



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

What Matters for Lightweight Multi-Label Chest X-Ray Classification? A Controlled Multi-Seed Study of Attention, Loss, Width and Resolution

Md Sakawat Hossain¹; Mst Kamrun Nahar Borsha²

¹School of Artificial Intelligence, Nanjing University of Information Science & Technology

²School of Computer Science & Technology, Nanjing University of Information Science & Technology

¹sakawathossain2000@gmail.com

DOI: <https://doi.org/10.47760/ijcsmc.2026.v15i10.001>

Abstract: Attention modules and imbalance-aware losses are often credited with gains in multi-label chest X-ray classification, but their effects are seldom separated from training length, model width, input resolution and seed-to-seed variation. We integrate three existing components, squeeze-and-excitation (SE) channel attention, the class-specific residual attention (CSRA) classifier and the asymmetric loss (ASL), into a lightweight residual network (SECA-Net) and evaluate them on the 14-label ChestMNIST benchmark in 87 training runs. A 2×2×2 factorial ablation with five seeds, a loss study, a width study, three lightweight reference models and a 64- and 128-pixel resolution study were all trained to convergence under one protocol. None of the three components changed macro-AUROC significantly (seed-blocked three-way ANOVA: SE $p = 0.84$, CSRA $p = 0.29$, ASL $p = 0.86$), and their combination did not significantly beat the baseline at width 64 (+0.0018, $p = 0.085$) or at 64 and 128 pixels. In contrast, input resolution raised macro-AUROC from 0.789 at 28 pixels to 0.815 at 128 pixels with the same 0.70 M-parameter network, training to convergence instead of eight epochs added 0.014, and widening the network from 16 to 64 base channels added 0.007. ASL and weighted cross-entropy left discrimination unchanged but made probabilities poorly calibrated (class-wise expected calibration error 0.31 versus 0.006 for cross-entropy), which per-class Platt scaling corrected. Differences of up to 0.01 AUROC between single runs can arise from seed-to-seed variation and GPU non-determinism alone, so multi-seed evaluation is essential. Code, trained weights and machine-readable results are released.

Keywords: Chest X-ray, multi-label classification, attention mechanism, asymmetric loss, ablation study, calibration, reproducibility, MedMNIST



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

I. INTRODUCTION

Chest radiography (CXR) is one of the most frequently performed imaging examinations, and deep convolutional neural networks (CNNs) have been studied extensively for its automated interpretation [1]. Public collections such as NIH ChestX-ray14 [2], CheXpert [3], MIMIC-CXR [4] and PadChest [5] made large-scale training possible, and retrospective studies such as CheXNet [6] and CheXNeXt [7] reported performance comparable to radiologists on selected findings. Translating such results into clinical benefit is much harder [8], and models trained at one institution often degrade at another [9], [10].

CXR classification is a *multi-label* problem: one radiograph may show several findings, while more than half of all images carry no positive label. The label distribution is long-tailed (in the ChestX-ray14 training split the most frequent finding is 97 times more common than the rarest), labels are correlated [11], [12], and findings occupy different anatomical regions, so one globally pooled feature vector mixes evidence for unrelated findings [13], [14]. Attention mechanisms and imbalance-aware losses are common responses to these properties, but their individual contributions are rarely isolated under controlled conditions.

Low-resolution benchmarks such as ChestMNIST [15] make controlled studies affordable. They also have clear limits: at 28×28 pixels a small nodule is smaller than one pixel, and the ChestX-ray14 labels were mined from reports and contain errors [16]. Two further problems are common in published results. Differences between methods are often reported from a single run, although run-to-run variation can be of the same size [17], and models are frequently trained for a short-fixed schedule instead of to convergence. Our own preliminary experiments illustrate both: after eight epochs the integrated network appeared to beat the baseline by 0.0026 macro-AUROC, a gain that disappeared once both models were trained to convergence (Section VI-A).

This paper is therefore designed as a controlled study. Its contributions are:

- We *integrate* three existing components into one lightweight network, referred to as SECA-Net for brevity: squeeze-and-excitation (SE) channel attention [18], the class-specific residual attention (CSRA) classifier [19], adopted without modification, and the asymmetric loss (ASL) [20]. No new module is claimed.
- We run a $2 \times 2 \times 2$ factorial ablation (backbone $\times$ classifier head $\times$ loss) with five seeds per configuration, a loss study, a width study (16 and 64 base channels), three lightweight reference models and a resolution study at 64 and 128 pixels (MedMNIST+ [21]): 87 runs, all trained to convergence under one protocol.
- We evaluate with validation-only thresholds, seed-paired tests, a blocked factorial ANOVA, ensemble-level bootstrap and DeLong tests [22] with Holm correction [23], pooled and class-wise calibration error [24], [25], and trivial baselines for decision-level metrics.
- We report a negative result: none of the three components improves discrimination; resolution, training length and width do. Code, weights and all results are released.

II. RELATED WORK

A. CNNs, Label Dependencies and Attention

Wang *et al.* [2] introduced ChestX-ray14 with report-mined labels and an ImageNet-pretrained ResNet-50 baseline; CheXNet [6] fine-tuned a DenseNet-121 [26]. Baltruschat *et al.* [27] showed that results depend markedly on the data split, so numbers from different splits are not comparable. Label dependencies were modelled with recurrent decoders [11] and graph convolutional networks [12]. Category-wise residual attention [13] and cross-attention networks [14] aim to localise the evidence for each finding. In general multi-label recognition, CSRA [19] adds a spatially attended score to the average-pooled score of each class at negligible cost, and SE blocks [18] re-weight channels inside the backbone.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

B. Imbalanced Labels and Calibration

Because every image contributes a negative example for almost every class, negative terms dominate the binary cross-entropy (BCE) gradient. Remedies include class-weighted BCE [2], focal loss [28] and ASL [20]. Long-tailed CXR classification was benchmarked by Holste *et al.* [29] and in the CXR-LT challenge [30], where re-weighting and loss design helped rare findings but gains were modest and dataset-dependent. The effect of these losses on probability calibration [24] is rarely reported.

C. Transformers, State-Space and Foundation Models

Vision transformers [31] and the Swin Transformer [32] have been applied to multi-label CXR classification, for example in SwinCheX [33]. Selective state-space models (Mamba) [34] were adapted to vision in Vision Mamba [35] and to medical images in MedMamba [36], and vision-language models such as CheXzero [37] detect pathologies without explicit labels. These models are usually pre-trained on large corpora and evaluated at 224×224 pixels or more. This study deliberately stays in the lightweight, train-from-scratch regime and asks which inexpensive design choices matter when compute is limited.

D. Lightweight Models and Low-Resolution Benchmarks

MobileNetV2 [38] and EfficientNet [39] are standard efficient CNN families. MedMNIST v2 [15] standardised 18 biomedical datasets at 28×28 pixels with fixed splits; for ChestMNIST it reports a test AUROC of 0.768 for ResNet-18 and 0.778 for Google AutoML Vision. MedMNIST+ [21] added 64-, 128- and 224-pixel versions and found that convolutional networks remain competitive with vision transformers at these sizes.

III. DATASET AND PREPROCESSING

A. ChestMNIST and MedMNIST+

ChestMNIST [15] is derived from NIH ChestX-ray14 [2]: 112,120 frontal radiographs of 30,805 patients with 14 report-mined labels. MedMNIST down-samples the 1024×1024 images and keeps the official patient-disjoint test list, giving 78,468 training, 11,219 validation and 22,433 test images. The 64- and 128-pixel versions of MedMNIST+ [21] contain the same images in the same splits. All files are used as distributed and verified against the official MD5 checksums. Fig. 1 shows representative 28×28 test images.

