Localization-Aware Chest X-ray Classification via Segmentation and Gradient-based Attention

Anonymous Authors

I. Evaluation metrics

In the classification tasks, instances where the lesion type is "No Finding" are designated as the negative class. For evaluating the classification results, we utilize the performance metrics provided by scikit-learn, including Accuracy, Precision, Recall, and F1-Score. Since the datasets have been balanced through cleaning to ensure representation across classes, both weighted and macro variants of the metrics are reported. Given the unique characteristics of medical data, we additionally incorporate sensitivity (recall), specificity, positive predictive value (PPV), false discovery rate (FDR), false omission rate (FOR), Youden's index (YI) and Discriminant power (DP) as evaluation metrics, which are formally defined as follows:

$$Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{FP + TN},$$

$$PPV = \frac{TP}{TP + FP},$$

$$FDR = \frac{FP}{FP + TP},$$

$$FOR = \frac{FN}{FN + TN}.$$

$$\text{YI} = \text{Sensitivity} + \text{Specificity} - 1$$

$$\text{DP} = \frac{\sqrt{3}}{\pi} \cdot \left(\log\left(\frac{\text{sensitivity}}{1 - \text{sensitivity}} \right) + \log\left(\frac{\text{specificity}}{1 - \text{specificity}} \right) \right)$$

II. Youden's index and Discriminant Power

In the following, we focus on two key evaluation metrics: *Youden's index*, which balances sensitivity and specificity, and *Discriminant Power* (DP), which quantifies a model's ability to discriminate between classes.

Results on the Ottawa Dataset. Table I summarizes the performance of our proposed loss compared to standard cross-entropy loss across three backbone architectures (PVT, VGG16, and ResNet50) and multiple binary classification tasks derived from the Ottawa dataset.

Across all binary tasks, the proposed loss consistently improves both Youden's index and DP compared to standard cross-entropy:

- **Effusion vs. No Finding:** PVT benefits most from the attention-guided loss, with Youden's index increasing from 0.257 to 0.449 and DP from 0.253 to 0.464. ResNet50 and VGG16 also show notable gains in DP and index, underscoring the robustness of our loss across architectures.
- **Pneumothorax vs. No Finding:** Improvements are again observed across all models, particularly for VGG16, where Youden's index improves from 0.267 to 0.370 and DP from 0.292 to 0.392, suggesting enhanced sensitivity-specificity balance for rare, subtle lesions.
- **Subcutaneous Emphysema vs. No Finding:** All three models show consistent gains with the proposed loss. PVT achieves the highest Youden's index (0.575) and DP (0.628), indicating better localization and classification in anatomically diffuse pathologies.
- **Multi-class classification (No Finding, Effusion, Pneumothorax):** While this setup presents greater classification complexity, the proposed loss still improves discriminative performance. Notably, DP improves from 0.212 to 0.368 for PVT and from 0.223 to 0.405 for VGG16, showing its effectiveness even in multi-class contexts.

Results on the NIH Dataset. Table II presents a comparison of our proposed attention-guided loss function with the standard cross-entropy loss across three backbone architectures (PVT, VGG16, and ResNet50) and multiple classification tasks on the NIH ChestX-ray14 dataset.

Overall, we observe that incorporating anatomical priors through our proposed loss leads to consistent improvements in both Youden's index and DP across binary and multi-class classification tasks:

- **Effusion vs. No Finding:** The PVT backbone shows the most notable gain in Youden's index, increasing from 0.634 to 0.709, and in DP, from 0.747 to 0.871. While VGG16 and ResNet50 both already perform well under cross-entropy, the proposed loss maintains or modestly improves Youden's index and sustains high discriminability.
- **Pneumothorax vs. No Finding:** The attention-aware loss offers notable improvements across all models. For instance, PVT's Youden's index rises from 0.542 to 0.608, and DP from 0.584 to 0.677. ResNet50, which has the lowest baseline, benefits significantly—its Youden's in-

dex increases from 0.453 to 0.576, and DP from 0.468 to 0.630—demonstrating that the proposed loss can mitigate performance limitations in weaker architectures.

