-label thresholds for the four most prevalent pathologies (test set, $n = 22{,}433$).

that preserves 2-D spatial structure through three convolutional blocks before global average pooling. Increasing dense-network resolution from 28×28 to 64×64 produces a negligible, slightly negative mean AUC change (-0.0039) despite $3.9\times$ more parameters, indicating that flattening discards the spatial information higher resolution would otherwise provide. In contrast, the CNN — with under one-fifth the parameters of the 28×28 dense model — achieves the highest mean AUC (0.7582) and mean F1 (0.2067) by preserving spatial locality through convolution rather than raw pixel count.

A diagnostic check confirmed the CNN checkpoint's reported AUC is not an artifact of unstabilized BatchNormalization running statistics: re-running one full unbiased forward pass over the training set in training mode (updating only BatchNorm running statistics, not learned weights) changed mean test AUC from 0.7582 to 0.7588 — a negligible $+0.0006$ shift, indicating the original checkpoint's BatchNorm statistics had already converged adequately at the restored epoch. We report this as a directly tested result rather than an assumed explanation, since the alternative (poorly learned convolutional features) would not have been ruled out without this check.

F. Explainability

Fig. 7 shows vanilla gradient saliency maps for Nodule and Infiltration, two of the three lowest-AUC pathologies under the CNN (the third being Pneumonia). For both diseases, saliency activation is diffuse and scattered across the lung fields rather than concentrated on a discrete, anatomically localized region, in both true-positive and true-negative examples, consistent with the 28×28 resolution's inability to resolve small, focal lesions.

TABLE V
MODEL COMPARISON SUMMARY (TEST SET, OPTIMIZED THRESHOLDS, MEAN ACROSS 14 LABELS)

Model	AUC	F1	Recall	Params
Dense (28×28)	0.7241	0.1786	0.3600	578,894
Dense (64×64)	0.7202	0.1803	0.3595	2,274,638
CNN (28×28)	**0.7582**	**0.2067**	0.3277	**111,886**

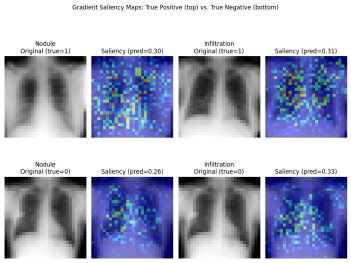

Fig. 7. Gradient saliency maps for Nodule and Infiltration (CNN model), true positive (top row) vs. true negative (bottom row) examples. Activation is diffuse rather than spatially localized for both labels.

V. DISCUSSION: CLINICAL FRAMING FOR RESOURCE-CONSTRAINED SCREENING

In a high-volume, low-resource screening context, a missed disease (false negative) typically carries far higher cost than an unnecessary radiologist review (false positive), favoring high-recall operating points provided precision remains workable — the trade-off underlying the field-deployed tuberculosis triage systems evaluated by Qin et al. [9]. Diseases with AUC > 0.75 (Edema, Cardiomegaly, Effusion) show the strongest discriminative ability, but even at optimized thresholds their precision remains modest (e.g. Effusion precision 0.299), so any "screening-ready" claim must be paired with actual precision/recall figures rather than AUC alone. Mid-range diseases (AUC 0.65–0.75: Pneumothorax, Consolidation, Atelectasis, Fibrosis, Emphysema, Hernia) are useful only as weak supplementary signal, while those below AUC 0.65 (Nodule, Infiltration, Mass) are close to chance at this resolution — precisely the small, focal findings lost when downsampling to 28×28, consistent with the diffuse saliency patterns in Fig. 7.

These results support a tiered deployment model for LMIC settings, mirroring low-cost, high-throughput triage pipelines already validated in field conditions [9]: (1) a low-resolution, low-compute CNN tier for coarse triage of high-prevalence, large-extent findings, deployable on the modest hardware typical of LMIC primary-care facilities; (2) escalation of ambiguous or focal-finding-suspected cases to a higher-resolution pipeline; (3) mandatory human-in-the-loop review of every positive flag, since precision at the F1-optimal operating point remains well below levels that would support autonomous diagnosis.

VI. LIMITATIONS

ChestMNIST labels are NLP-extracted from radiology reports and carry inherent label noise [8]. The $28 \times 28/64 \times 64$ resolutions studied are far below clinical radiograph resolution ($\geq 1024 \times 1024$); absolute performance should not be extrapolated to full-resolution deployment. The CNN's smaller parameter count is an uncontrolled confound between architecture and capacity; a parameter-matched CNN would be needed to attribute the full gap to architecture alone. The resolution comparison reflects a single run per resolution, so small AUC differences should be read as within run-to-run variation rather than a confirmed null result. Bootstrap confidence intervals characterize sampling variability on this test split only and do not account for distributional shift to other populations or imaging equipment.

VII. CONCLUSION

A class-weighted focal loss with square-root-dampened inverse-frequency weighting, combined with per-label F1-optimal decision thresholds, corrects the illusory high-accuracy/near-zero-recall failure mode of naively trained multi-label chest radiograph classifiers under $96.6\times$ class imbalance, raising mean test F1 by $31\times$ over the fixed-threshold baseline. A controlled ablation confirms the loss design contributes genuine recall gains beyond threshold tuning alone, particularly for the rarest pathology (Hernia). Across three architectures trained under an identical protocol, a compact CNN with $5.2\times$ fewer parameters than the 28×28 dense baseline achieves the best mean AUC (0.7582) and F1 (0.2067), indicating that preserved spatial structure is more valuable than raw input resolution at this dataset scale; a BatchNorm-recalibration check confirms this is not an artifact of an undertrained checkpoint. All 14 pathologies show bootstrap-confirmed discriminative signal above chance, but precision at clinically actionable operating points remains low enough that any LMIC screening deployment should be tiered and human-in-the-loop rather than autonomous.

REFERENCES

[1] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017.

[2] X. Wang *et al.*, "ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2097–2106.

[3] J. Yang *et al.*, "MedMNIST v2 — A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, p. 41, 2023.

[4] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 590–597, Jul. 2019, doi: 10.1609/aaai.v33i01.3301590.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.

[6] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, 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 *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 70, 2017, pp. 1321–1330.

[8] L. Oakden-Rayner, "Exploring the ChestXray14 dataset: problems," *laurenoakdenrayner.com*, Dec. 18, 2017. [Online]. Available: https://laurenoakdenrayner.com/2017/12/18/the-chestxray14-dataset-problems/

[9] Z. Z. Qin *et al.*, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Scientific Reports*, vol. 9, p. 15000, 2019.