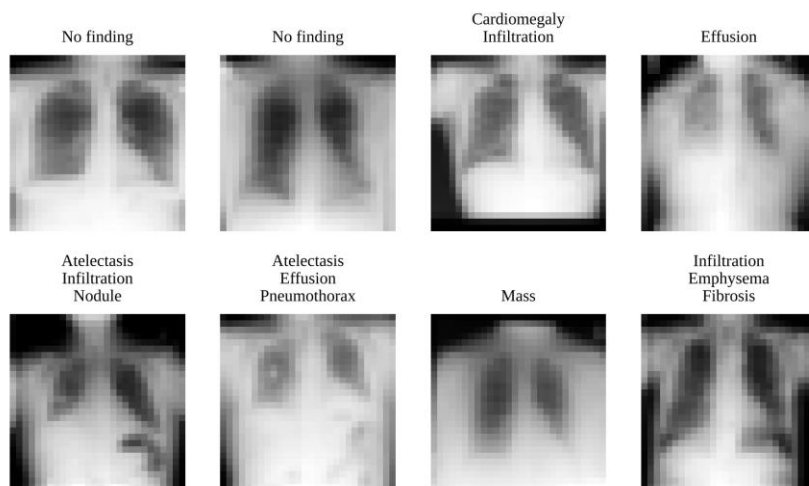


Fig. 1 Representative ChestMNIST test images (28×28 pixels) with their ground-truth labels.



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X
Impact Factor: 7.056

B. Label Statistics

Table I lists the positive images per class. Infiltration appears in 17.7% of training images and Hernia in only 0.18%. 54.0% of training images have no positive label and the mean label cardinality is 0.72. One consequences matter for evaluation. An all-negative predictor already matches every label of 53.2% of the test images and attains a per-label accuracy of 0.947, so exact-match and accuracy figures must be read against these trivial baselines. Fig. 2 (right) shows the conditional co-occurrence $P(\text{column} | \text{row})$; the strongest dependencies are $P(\text{Infiltration} | \text{Edema}) = 0.43$; $P(\text{Infiltration} | \text{Pneumonia}) = 0.42$; $P(\text{Effusion} | \text{Cardiomegaly}) = 0.38$.

TABLE I
 POSITIVE IMAGES PER CLASS IN THE CHESTMNIST SPLITS

Pathology	Train	Train %	Val	Val %	Test	Test %
Atelectasis	7996	10.19	1119	9.97	2420	10.79
Cardiomegaly	1950	2.49	240	2.14	582	2.59
Effusion	9261	11.80	1292	11.52	2754	12.28
Infiltration	13914	17.73	2018	17.99	3938	17.55
Mass	3988	5.08	625	5.57	1133	5.05
Nodule	4375	5.58	613	5.46	1335	5.95
Pneumonia	978	1.25	133	1.19	242	1.08
Pneumothorax	3705	4.72	504	4.49	1089	4.85
Consolidation	3263	4.16	447	3.98	957	4.27
Edema	1690	2.15	200	1.78	413	1.84
Emphysema	1799	2.29	208	1.85	509	2.27
Fibrosis	1158	1.48	166	1.48	362	1.61
Pleural Thickening	2279	2.90	372	3.32	734	3.27
Hernia	144	0.18	41	0.37	42	0.19
No finding (all-zero)	42405	54.04	6079	54.18	11928	53.17
Images (total)	78468	100.00	11219	100.00	22433	100.00

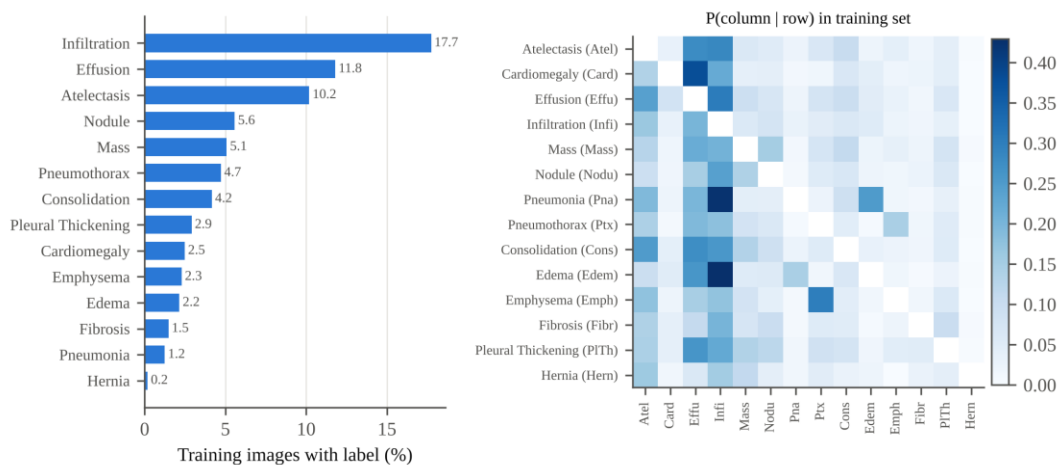


Fig. 2 Left: percentage of training images containing each finding. Right: conditional label co-occurrence $P(\text{column} | \text{row})$ in the training set.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

C. Preprocessing and Augmentation

Intensities are scaled to $[0, 1]$ and standardised with the training mean and standard deviation. Training batches are augmented with a random affine transformation (rotation $\pm 10^\circ$, scaling $\pm 10\%$, translation $\pm 8\%$) and brightness/contrast jitter of $\pm 10\%$. Horizontal flipping is not used because it reverses cardiac laterality. No augmentation is applied at validation or test time.

IV. METHODS

A. Problem Formulation

Let x be a radiograph and $y = (y_1, \dots, y_K) \in \{0, 1\}^K$ its labels over $K = 14$ findings. A network maps x to logits $z \in \mathbb{R}^K$, and each finding receives an independent probability

$$p_k = \sigma(z_k) = \frac{1}{1 + e^{-z_k}}, \quad k = 1, \dots, K. \quad (1)$$

Fig. 3 shows how the three integrated components fit together; each is switched on and off independently in the factorial ablation (Section V).

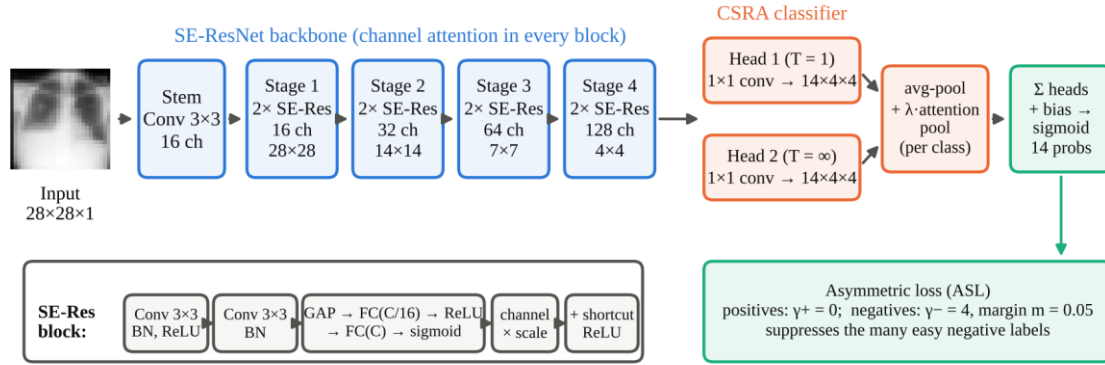


Fig. 3 The integrated network (SECA-Net) at 28×28 input and width 16. A residual backbone with squeeze-and-excitation (SE) blocks produces a 4×4 feature map (8×8 at 64 and 128 pixels), which a two-head class-specific residual attention (CSRA) classifier [19] turns into 14 logits; training uses the asymmetric loss (ASL) [20]. None of the three components is new.