- **Multi-class classification (No Finding, Effusion, Pneumothorax):** Despite the increased complexity of the multi-class task, our loss yields improvements or comparable performance across all backbones. For PVT, Youden's index improves from 0.477 to 0.534 and DP from 0.525 to 0.603. VGG16 also benefits slightly, whereas ResNet50 shows marginal change, highlighting the robustness of the attention-guided supervision in more challenging scenarios.

Taken together, these results confirm that guiding model attention toward medically relevant anatomical regions improves both the sensitivity-specificity tradeoff and overall classification discriminability, even on a large-scale and heterogeneous dataset like NIH ChestX-ray14.

TABLE I
RESULTS OF COMPARISON FOR OTTAWA DATASET (CROSS-ENTROPY/PROPOSED LOSS) OVER THREE MODELS, THE EXPERIMENTAL SETTING USED FOR ATTENTION LOSS IS THE BEST-PERFORMING (F1 SCORE) CONFIGURATION FROM ALL SETTINGS OF EACH MODEL, DP REPRESENTS DISCRIMINANT POWER.

Ottawa: No Finding vs Effusion

Metric	PVT	VGG16	ResNet50
Sensitivity	0.654/0.705	0.782/0.692	0.744/0.692
Specificity	0.603/0.744	0.564/0.718	0.564/0.718
Youden's index	0.257/0.449	0.410/0.384	0.308/0.410
DP	0.253/0.464	0.367/0.418	0.317/0.418

Ottawa: No Finding vs Pneumothorax

Metric	PVT	VGG16	ResNet50
Sensitivity	0.517/0.638	0.784/0.784	0.707/0.672
Specificity	0.612/0.647	0.483/0.586	0.397/0.612
Youden's index	0.129/0.285	0.267/0.370	0.104/0.284
DP	0.125/0.281	0.292/0.392	0.111/0.281

Ottawa: No Finding vs Subcutaneous emphysema

Metric	PVT	VGG16	ResNet50
Sensitivity	0.697/0.780	0.727/0.735	0.621/0.689
Specificity	0.811/0.795	0.773/0.780	0.712/0.712
Youden's index	0.508/0.575	0.500/0.515	0.333/0.401
DP	0.548/0.628	0.528/0.547	0.335/0.407

Ottawa: No finding, Effusion, Pneumothorax

Metric	PVT	VGG16	ResNet50
Sensitivity	0.457/0.551	0.474/0.543	0.496/0.513
Specificity	0.729/0.776	0.737/0.771	0.748/0.756
Youden's index	0.186/0.327	0.212/0.314	0.244/0.269
DP	0.212/0.368	0.223/0.405	0.263/0.284

III. ATTENTION LOSS: ABLATION AND SENSITIVITY STUDY

A. Ottawa dataset

The experimental results presented in Tables III through VI provide a detailed evaluation of the proposed attention-augmented loss across multiple classification tasks using the Ottawa dataset. The central aim of this ablation study is to assess whether incorporating the attention loss (, setting $\lambda \neq 1$) can yield better performance than the baseline cross-entropy loss ($\lambda = 1$). Moreover, these results serve as a sensitivity study on λ, which reveal whether classification performance is stable across different configurations of this parameter.

TABLE II
RESULTS OF COMPARISON FOR NIH DATASET (CROSS-ENTROPY/PROPOSED LOSS) OVER THREE MODELS, THE EXPERIMENTAL SETTING USED FOR ATTENTION LOSS IS THE BEST-PERFORMING (F1 SCORE) CONFIGURATION FROM ALL SETTINGS OF EACH MODEL, DP REPRESENTS DISCRIMINANT POWER.