B. Backbone and Width

The backbone follows the ResNet-18 topology [40] adapted to small inputs as in MedMNIST [15]: a 3×3 stem convolution (stride 1 up to 64 pixels, stride 2 at 128 pixels) without max-pooling, followed by four stages of two basic residual blocks with w , $2w$, $4w$ and $8w$ channels. With $w = 64$ this is the standard ResNet-18 (11.17 M parameters); with $w = 16$ it has 0.70 M. Width alone controls model size, so size reductions are attributed to width and not to the attention modules. SE and CSRA together add 10,016 parameters at $w = 16$ and 97,280 at $w = 64$.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

C. Squeeze-and-Excitation Channel Attention

In the SE variant every residual block contains an SE unit [18], which re-weights the channels of the residual branch $U \in \mathbb{R}^{C \times H \times W}$ before the shortcut addition:

$$s = \sigma \left(W_2 \delta \left(W_1 \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W U_{:,i,j} \right) \right), \quad \tilde{U}_c = s_c U_c, \quad (2)$$

where δ is the ReLU and W_1, W_2 form a bottleneck with reduction ratio 16 (minimum width 8).

D. Class-Specific Residual Attention Classifier

The standard head applies global average pooling (GAP) and a linear layer. The CSRA head [19] is adopted unchanged. For feature vectors x_j at the hw positions of the final map and a classifier weight m_k per class, the class-specific score at position j is

$$s_{k,j} = \frac{m_k^T x_j}{\|m_k\|_2}, \quad (3)$$

and the head combines an average-pooled score with an attention-pooled score:

$$g_k = \frac{1}{hw} \sum_{j=1}^{hw} s_{k,j}, \quad a_k(T) = \sum_{j=1}^{hw} \frac{e^{T s_{k,j}}}{\sum_{l=1}^{hw} e^{T s_{k,l}}} s_{k,j}, \quad (4)$$

$$z_k = \sum_{n=1}^2 \left(g_{k,n} + \lambda a_{k,n}(T_n) \right) + b_k, \quad (5)$$

with two heads ($T_1 = 1$ and $T_2 = \infty$, i.e. max-pooling), $\lambda = 0.1$ and a learnable bias b_k . At 28×28 the final map has only 4×4 positions, so the spatial attention is necessarily coarse: it can favour a region of the thorax but cannot localise small lesions.

E. Loss Functions

Besides BCE, the ablation uses ASL [20], which keeps the full cross-entropy gradient for positive labels and down-weights and thresholds easy negatives:

$$\mathcal{L}_{ASL} = - \sum_{k=1}^K [y_k (1 - p_k)^{\gamma_{pos}} \log p_k + (1 - y_k) p_{m,k}^{\gamma_{neg}} \log(1 - p_{m,k})], \quad p_{m,k} = \max(p_k - m, 0), \quad (6)$$

with $\gamma_{pos} = 0$, $\gamma_{neg} = 4$ and $m = 0.05$ as recommended by its authors. Weighted BCE (positive weight $N_{neg,k} / N_{pos,k}$, capped at 50) and focal loss [28] ($\gamma = 2$, $\alpha = 0.25$) are evaluated in a separate loss study.

F. Reference Models

MobileNetV2 [38], EfficientNet-B0 [39] and a plain four-block CNN are trained from scratch with BCE under the same protocol. For the torchvision models the first convolution is replaced by a single-channel convolution with stride 1, so that 28-pixel inputs are not reduced to a single spatial position.

G. Optimisation, Thresholds and Calibration

All models use AdamW [41] with a two-epoch linear warm-up and cosine decay [42],

$$\eta_t = \eta_{max} \cdot 1/2 \left(1 + \cos \left(\pi \frac{t - t_w}{T_{tot} - t_w} \right) \right), \quad (7)$$

an exponential moving average (EMA) of the weights (decay 0.999) [43] and mixed precision. Training runs for at most 100 epochs and stops when the validation macro-AUROC of the EMA weights has not improved for 20 epochs; the best validation checkpoint is kept. Per-class thresholds maximise the F1-score on the validation set and are then frozen for the test set,



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

$$\hat{\tau}_k = \arg \max_{\tau} F1_k(\tau; \mathcal{D}_{val}). \quad (8)$$

Because the thresholds are fitted on the validation set, validation F1 is optimistic by construction and only test F1 is used as a performance estimate. Probabilities are optionally re-calibrated with per-class Platt scaling [44] fitted on the validation set,

$$\tilde{p}_k = \sigma \left(a_k \log \frac{p_k}{1-p_k} + c_k \right), \quad (9)$$

which is monotone for $a_k > 0$ and therefore leaves every per-class AUROC unchanged.

V. EXPERIMENTAL SETUP

A. Experiment Grid

TABLE II
EXPERIMENT GRID

Group	Purpose	Configurations	Seeds	Runs
A	Factorial ablation, 28 px, width 16	{ResNet, SE-ResNet} $\times$ {GAP, CSRA} $\times$ {BCE, ASL}	5	40
B	Loss study	SE-ResNet + CSRA with weighted BCE, focal loss	5	10
C	Width study	ResNet-GAP-BCE and SECA-Net at width 64	5	10
D	Lightweight reference models	MobileNetV2, EfficientNet-B0, plain CNN (BCE)	5	15
E	Resolution study	ResNet-GAP-BCE and SECA-Net at 64 and 128 px	3	12
	Total			87

Table II summarises the experiments. Group A crosses the three components, so each main effect is estimated over the four settings of the other two factors and interactions can be tested directly. Groups B to E reuse the same pipeline; Table III lists the shared settings.

TABLE III
TRAINING CONFIGURATION SHARED BY ALL MODELS

Setting	Value
Hardware / software	NVIDIA Tesla T4 (Google Colab: 39 runs; Kaggle: 48 runs); PyTorch 2.11.0, Python 3.13.15
Input	28 $\times$ 28 (groups A–D); 64 $\times$ 64 and 128 $\times$ 128 (group E); one channel; standardised
Optimiser	AdamW, peak learning rate 0.002, weight decay 0.0005, batch size 128
Schedule	2-epoch warm-up + cosine decay; up to 100 epochs (60 at 64/128 px)
Early stopping	Patience 20 epochs on validation macro-AUROC (15 at 64/128 px)
Regularisation	Affine + intensity augmentation; dropout 0.1; EMA 0.999; mixed precision
Initialisation	Random for every model (no ImageNet pre-training)
Thresholds / calibration	Per-class max-F1 on validation; Platt scaling fitted on validation
Seeds	0–4 (0–2 at 64/128 px); seed i gives the same data order for every configuration

B. Evaluation Metrics

The primary metric is macro-AUROC [45], the mean of the 14 per-class AUROCs. We also report mean average precision (mAP) and, at the validation-tuned thresholds, macro-F1, Hamming loss and exact-match (subset) accuracy, always next to the all-negative baselines. Calibration is measured with the pooled expected calibration error over all image-label pairs,

$$ECE = \sum_{b=1}^{15} \frac{|B_b|}{N} |\bar{p}(B_b) - \bar{y}(B_b)|, \quad (10)$$

and, because about 95% of all pairs are easy negatives that keep the pooled value small, with the class-wise ECE [25], the mean of the per-class values,



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

$$ECE_{cw} = \frac{1}{K} \sum_{k=1}^K \sum_{b=1}^{15} \frac{|B_{k,b}|}{N} |\bar{p}(B_{k,b}) - \bar{y}(B_{k,b})|. \quad (11)$$

C. Statistical Analysis

Each configuration is summarised by its mean $\pm$ standard deviation (SD) over seeds. The factorial ablation is analysed with a three-way ANOVA with all interactions and the seed as a blocking factor. Pairs of configurations are compared in three ways: (i) a paired t-test over seeds (training variability); (ii) the seed-averaged probabilities of each configuration form an ensemble, and two ensembles are compared on the same test images with a paired bootstrap of macro-AUROC (10,000 resamples [46]); (iii) per-class DeLong tests [22], [47] between the ensembles, Holm-adjusted for the 14 classes [23]. Exact p-values are reported.