NIH: No Finding vs Effusion

Metric	PVT	VGG16	ResNet50
Sensitivity	0.752/0.810	0.861/0.823	0.746/0.828
Specificity	0.882/0.899	0.807/0.853	0.872/0.824
Youden's index	0.634/0.709	0.668/0.676	0.618/0.652
DP	0.747/0.871	0.779/ 0.750	0.789/0.746

NIH: No Finding vs Pneumothorax

Metric	PVT	VGG16	ResNet50
Sensitivity	0.745/0.815	0.795/0.795	0.715/0.806
Specificity	0.797/0.793	0.736/0.788	0.738/0.770
Youden's index	0.542/0.608	0.531/0.583	0.453/0.576
DP	0.584/0.677	0.570/0.639	0.468/0.630

NIH: No finding, Effusion, Pneumothorax

Metric	PVT	VGG16	ResNet50
Sensitivity	0.651/0.689	0.674/0.683	0.631/0.632
Specificity	0.826/0.845	0.837/0.842	0.815/0.816
Youden's index	0.477/0.534	0.511/0.525	0.446/0.449
DP	0.525/0.603	0.569/0.590	0.498/0.492

Across all tasks, binary classification (Effusion, Pneumothorax, and Subcutaneous Emphysema vs. No Finding) and the multi-class setting, we observe consistent evidence that configurations using $\lambda \neq 1$ often outperform the baseline $\lambda = 1$, thereby validating the benefit of the attention loss mechanism.

a) Effusion vs. No Finding: For the PVT model, the best results are achieved with $\lambda = 0.5$, outperforming $\lambda = 1$ across multiple metrics including accuracy (0.692 vs. 0.628), F1-score (0.692 vs. 0.628), and AUC (0.752 vs. 0.743). Similarly, the VGG16 model achieves peak performance with the adaptive loss ($\lambda = *$), showing a marked improvement in AUC (0.760 vs. 0.726) and F1-score (0.692 vs. 0.669).

b) Pneumothorax vs. No Finding: In this setting, the PVT model attains its best overall accuracy and F1-score using the adaptive $\lambda = *$ configuration, improving from 0.564 (F1-score at $\lambda = 1$) to 0.642. For the VGG16 model, a fixed attention weight of $\lambda = 0.5$ leads to the best overall metrics, including the highest F1-score (0.681) and AUC (0.714), surpassing the baseline values.

c) Subcutaneous Emphysema vs. No Finding: The PVT model shows optimal performance at $\lambda = 0.25$, reaching the

highest F1-score (0.757) and AUC (0.831), both higher than those at $\lambda = 1$. VGG16 again benefits from the adaptive setting ($\lambda = *$), particularly in PPV (0.847 vs. 0.762) and specificity (0.886 vs. 0.773), demonstrating stronger discriminative ability.

d) Multi-Class Classification: In the multi-class setting, the adaptive loss ($\lambda = *$) for the PVT model results in the best overall accuracy and F1-score (0.496 and 0.488, respectively), outperforming the baseline by a noticeable margin. This further confirms the flexibility and robustness of the proposed loss formulation.

The consistent pattern across datasets and models reveals that the attention-enhanced loss function is not only beneficial but also stable across different configurations. In nearly all classification settings, at least one $\lambda \neq 1$ configuration achieves better results than $\lambda = 1$, thereby empirically validating the core hypothesis of this study.

Furthermore, the success of both fixed-weight ($\lambda = 0.25$, $0.5, 0.75$) and adaptive ($\lambda = *$) configurations suggests that the attention mechanism contributes to improved representation learning, likely by guiding the model to focus on more informative features during training. This is especially valuable in medical imaging tasks, where subtle patterns are often decisive.

By consistently outperforming the standard cross-entropy loss in multiple classification tasks and across various deep learning architectures, the results advocate for the broader adoption of such attention-augmented losses in settings where model interpretability and discriminative accuracy are critical.

B. NIH dataset

The experimental results presented in Tables VII through IX provide a detailed evaluation of the proposed attention-augmented loss across multiple classification tasks using the NIH dataset.

a) Effusion vs No Finding: The PVT model consistently outperformed both VGG16 and ResNet50 across most metrics when using the adaptive custom loss ($\lambda = *$), achieving the highest accuracy, precision, F1-score, AUC, specificity, PPV, and lowest FDR. This indicates the model's strong ability to discriminate between the two classes when guided by the adaptive loss. Notably, PVT's performance peaked with $\lambda = *$ for almost every metric except sensitivity, which was highest at $\lambda = 0.5$.

VGG16, by contrast, achieved its best performance under the traditional cross-entropy loss ($\lambda = 1$), especially excelling in recall-oriented metrics such as sensitivity and false omission rate (FOR). This suggests that VGG16 may be more effective in settings where minimizing false negatives is prioritized.

ResNet50 performed moderately, with its best sensitivity and FOR observed at $\lambda = 0.25$, but it lagged behind PVT in most other metrics, indicating that while some improvements were seen with reduced λ, the architecture was less responsive to the adaptive loss overall.

Pneumothorax vs No Finding: In this task, the PVT model achieved the best overall performance with $\lambda = 0.75$, slightly outperforming $\lambda = *$ in key metrics such as accuracy, precision, F1-score, AUC, PPV, and specificity. This suggests that a moderately reduced λ—rather than the fully adaptive variant—may be optimal for PVT in detecting pneumothorax, potentially due to differences in class separability or noise distribution compared to the effusion task.

VGG16 again showed strong sensitivity at both $\lambda = 1$ and $\lambda = *$, maintaining a high FOR and a decent AUC. It remained robust across varying λ values, although adaptive loss did not consistently improve its performance.

ResNet50 achieved its best sensitivity and FOR using the adaptive loss ($\lambda = *$), but its overall AUC and specificity remained lower compared to the other models. As in the effusion task, ResNet50 appears to benefit marginally from adaptive loss but still underperforms relative to PVT.

Multi-Class Classification: In the more challenging multi-class setting, the PVT model again emerged as the top performer, with $\lambda = *$ producing the highest accuracy, precision, and F1-score, while AUC was maximized at $\lambda = 0.5$. This supports the conclusion that the adaptive loss is particularly well-suited for more complex classification scenarios involving multiple disease types.

VGG16, in contrast, saw its best results at $\lambda = 1$, and its performance degraded as λ decreased or was replaced by the adaptive version. This further supports the notion that VGG16 responds better to standard cross-entropy in both binary and multi-class tasks, particularly for recall-heavy metrics.

ResNet50 continued to trail the other models in this task. While it showed slight improvements in accuracy and F1 at $\lambda = *$, the gains were minimal, and its overall AUC was still lower than PVT's, indicating limited benefit from the adaptive loss.

Overall, results show that PVT benefits most significantly from the adaptive custom loss, especially in complex classification settings. It consistently achieved the best overall metrics across tasks, especially in precision, AUC, and F1-score. VGG16 performed best under cross-entropy loss, especially excelling in sensitivity and false omission rate. This model may be preferable in clinical scenarios where false negatives must be minimized. ResNet50 showed modest improvements in certain metrics (e.g., sensitivity and FOR) with lower λ or adaptive loss, but it was overall less competitive compared to PVT and VGG16.

While adaptive loss ($\lambda = *$) proved highly effective for PVT, the optimal λ varied across tasks and models, highlighting the need for task-specific tuning.