VI. RESULTS

Fig. 4 summarises every effect measured in this study on one scale; the following subsections give the details.

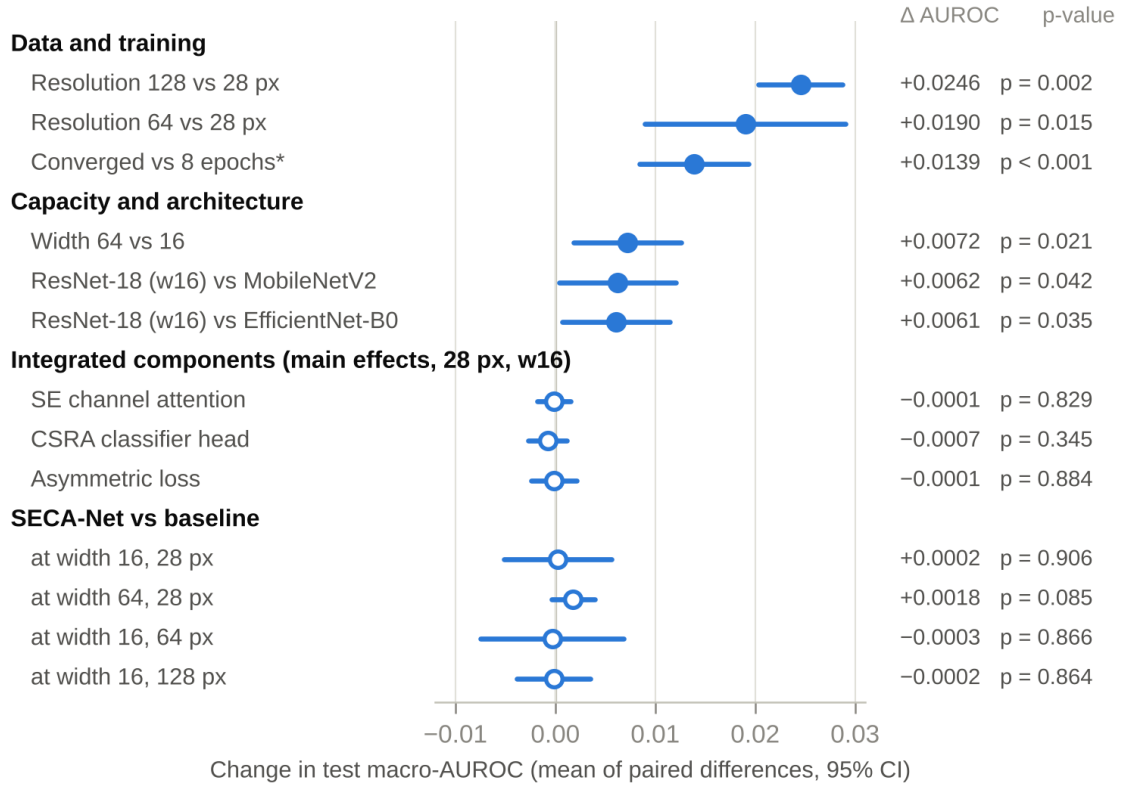


Fig. 4 Change in test macro-AUROC for each factor (mean of seed-paired differences with 95% confidence interval; filled markers: interval excludes zero). Five seed pairs per comparison, three at 64 and 128 pixels. *The eight-epoch preliminary models (three seeds, CPU) are compared unpaired with the converged models.

A. Training to Convergence

With early stopping the baseline reached its best validation AUROC after 14–24 epochs and SECA-Net after 19–26 epochs (Fig. 5). The baseline's converged test macro-AUROC of 0.7892 ± 0.0041 is 0.0139 higher than



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

the 0.7753 ± 0.0022 of the eight-epoch preliminary models (three seeds; Welch $p < 0.001$). The apparent advantage of SECA-Net at eight epochs (0.7779 versus 0.7753) did not survive convergence: 0.7894 ± 0.0011 versus 0.7892 ± 0.0041 (seed-paired difference $+0.0002$, $p = 0.91$). Attention and the asymmetric loss mainly changed how fast the models learned, not where they converged. The models also differed in what happened after their peak: the width-64 network over-fitted much faster than the width-16 network (Fig. 5), so a fixed schedule ranks models by when they peak rather than by how well they can perform.

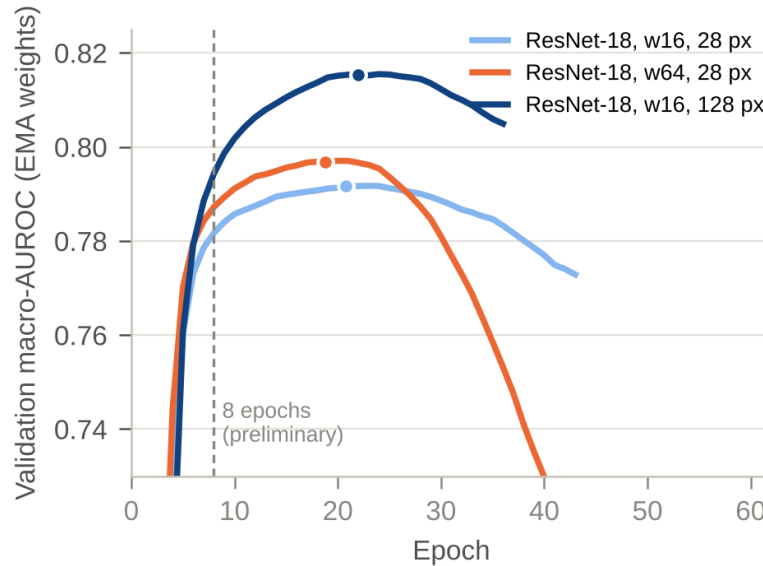
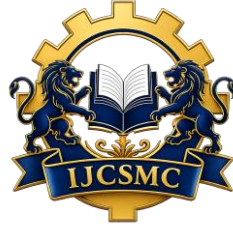


Fig. 5 Validation macro-AUROC per epoch (EMA weights, mean over seeds); dots mark the mean best epoch. The dashed line marks the eight-epoch schedule of the preliminary experiments.

B. Factorial Ablation of the Three Components

TABLE IV
FACTORIAL ABLATION AND LOSS STUDY AT 28 PIXELS, WIDTH 16 (TEST SET, MEAN $\pm$ SD OVER 5 SEEDS; ENS. = 5-SEED ENSEMBLE)

Backbone	Head	Loss	Macro-AUROC	mAP	Macro-F1	cw-ECE	Ens. AUROC
ResNet	GAP	BCE	0.7892 ± 0.0041	0.1786 ± 0.0028	0.2322 ± 0.0040	0.006	0.7978
ResNet	GAP	ASL	0.7911 ± 0.0020	0.1790 ± 0.0014	0.2340 ± 0.0012	0.309	0.7996
ResNet	CSRA	BCE	0.7897 ± 0.0013	0.1790 ± 0.0012	0.2326 ± 0.0016	0.007	0.7983
ResNet	CSRA	ASL	0.7894 ± 0.0023	0.1785 ± 0.0010	0.2330 ± 0.0015	0.309	0.7988
SE-ResNet	GAP	BCE	0.7914 ± 0.0013	0.1796 ± 0.0016	0.2364 ± 0.0012	0.006	0.7997
SE-ResNet	GAP	ASL	0.7890 ± 0.0011	0.1789 ± 0.0012	0.2334 ± 0.0027	0.310	0.7969
SE-ResNet	CSRA	BCE	0.7891 ± 0.0025	0.1784 ± 0.0022	0.2346 ± 0.0026	0.006	0.7992
SE-ResNet	CSRA	ASL	0.7894 ± 0.0011	0.1792 ± 0.0011	0.2344 ± 0.0015	0.313	0.7973
SE-ResNet	CSRA	W-BCE	0.7900 ± 0.0020	0.1730 ± 0.0024	0.2287 ± 0.0032	0.290	0.7972
SE-ResNet	CSRA	Focal	0.7907 ± 0.0019	0.1781 ± 0.0028	0.2340 ± 0.0053	0.097	0.7981



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

Table IV lists the eight cells of the ablation and the loss study. All ten configurations lie within 0.0025 macro-AUROC of each other (0.7890 to 0.7914), less than the seed SD of the baseline (0.0041). Table V gives the blocked ANOVA. No main effect and no interaction were significant for macro-AUROC or mAP; the largest main effect was that of the CSRA head (-0.0007). The highest mean was obtained by SE-ResNet + GAP + BCE (0.7914 ± 0.0013), but its advantage over the baseline ($+0.0022$, $p = 0.32$) is within noise. The integrated network itself (SE + CSRA + ASL) scored 0.7894 ± 0.0011 .

TABLE V
THREE-WAY ANOVA OF THE FACTORIAL ABLATION WITH SEED AS BLOCK (40 RUNS; DF = 1 PER TERM)

Term	Δ AUROC	F (AUROC)	p (AUROC)	F (mAP)	p (mAP)
SE	-0.0001	0.04	0.844	0.26	0.615
CSRA	-0.0007	1.14	0.294	0.28	0.603
ASL	-0.0001	0.03	0.857	0.00	0.967
SE $\times$ CSRA	–	0.06	0.815	0.18	0.674
SE $\times$ ASL	–	1.92	0.177	0.02	0.898
CSRA $\times$ ASL	–	0.03	0.858	0.12	0.734
SE $\times$ CSRA $\times$ ASL	–	3.24	0.083	1.29	0.265

C. Loss Functions and Calibration

The four losses gave the same discrimination (Table IV), but very different probabilities (Fig. 6). With BCE the class-wise ECE was 0.0063; focal loss raised it to 0.097, weighted BCE to 0.290 and ASL to 0.313. Weighted BCE and ASL over-predict in every probability bin, so their pooled and class-wise ECE both equal the mean over-prediction, which explains why the two measures coincide for them. This is expected: both losses down-weight negatives and therefore move all probabilities upwards. As a result, the standard MedMNIST accuracy at a 0.5 threshold fell from 0.948 (BCE) to 0.877 (ASL) and 0.735 (weighted BCE), below the all-negative baseline of 0.947, although ranking quality was unchanged. Weighted BCE also lowered mAP (-0.0054 , $p < 0.001$). Per-class Platt scaling fitted on the validation set reduced the class-wise ECE of every loss to 0.004–0.006. A small pooled ECE after re-calibration mainly reflects the 95% of image-label pairs that are easy negatives; it says little about the rare positives, which is why the class-wise ECE is reported.

Miscalibration also matters when probabilities are combined. Scoring each image for the presence of any finding with $1 - \prod_k (1 - p_k)$ gave a normal-versus-abnormal AUROC of 0.753 for the SECA-Net ensemble and 0.764 for the BCE baseline (DeLong $p < 0.001$), although their per-class AUROCs were nearly identical: the combination rule depends on the probability scale, which ASL distorts.

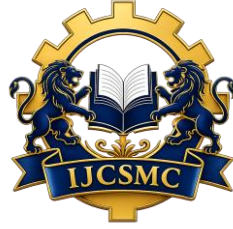
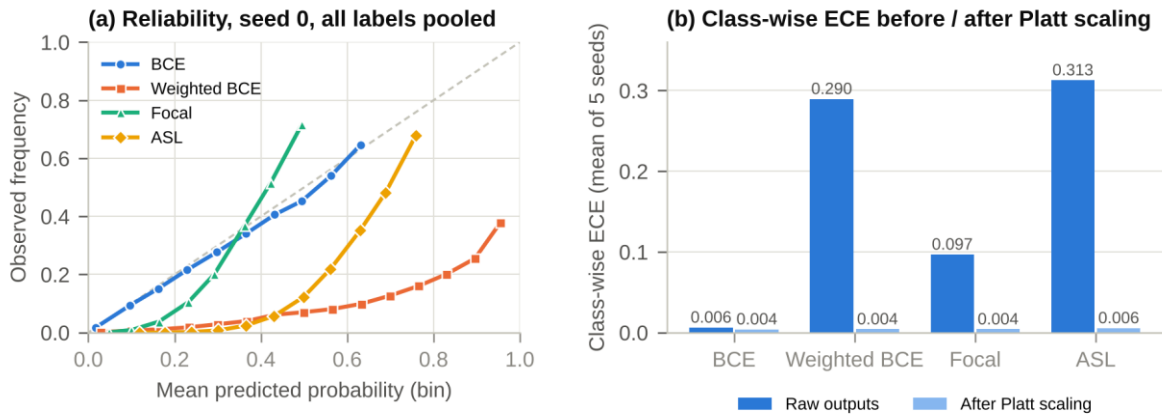


Fig. 6 Calibration of SE-ResNet + CSRA with four losses at 28 pixels. (a) Reliability diagram over all image-label pairs (seed 0; bins with at least 200 pairs). (b) Class-wise ECE before and after per-class Platt scaling fitted on the validation set.

D. Width and Lightweight Reference Models

TABLE VI
WIDTH, REFERENCE MODELS AND RESOLUTION (TEST SET, MEAN $\pm$ SD OVER SEEDS; 5 SEEDS AT 28 PX, 3 AT 64/128 PX). MS/IMG: T4
INFERENCE TIME; EPOCHS: MEAN BEST EPOCH

Model	Input	Params	MACs	Macro-AUROC	mAP	Ens. AUROC	ms/img	Epochs
Plain CNN	28	0.30 M	9 M	0.7821 ± 0.0024	0.1718 ± 0.0017	0.7919	0.02	20
MobileNetV2	28	2.24 M	22 M	0.7830 ± 0.0016	0.1753 ± 0.0015	0.7922	0.18	27
EfficientNet-B0	28	4.02 M	29 M	0.7831 ± 0.0015	0.1730 ± 0.0015	0.7912	0.19	34
ResNet-18, w16	28	0.70 M	29 M	0.7892 ± 0.0041	0.1786 ± 0.0028	0.7978	0.04	21
SECA-Net, w16	28	0.71 M	29 M	0.7894 ± 0.0011	0.1792 ± 0.0011	0.7973	0.04	22
ResNet-18, w64	28	11.17 M	456 M	0.7964 ± 0.0009	0.1887 ± 0.0025	0.8029	0.22	19
SECA-Net, w64	28	11.27 M	456 M	0.7982 ± 0.0026	0.1884 ± 0.0016	0.8048	0.22	21
ResNet-18, w16	64	0.70 M	139 M	0.8095 ± 0.0022	0.2013 ± 0.0024	0.8180	0.10	22
SECA-Net, w16	64	0.71 M	139 M	0.8092 ± 0.0006	0.2044 ± 0.0018	0.8176	0.11	24
ResNet-18, w16	128	0.70 M	139 M	0.8150 ± 0.0007	0.2112 ± 0.0022	0.8222	0.11	22
SECA-Net, w16	128	0.71 M	139 M	0.8149 ± 0.0022	0.2131 ± 0.0016	0.8230	0.13	24

Table VI compares model sizes. Widening the ResNet from 16 to 64 base channels (0.70 M to 11.2 M parameters, 16 $\times$ the multiply-accumulates) raised macro-AUROC by 0.0072 ($p = 0.021$, 5 of 5 seeds) and mAP by 0.0101. At width 64 SECA-Net differed from the baseline by +0.0018 in macro-AUROC ($p = 0.085$, 4 of 5 seeds) and -0.0003 in mAP; the width $\times$ attention interaction was not significant ($F = 0.46$, $p = 0.51$). The trend is consistent with attention needing more capacity, but the evidence is weak and this study cannot say which of the three components would be responsible.