TABLE III

RESULTS WITH DIFFERENT λ VALUES (BINARY CLASSIFICATION - OTTAWA DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Effusion vs No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	0.628	0.660	**0.692**	0.654	0.628
	Prec	0.629	0.662	**0.693**	0.659	0.630
	F1	0.628	0.659	**0.692**	0.651	0.627
	AUC	*0.743*	0.723	**0.752**	0.678	0.683
	PPV	0.622	0.681	**0.703**	0.630	0.614
	Sensitivity	0.654	0.603	0.718	**0.744**	0.692
	Specificity	*0.603*	**0.718**	0.718	0.564	0.564
	FDR	0.378	0.319	**0.297**	0.370	0.386
	FOR	0.365	0.356	0.317	**0.313**	0.353
VGG16	Acc / Rec	*0.673*	0.667	0.673	0.654	**0.692**
	Prec	*0.682*	0.667	0.674	0.662	**0.692**
	F1	*0.669*	0.666	0.672	0.649	**0.692**
	AUC	0.726	0.735	0.745	0.745	**0.760**
	PPV	*0.642*	0.676	0.659	0.625	**0.697**
	Sensitivity	***0.782***	0.641	0.718	0.769	0.679
	Specificity	0.564	0.692	0.628	0.538	**0.705**
	FDR	*0.358*	0.324	0.341	0.375	**0.303**
	FOR	***0.279***	0.341	0.310	0.300	0.313
ResNet50	Acc / Rec	0.654	0.635	0.628	**0.673**	0.667
	Prec	0.659	0.638	0.630	**0.673**	0.669
	F1	0.651	0.632	0.627	**0.673**	0.665
	AUC	0.736	0.672	**0.739**	0.737	0.711
	PPV	0.630	0.615	0.616	**0.667**	0.648
	Sensitivity	**0.744**	0.718	0.679	0.692	0.731
	Specificity	0.564	0.551	0.577	**0.654**	0.603
	FDR	0.370	0.385	0.384	**0.333**	0.352
	FOR	0.313	0.338	0.357	0.320	**0.309**

TABLE IV

RESULTS WITH DIFFERENT λ VALUES (BINARY CLASSIFICATION - OTTAWA DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Pneumothorax vs No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	0.565	0.578	0.595	0.595	**0.642**
	Prec	0.565	0.578	0.595	0.596	**0.642**
	F1	0.564	0.576	0.594	0.594	**0.642**
	AUC	0.626	0.632	**0.654**	0.623	0.653
	PPV	0.571	0.587	0.602	0.587	**0.643**
	Sensitivity	0.517	0.526	0.560	**0.638**	**0.638**
	Specificity	*0.612*	0.629	0.629	0.552	**0.647**
	FDR	0.429	0.413	0.398	0.413	**0.357**
	FOR	0.441	0.430	0.411	0.396	**0.359**
VGG16	Acc / Rec	*0.634*	0.642	**0.681**	0.647	0.629
	Prec	*0.647*	0.643	**0.681**	0.648	0.629
	F1	*0.625*	0.642	**0.681**	0.646	0.629
	AUC	*0.707*	0.681	0.714	**0.740**	0.680
	PPV	*0.603*	0.636	**0.684**	0.663	0.627
	Sensitivity	***0.784***	0.664	0.672	0.595	0.638
	Specificity	0.483	0.621	0.690	**0.698**	0.621
	FDR	*0.397*	0.364	**0.316**	0.337	0.373
	FOR	***0.309***	0.351	0.322	0.367	0.368
ResNet50	Acc / Rec	0.552	0.547	0.504	**0.608**	0.556
	Prec	0.557	0.549	0.505	**0.609**	0.558
	F1	0.541	0.544	0.496	**0.607**	0.553
	AUC	0.591	0.569	0.504	**0.627**	0.583
	PPV	0.539	0.540	0.506	**0.619**	0.567
	Sensitivity	**0.707**	0.638	0.379	0.560	0.474
	Specificity	0.397	0.457	0.629	**0.655**	0.638
	FDR	0.461	0.460	0.494	**0.381**	0.433
	FOR	0.425	0.442	0.497	**0.402**	0.452