All three reference models scored lower than the width-16 ResNet: MobileNetV2 (0.7830 ± 0.0016 , $p = 0.042$ for the difference), EfficientNet-B0 (0.7831 ± 0.0015 , $p = 0.035$) and the plain CNN (0.7821 ± 0.0024 , $p = 0.008$), although MobileNetV2 and EfficientNet-B0 have 2.2 M and 4.0 M parameters. Their depthwise designs were developed for ImageNet-sized inputs and pre-training; trained from scratch on 28-pixel images they were also slower per image on the T4 GPU.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

E. Input Resolution

Resolution was the largest effect in the study. The same 0.70 M-parameter ResNet reached 0.8095 ± 0.0022 at 64 pixels and 0.8150 ± 0.0007 at 128 pixels, gains of $+0.0190$ ($p = 0.015$) and $+0.0246$ ($p = 0.002$) over the same three seeds at 28 pixels (seed-paired); mAP rose from 0.179 to 0.211. Because the stem uses stride 2 at 128 pixels, the 64- and 128-pixel models have the same cost (139 M multiply-accumulates), so the gain from 64 to 128 pixels ($+0.0055$) came at no extra compute. SECA-Net again matched but did not beat the baseline (-0.0003 at 64 and -0.0002 at 128 pixels), although its CSRA head now attends over an 8×8 map. Fig. 7 shows that the gain is largest for hernia ($+0.057$), emphysema ($+0.052$), nodule ($+0.042$), mass ($+0.034$) and smallest for infiltration ($+0.012$), edema ($+0.009$), consolidation ($+0.007$): findings defined by small structures or fine texture benefit most from higher resolution, while large, diffuse findings are already visible at 28 pixels.

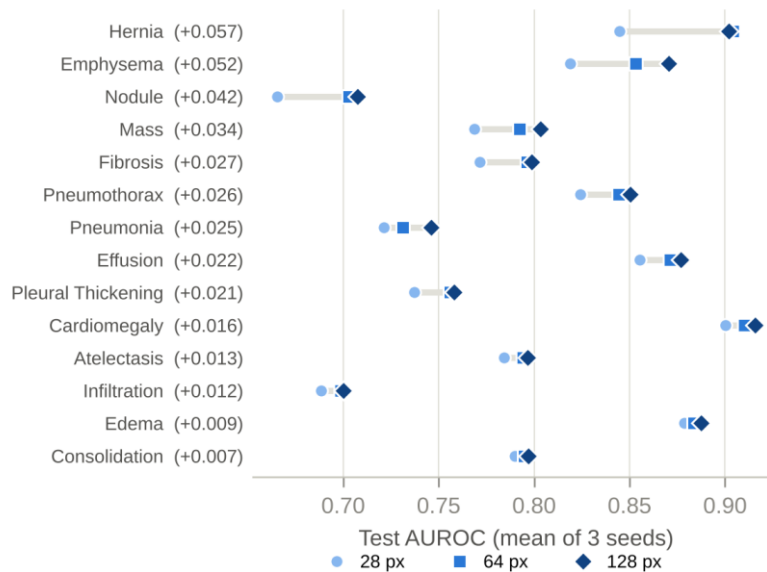


Fig. 7 Per-class test AUROC of the width-16 ResNet at 28, 64 and 128 pixels (mean of three seeds), sorted by the gain from 28 to 128 pixels (in brackets).

F. Seed Ensembles and Per-Class Tests

TABLE VII

COMPARISON OF SEED ENSEMBLES ON THE TEST SET (PAIRED BOOTSTRAP, 10,000 RESAMPLES). *CLASSES WITH A SIGNIFICANT DELONG DIFFERENCE AFTER HOLM CORRECTION

Comparison (A vs B)	Ens. AUROC A	Ens. AUROC B	Δ [95% CI]	p (boot.)	Classes*
SECA-Net vs ResNet, w16, 28 px	0.7973	0.7978	-0.0006 [-0.0024, +0.0012]	0.5224	0
SECA-Net vs ResNet, w64, 28 px	0.8048	0.8029	+0.0019 [+0.0001, +0.0037]	0.0358	0
SECA-Net vs ResNet, w16, 64 px	0.8176	0.8180	-0.0004 [-0.0029, +0.0019]	0.7792	0
SECA-Net vs ResNet, w16, 128 px	0.8230	0.8222	+0.0008 [-0.0010, +0.0026]	0.3650	0
ResNet w64 vs w16, 28 px	0.8029	0.7978	+0.0051 [+0.0032, +0.0070]	< 0.0002	2
SECA-Net w64 vs w16, 28 px	0.8048	0.7973	+0.0076 [+0.0054, +0.0100]	< 0.0002	4
SECA-Net vs MobileNetV2, 28 px	0.7973	0.7922	+0.0051 [+0.0033, +0.0070]	< 0.0002	1



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X
Impact Factor: 7.056

Averaging the seeds of a configuration raised macro-AUROC by about 0.009 (Table VII). At width 16 the SECA-Net and baseline ensembles differed by -0.0006 ($p = 0.52$), and no class was significant after Holm correction. At width 64 they differed by $+0.0019$ (95% CI $+0.0001$ to $+0.0037$, $p = 0.0358$), and no single class differed after Holm correction. The bootstrap treats the test images as the random sample and the trained models as fixed; together with the non-significant seed-level test, it indicates a small effect at width 64 that is not robust to training randomness. The width effect, in contrast, was significant at both levels.

TABLE VIII
PER-CLASS TEST AUROC OF THE SEED ENSEMBLES (28 px UNLESS STATED). NO CLASS DIFFERED SIGNIFICANTLY FROM THE PAIRED RESNET AFTER HOLM CORRECTION (DELONG)

Finding	Test %	ResNet w16	SECA w16	ResNet w64	SECA w64	ResNet 128	SECA 128
Atelectasis	10.79	0.790	0.789	0.793	0.793	0.802	0.803
Cardiomegaly	2.59	0.907	0.907	0.906	0.906	0.921	0.919
Effusion	12.28	0.860	0.860	0.863	0.864	0.881	0.881
Infiltration	17.55	0.692	0.692	0.694	0.697	0.704	0.706
Mass	5.05	0.778	0.777	0.788	0.793	0.812	0.816
Nodule	5.95	0.673	0.673	0.678	0.679	0.717	0.722
Pneumonia	1.08	0.728	0.736	0.736	0.738	0.753	0.755
Pneumothorax	4.85	0.834	0.835	0.838	0.839	0.857	0.856
Consolidation	4.27	0.794	0.795	0.796	0.798	0.801	0.801
Edema	1.84	0.884	0.884	0.885	0.886	0.892	0.890
Emphysema	2.27	0.831	0.832	0.834	0.832	0.880	0.879
Fibrosis	1.61	0.785	0.780	0.790	0.787	0.809	0.807
Pleural Thickening	3.27	0.745	0.744	0.749	0.751	0.764	0.767
Hernia	0.19	0.868	0.856	0.889	0.905	0.918	0.921
Mean		0.798	0.797	0.803	0.805	0.822	0.823

Table VIII gives the per-class AUROCs. At width 16 SECA-Net was higher than the baseline for 4 of the 14 findings, equal for 5 and lower for 5 (to three decimals), a pattern compatible with chance. The weakest findings were the same for every model: nodule, infiltration and pneumonia stayed below 0.74 at 28 pixels, while cardiomegaly, effusion, edema and hernia exceeded 0.85. Per-class differences between configurations should therefore be read together with the Holm-corrected tests, not class by class.

G. Decision-Level Metrics and Trivial Baselines

TABLE IX
DECISION-LEVEL METRICS AT VALIDATION-TUNED THRESHOLDS (TEST SET, MEAN OVER SEEDS) COMPARED WITH THE ALL-NEGATIVE PREDICTOR

Model	Exact match	ACC@0.5	Hamming	Macro-F1 (val)	Macro-F1 (test)	Recall	Specif.
All-negative predictor	0.5317	0.9474	0.0526	0	0	0	1
ResNet-18, w16, 28 px	0.3526	0.9480	0.1032	0.2517	0.2322	0.337	0.918
SECA-Net, w16, 28 px	0.3463	0.8772	0.1068	0.2482	0.2344	0.349	0.915
ResNet-18, w64, 28 px	0.3518	0.9479	0.0990	0.2638	0.2454	0.340	0.922
ResNet-18, w16, 128 px	0.4007	0.9481	0.0875	0.2849	0.2686	0.340	0.936

Table IX puts the threshold-based metrics in context. With per-class F1-optimal thresholds every model had a lower exact-match accuracy (for example 0.353) than the all-negative predictor (0.532) and a higher Hamming loss: F1-optimal thresholds deliberately trade many false positives on common negatives for sensitivity to rare positives. Exact match and per-label accuracy therefore do not measure useful performance on this dataset and should never be reported without the trivial baseline. Test macro-F1 (0.232) was consistently below validation



International Journal of Computer Science and Mobile Computing
A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19
(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X
Impact Factor: 7.056

macro-F1 (0.252), because the thresholds were tuned on the validation set; only the test values are performance estimates.

H. Attention Maps

Fig. 8 shows CSRA attention maps for confidently detected test images, next to Grad-CAM [48] maps of the GAP baseline. At 28 pixels the maps have 4×4 cells. For cardiomegaly and effusion, they mostly favour the lower and central thorax, which is plausible, but they often highlight the mediastinum or the region below the diaphragm rather than the lung fields, and the width-16 and width-64 models often disagree. Because ChestMNIST has no lesion annotations, these maps cannot be validated quantitatively; we show them only to illustrate what the head can represent, and we make no claim that the attention localises disease or improves interpretability.

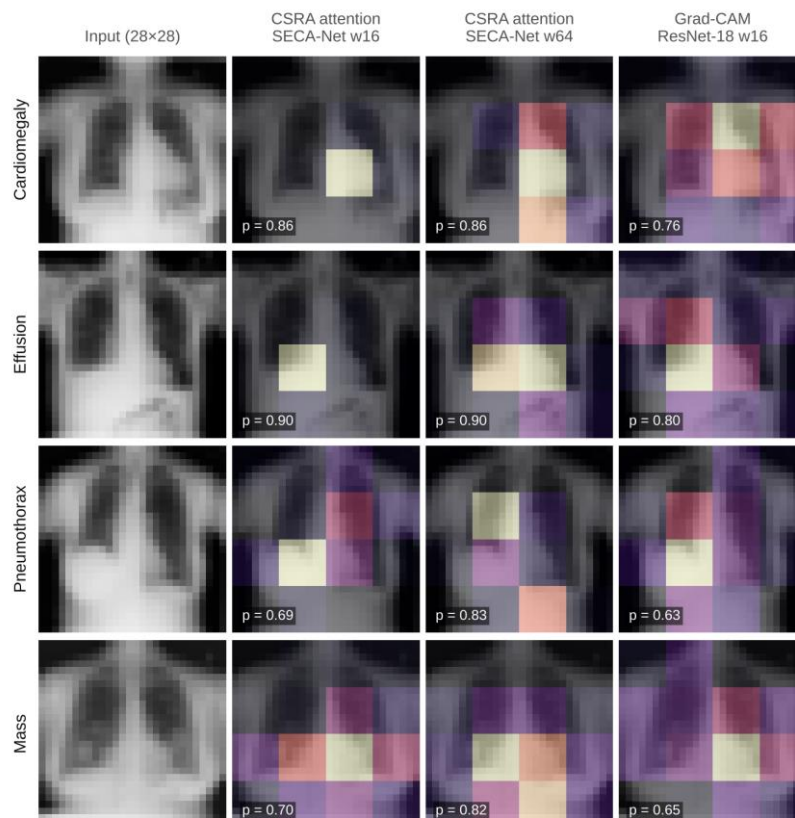


Fig. 8 CSRA attention (T = 1 head) of SECA-Net at widths 16 and 64, and Grad-CAM of the width-16 ResNet baseline, for the most confident true-positive test images of four findings; p is the predicted probability. Qualitative illustration only.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

I. Comparison with Published ChestMNIST Results

TABLE X
PUBLISHED MEDMNIST V2 BENCHMARK RESULTS AND THIS STUDY (SAME OFFICIAL TEST SPLIT)

Method	Input	Params	Test macro-AUROC
ResNet-18 [15]	28	11.2 M	0.768
ResNet-18 [15]	224	11.2 M	0.773
ResNet-50 [15]	28	23.5 M	0.769
Google AutoML Vision [15]	28	–	0.778
ResNet-18, w16 (ours, 5 seeds)	28	0.70 M	0.7892 $\pm$ 0.0041
ResNet-18, w64 (ours, 5 seeds)	28	11.2 M	0.7964 $\pm$ 0.0009
ResNet-18, w16 (ours, 3 seeds)	128	0.70 M	0.8150 $\pm$ 0.0007

Table X lists the MedMNIST v2 benchmark [15]. The width-16 ResNet at 28 pixels (0.789) exceeds the published 28-pixel ResNet-18 (0.768) with 16 times fewer parameters, and the 128-pixel model reaches 0.815. We did not re-run the benchmark protocol, so we attribute this gap to differences in training (augmentation, weight averaging and model selection by validation AUROC) and not to any architectural component; the benchmark models were also selected and reported from single runs.

VII. DISCUSSION

A. What Mattered

On ChestMNIST, the inexpensive choices that mattered were, in order of size: input resolution, training to convergence and network width. The three integrated components, SE channel attention, the CSRA head and ASL, did not change discrimination at width 16, and their combination gave no reliable gain at width 64 or at higher resolution. Three explanations are plausible. First, at 28 pixels the final feature map has 4×4 cells, so CSRA can do little more than global pooling; at 128 pixels it has 8×8 cells but still gave no gain. Second, ASL and focal loss mainly re-weight easy negatives; with AUROC as the target and a well-tuned BCE baseline, re-weighting changes the scale of the probabilities more than their ranking. Third, a width-16 network may lack the capacity to exploit channel re-weighting, which is consistent with the small, non-robust gain at width 64.

B. Reproducibility and Effect Sizes

Training the same seed twice on the same GPU type changed macro-AUROC by up to 0.0026 (3 repeated runs), because cuDNN kernel selection and mixed-precision arithmetic are not deterministic. Seed SDs were 0.001–0.004, and the five seeds of the baseline alone spanned 0.0102. Single-run differences below about 0.01 macro-AUROC should therefore be confirmed with several seeds before they are interpreted.