TABLE V

RESULTS WITH DIFFERENT λ VALUES (BINARY CLASSIFICATION - OTTAWA DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Subcutaneous emphysema vs No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	*0.754*	0.735	0.727	**0.758**	0.739
	Prec	*0.757*	0.740	0.727	**0.759**	0.740
	F1	*0.753*	0.733	0.727	**0.757**	0.738
	AUC	*0.828*	0.810	0.811	**0.831**	0.799
	PPV	***0.786***	0.777	0.721	0.779	0.756
	Sensitivity	0.697	0.659	**0.742**	0.720	0.705
	Specificity	***0.811***	**0.811**	0.712	0.795	0.773
	FDR	***0.214***	0.223	0.279	0.221	0.244
	FOR	0.272	0.296	0.266	**0.261**	0.277
VGG16	Acc / Rec	0.750	0.742	0.742	0.731	**0.758**
	Prec	0.751	0.742	0.751	0.731	**0.776**
	F1	0.750	0.742	0.740	0.731	**0.753**
	AUC	0.790	0.811	**0.829**	0.813	0.811
	PPV	0.762	0.739	0.796	0.729	**0.847**
	Sensitivity	*0.727*	**0.750**	0.652	0.735	0.629
	Specificity	0.773	0.735	0.833	0.727	**0.886**
	FDR	0.238	0.261	0.204	0.271	**0.153**
	FOR	*0.261*	**0.254**	0.295	0.267	0.295
ResNet50	Acc / Rec	0.667	**0.678**	0.621	0.652	0.652
	Prec	0.668	**0.683**	0.621	0.657	0.652
	F1	0.666	**0.676**	0.621	0.649	0.651
	AUC	0.707	0.712	0.667	**0.715**	0.696
	PPV	0.683	**0.712**	0.619	0.685	0.664
	Sensitivity	0.621	0.598	**0.629**	0.561	0.614
	Specificity	0.712	**0.758**	0.614	0.742	0.689
	FDR	0.317	**0.288**	0.381	0.315	0.336
	FOR	0.347	**0.346**	0.377	0.372	0.359

TABLE VI

RESULTS WITH DIFFERENT λ VALUES (MULTI-CLASS CLASSIFICATION - OTTAWA DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Effusion, Pneumothorax and No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	0.457	0.491	0.453	0.474	**0.496**
	Prec	0.465	0.488	0.462	0.472	**0.491**
	F1	0.443	**0.488**	0.453	0.470	**0.488**
	AUC	0.643	0.663	0.644	0.650	**0.677**
VGG16	Acc / Rec	0.474	**0.512**	**0.512**	0.496	0.474
	Prec	0.466	**0.518**	0.516	0.486	0.474
	F1	0.459	**0.517**	0.507	0.482	0.445
	AUC	*0.697*	0.709	**0.722**	0.688	0.667
ResNet50	Acc / Rec	***0.496***	0.419	0.466	0.487	0.453
	Prec	***0.493***	0.407	0.448	0.486	0.437
	F1	***0.488***	0.408	0.450	0.484	0.441
	AUC	**0.683**	0.610	0.675	0.679	0.674

TABLE VII

RESULTS WITH DIFFERENT λ VALUES (BINARY CLASSIFICATION - NIH DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Effusion vs No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	0.817	0.822	0.828	0.821	**0.855**
	Prec	0.823	0.824	0.828	0.821	**0.857**
	F1	0.817	0.821	0.828	0.821	**0.854**
	AUC	0.884	0.896	0.896	0.894	**0.909**
	PPV	*0.865*	0.851	0.838	0.829	**0.889**
	Sensitivity	0.752	0.780	**0.813**	0.809	0.810
	Specificity	*0.882*	0.863	0.843	0.833	**0.899**
	FDR	*0.135*	0.149	0.162	0.171	**0.111**
	FOR	0.219	0.203	0.182	0.186	**0.174**
VGG16	Acc / Rec	***0.834***	0.821	0.821	0.820	0.825
	Prec	***0.835***	0.822	0.822	0.824	0.826
	F1	***0.834***	0.821	0.821	0.820	0.825
	AUC	***0.907***	0.900	0.896	0.900	0.891
	PPV	0.817	0.843	0.840	**0.858**	0.846
	Sensitivity	***0.861***	0.789	0.793	0.769	0.794
	Specificity	0.807	0.853	0.850	**0.872**	0.856
	FDR	0.183	0.157	0.160	**0.142**	0.154
	FOR	***0.147***	0.198	0.196	0.210	0.194
ResNet50	Acc / Rec	0.809	0.808	0.784	**0.815**	0.809
	Prec	0.814	0.809	0.788	**0.815**	0.810
	F1	0.808	0.808	0.784	**0.815**	0.809
	AUC	0.870	0.863	0.860	**0.874**	0.872
	PPV	**0.854**	0.803	0.819	0.805	0.821
	Sensitivity	0.746	0.818	0.731	**0.831**	0.790
	Specificity	**0.872**	0.799	0.838	0.799	0.828
	FDR	**0.146**	0.197	0.181	0.195	0.179
	FOR	0.226	0.186	0.243	**0.175**	0.202