A null result is only informative if the study could have detected a meaningful effect. With five seeds, the 95% confidence intervals of the three main effects all lie within ± 0.0027 macro-AUROC (Fig. 4), and a seed-paired t-test with 80% power would have detected a difference of about 0.0026 for a main effect and 0.0072 for SECA-Net versus the baseline. Effects of the size often reported for attention modules (0.01 or more) are therefore ruled out in this setting; effects of a few thousandths cannot be excluded.

C. Relation to Previous Reports

Attention-based ChestX-ray14 models such as category-wise residual attention [13] and cross-attention networks [14] reported gains over their baselines with ImageNet-pre-trained backbones at 224 pixels or more, and CSRA [19] and ASL [20] were developed for natural images with large pre-trained models. Our results do not contradict those reports; they show that the gains do not transfer automatically to small networks trained from scratch on low-resolution radiographs, and that a short training schedule can produce apparent gains that



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

vanish at convergence. The finding that resolution matters more than architecture agrees with the MedMNIST+ study [21] and with the general observation that benchmark improvements often shrink under controlled, multi-seed evaluation [17].

D. Recommendations

From these results we suggest the following practice for lightweight multi-label CXR studies:

- train every compared model to convergence with model selection on the validation set, not for a fixed short schedule;
- report the mean and SD over at least three, preferably five, seeds, and test differences with seed-paired and ensemble-level tests;
- separate width and resolution from architectural changes, since both have larger effects than the modules usually studied;
- report class-wise calibration and re-calibrate the outputs of re-weighting losses before using their probabilities, and never report exact match or per-label accuracy without the all-negative baseline.

E. Clinical Relevance

The study is a methodological benchmark, not a clinical evaluation. The images are down-sampled to at most 128 pixels, the labels were mined from reports and are noisy [16], the data come from one institution and no external test set was used [9]. Macro-AUROC values around 0.8 are far from what would be needed for diagnostic use, and the probabilities produced by ASL and weighted BCE would be misleading without re-calibration. The results should be read as guidance on experimental design for lightweight models, not as evidence of clinical utility.

F. Limitations

The study uses one dataset, low resolutions and training from scratch; conclusions may differ for pre-trained backbones, full-resolution images or other label sets. The width-64 and resolution studies compare only the baseline with the full integrated network, so the contribution of each component at width 64 is unknown, and the resolution study used three seeds. Hyper-parameters (λ , the ASL margins, the learning rate) were taken from the original papers and not tuned per configuration, which could disadvantage some components. Attention maps were inspected only qualitatively. Finally, transformers and state-space models were discussed but not trained.

VIII. CONCLUSION

We integrated SE channel attention, the CSRA classifier and the asymmetric loss into a lightweight residual network and evaluated them on ChestMNIST in a controlled study of 87 runs. None of the components improved macro-AUROC or mAP; ASL and weighted BCE degraded calibration, which Platt scaling repaired. Resolution, training to convergence and width produced gains more than ten times larger than any component effect, and a 0.70 M-parameter ResNet trained to convergence exceeded the published ChestMNIST benchmark. Future work should test the components at width 64 one at a time, use pre-trained backbones and full-resolution images, and validate on an external dataset. Code, weights and all results are available at <https://github.com/Sakawat-hossain/lightweight-cxr-ablation>.

DATA AND CODE AVAILABILITY

ChestMNIST and its 64- and 128-pixel MedMNIST+ versions are available from the official MedMNIST record (Zenodo, record 10519652) under the licence stated by MedMNIST. The training and analysis code, the experiment runner used on Google Colab and Kaggle, the trained weights of the main models, the per-run



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

results files (configuration, software versions, training history, validation and test metrics with bootstrap intervals) and the test-set predictions of all 87 runs are released at <https://github.com/Sakawat-hossain/lightweight-cxr-ablation>. Every number in this paper is computed from these files by the released analysis scripts.

REFERENCES

- [1]. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017.
- [2]. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097-2106.
- [3]. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 590-597.
- [4]. A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, art. 317, 2019.
- [5]. A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, art. 101797, 2020.
- [6]. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.
- [7]. P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, no. 11, e1002686, 2018.
- [8]. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, art. 195, 2019.
- [9]. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, e1002683, 2018.
- [10]. J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction," in *Proc. Medical Imaging with Deep Learning (MIDL)*, PMLR, vol. 121, 2020, pp. 136-155.
- [11]. L. Yao, E. Poblens, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.
- [12]. B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, "Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2292-2302, 2020.
- [13]. Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, pp. 259-266, 2020.
- [14]. C. Ma, H. Wang, and S. C. H. Hoi, "Multi-label thoracic disease image classification with cross-attention networks," in *Medical Image Computing and Computer Assisted Intervention (MICCAI 2019)*, Springer, 2019, pp. 730-738.
- [15]. 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," *Scientific Data*, vol. 10, art. 41, 2023.
- [16]. L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proc. ACM Conf. Health, Inference, and Learning (CHIL)*, 2020, pp. 151-159, doi: 10.1145/3368555.3384468.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

- [17].X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, et al., "Accounting for variance in machine learning benchmarks," in *Proc. Machine Learning and Systems (MLSys)*, vol. 3, 2021.
- [18].J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132-7141.
- [19].K. Zhu and J. Wu, "Residual attention: A simple but effective method for multi-label recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 184-193.
- [20].T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 82-91.
- [21].S. Doerrich, F. Di Salvo, J. Brockmann, and C. Ledig, "Rethinking model prototyping through the MedMNIST+ dataset collection," *Scientific Reports*, vol. 15, art. 7669, 2025, doi: 10.1038/s41598-025-92156-9.
- [22].E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837-845, 1988.
- [23].S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65-70, 1979.
- [24].C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321-1330.
- [25].M. Kull, M. Perello-Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [26].G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700-4708.
- [27].I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Scientific Reports*, vol. 9, art. 6381, 2019.
- [28].T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980-2988.
- [29].G. Holste, S. Wang, Z. Jiang, T. C. Shen, G. Shih, R. M. Summers, Y. Peng, and Z. Wang, "Long-tailed classification of thorax diseases on chest X-ray: A new benchmark study," in *Data Augmentation, Labelling, and Imperfections (DALI 2022), MICCAI Workshop, Lecture Notes in Computer Science*, vol. 13567, Springer, 2022, doi: 10.1007/978-3-031-17027-0_3.
- [30].G. Holste, Y. Zhou, S. Wang, A. Jaiswal, M. Lin, S. Zhuge, et al., "Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge," *Medical Image Analysis*, vol. 97, art. 103224, 2024.
- [31].A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- [32].Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012-10022.
- [33].S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban, "SwinCheX: Multi-label classification on chest X-ray images with transformers," *arXiv preprint arXiv:2206.04246*, 2022.
- [34].A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [35].L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. Int. Conf. Machine Learning (ICML)*, PMLR, vol. 235, 2024.
- [36].Y. Yue and Z. Li, "MedMamba: Vision Mamba for medical image classification," *arXiv preprint arXiv:2403.03849*, 2024.
- [37].E. Tiu, E. Talus, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, pp. 1399-1406, 2022.
- [38].M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510-4520.
- [39].M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105-6114.



International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

Md Sakawat Hossain *et al*, Volume 15, Issue 10, October 2026, pg. 1-19

(An Open Accessible, ISI Indexed, Fully Refereed and Peer Reviewed Journal)

ISSN: 2320-088X

Impact Factor: 7.056

- [40].K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.
- [41].I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.
- [42].I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.
- [43].B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838-855, 1992.
- [44].J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61-74.
- [45].T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
- [46].B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.
- [47].X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389-1393, 2014.
- [48].R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618-626.