TABLE VIII

RESULTS WITH DIFFERENT λ VALUES (BINARY CLASSIFICATION - NIH DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Pneumothorax vs No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	*0.771*	**0.793**	0.784	0.759	0.790
	Prec	*0.772*	**0.793**	0.784	0.760	0.791
	F1	*0.771*	**0.793**	0.783	0.758	0.790
	AUC	*0.842*	**0.872**	0.858	0.841	0.868
	PPV	*0.786*	**0.804**	0.769	0.778	0.796
	Sensitivity	0.745	0.774	**0.811**	0.724	0.781
	Specificity	*0.797*	**0.811**	0.756	0.793	0.800
	FDR	*0.214*	**0.196**	0.231	0.222	0.204
	FOR	0.242	0.218	**0.200**	0.258	0.215
VGG16	Acc / Rec	0.765	0.763	0.768	0.776	**0.778**
	Prec	0.766	0.764	0.768	0.776	**0.778**
	F1	0.765	0.763	0.768	0.776	**0.778**
	AUC	0.841	0.828	0.848	0.848	**0.860**
	PPV	0.751	0.753	**0.771**	0.768	0.769
	Sensitivity	***0.795***	0.784	0.761	0.790	**0.795**
	Specificity	0.736	0.743	**0.774**	0.761	0.761
	FDR	0.249	0.247	**0.229**	0.232	0.231
	FOR	*0.218*	0.226	0.236	0.216	**0.212**
ResNet50	Acc / Rec	0.727	**0.749**	0.731	0.738	0.743
	Prec	0.727	**0.751**	0.733	0.739	0.746
	F1	0.727	**0.749**	0.731	0.738	0.742
	AUC	0.817	**0.821**	0.807	0.818	0.804
	PPV	0.732	**0.733**	0.713	0.723	0.719
	Sensitivity	0.715	0.784	0.774	0.772	**0.797**
	Specificity	**0.738**	0.715	0.688	0.704	0.688
	FDR	0.268	**0.267**	0.287	0.277	0.281
	FOR	0.278	0.232	0.247	0.244	**0.228**

TABLE IX

RESULTS WITH DIFFERENT λ VALUES (MULTI-CLASS CLASSIFICATION - NIH DATASET). $\lambda = 1$: CROSS-ENTROPY LOSS, $\lambda = *$: ADAPTIVE CUSTOM LOSS, BOLD: BEST RESULTS WITHIN THE SAME MODEL, UNDERLINE: BEST RESULTS OVER ALL MODELS, ITALICS: BEST RESULTS OVER ALL MODELS PRIOR TO "ATTENTION LOSS" BEING USED.

Effusion, Pneumothorax and No Finding

Model	Metric	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = *$
PVT	Acc / Rec	0.651	0.651	0.670	0.669	**0.677**
	Prec	0.651	0.651	0.672	0.669	**0.678**
	F1	0.652	0.651	0.671	0.669	**0.677**
	AUC	0.812	0.819	**0.843**	0.828	0.839
VGG16	Acc / Rec	***0.674***	0.653	0.640	0.660	0.641
	Prec	***0.674***	0.657	0.641	0.660	0.641
	F1	***0.674***	0.651	0.640	0.660	0.641
	AUC	***0.831***	0.822	0.820	0.822	0.817
ResNet50	Acc / Rec	0.631	0.610	0.601	0.620	**0.632**
	Prec	0.631	0.609	0.598	0.622	**0.632**
	F1	0.626	0.609	0.598	0.609	**0.630**
	AUC	**0.800**	0.781	0.781	0.793	0.794