UDC 004.932.2:004.8:616-073.75
DOI https://doi.org/10.32782/3041-2080/2026-7-6

PNEUMOTHORAX SEGMENTATION ON CHEST X-RAYS: PROGRESSIVE FINE-TUNING AND THRESHOLD-BASED PREDICTION REFINEMENT

Tanasuik Dmytro Olehovich,
PhD Student at the Department of Automation of Production Processes
Donbass State Engineering Academy
ORCID ID: 0009-0005-3413-3324

Razzhyvin Oleksii Valeriiovych,
PhD in Technical Sciences,
Associate Professorat at the Department of Automation of Production Processes
Donbass State Engineering Academy
ORCID ID: 0000-0002-1371-2651

Pneumothorax segmentation on chest radiographs remains a critical challenge for computer-aided triage, as small pleural separations are difficult to reliably localize. This study presents a segmentation method based on a U-Net-style deep neural network with a pretrained convolutional encoder, progressively fine-tuned on the SIIM-ACR Pneumothorax dataset. Training utilizes a two-stage resolution approach – starting at 512×512 for stable localization, then refining at 1024×1024 to recover thin pleural boundaries. Optimization incorporates staged encoder unfreezing, cyclic learning-rate scheduling, and a loss curriculum transitioning from weighted binary cross-entropy to soft Dice and symmetric Lovasz objectives.

The experimental analysis is conducted in three stages. First, three encoder backbones (ResNet34, EfficientNet, and SE-ResNeXt-50) are compared. Second, the best-performing backbone is evaluated across confidence thresholds to determine the optimal threshold-only operating point. Third, the model is assessed under joint confidence-threshold and minimum-mask-size control, where excessively small predicted masks are reset to empty. This structure reflects the SIIM-ACR evaluation protocol, where correct empty-mask predictions on negative studies heavily impact the overall mean Dice score.

Results identify SE-ResNeXt-50 as the strongest backbone. While a 0.70 confidence threshold maximizes threshold-only performance, the strongest overall result is achieved by combining a 0.55 confidence threshold with a minimum positive-pixel threshold of 2000. These findings demonstrate that explicit operating-point design – converting continuous probability maps into binary clinical decisions – is essential for pneumothorax segmentation and should be reported alongside architecture and training methodologies.

Key words: *pneumothorax, chest radiography, medical image segmentation, U-Net, SE-ResNeXt-50, confidence threshold, post-processing.*

Танасюк Дмитро, Разживін Олексій. Сегментація пневмотораксу на рентгенограмах грудної клітки: прогресивне донавчання та удосконалення прогнозу на основі порогової обробки

У статті розглянуто практичний підхід до сегментації пневмотораксу на рентгенограмах органів грудної клітки з використанням згорткової архітектури нейронної мережі типу U-Net із попередньо натренованим енкодером. Оскільки рентгенографія залишається найдоступнішим методом візуалізації, а виявлення невеликих колапсів легені є складним навіть для фахівців, розробка таких систем є вкрай актуальною. Запропонована методика поєднує поетапне донавчання моделі з прогресивною зміною просторової роздільної здатності від 512×512 до 1024×1024. Процес оптимізації включає поступове розморожування енкодера нейронної мережі, циклічний графік швидкості навчання та послідовне застосування функцій втрат: weighted binary cross-entropy, soft Dice і symmetric Lovasz.

Експериментальне дослідження організовано у три послідовні етапи. Спочатку за однакової схеми навчання порівнюються три варіанти архітектури нейронних мереж (енкодери): ResNet34, EfficientNet та SE-ResNeXt-50. На другому етапі для найкращої архітектури аналізується залежність якості сегментації від порога бінаризації карти ймовірностей. На третьому етапі модель оцінюється за спільного контролю порога впевненості та мінімальної кількості позитивних пікселів, де занадто малі маски примусово скидаються до порожніх. Це відображає специфіку протоколу оцінювання SIIM-ACR, у якому правильне передбачення порожньої маски для негативних випадків має вирішальний вплив на підсумковий показник Dice.

Результати доводять, що серед розглянутих варіантів SE-ResNeXt-50 забезпечує найкращий баланс. В аналізі лише за порогом упевненості оптимальним є значення 0.70. Проте найкращий підсумковий результат досягається за комбінованої схеми: поріг упевненості 0.55 разом із мінімальним порогом позитивних пікселів 2000. Це ефективніше пригнічує хибнопозитивні дрібні області, не погіршуючи сегментацію значущих випадків. Отримані дані свідчать, що вибір робочої точки є невід'ємною складовою алгоритму сегментації, яку необхідно звітувати нарівні з архітектурою та деталями навчання.

Ключові слова: *пневмоторакс, рентгенографія органів грудної клітки, сегментація медичних зображень, U-Net, SE-ResNeXt-50, робоча точка, поріг впевненості, післяобробка.*

Introduction. Pneumothorax is a clinically significant condition in which air accumulates in the pleural cavity and may lead to partial or complete lung collapse. In emergency and routine radiological practice, chest radiography remains one of the fastest and most accessible diagnostic tools, yet subtle cases are often difficult to identify because the main visual marker may be represented only by a thin displaced pleural line. This creates an important practical task for computer vision systems that can assist with localization and segmentation of pneumothorax regions on chest radiographs [1; 2].

The SIIM-ACR Pneumothorax Segmentation Challenge further highlighted the applied relevance of this task by establishing a large benchmark dataset and an evaluation protocol in which correct prediction of an empty mask on a negative study is rewarded with the maximum case-level score [1; 2]. Consequently, the final quality of a segmentation system depends not only on the network architecture and training strategy, but also on the decision rule used to transform a continuous probability map into a binary mask. This makes the problem relevant both from the viewpoint of medical image segmentation and from the viewpoint of practical deployment of diagnostic support systems.

From an applied perspective, the task is difficult not only because of the visual subtlety of many positive cases, but also because the clinically relevant target is structurally atypical for standard semantic segmentation. Pneumothorax often forms a thin peripheral region adjacent to the pleural boundary rather than a large compact object in the center of the image. As a result, minor errors in confidence distribution may change the binary decision from a plausible contour to an empty mask or vice versa. This characteristic makes the problem especially sensitive to the interaction between training loss, resolution, and inference thresholding, and partly explains why challenge solutions often combined conventional encoder–decoder models with carefully tuned decision rules [2; 3].

Analysis of recent research and publications. U-Net-like encoder–decoder architectures remain a widely used basis for medical image segmentation and have also shown strong performance for pneumothorax segmentation on chest radiographs [4; 6]. In the SIIM-ACR context, practical systems have combined U-Net decoders with pretrained backbones such as ResNet, EfficientNet, and SE-ResNeXt, while challenge-winning solutions often relied on ensembling, staged optimization, and carefully tuned post-processing [2; 3; 4].

Another important research direction concerns the choice of objective functions and structural priors. Overlap-oriented losses such as Dice and Lovasz-based surrogates have been used to improve region consistency, whereas boundary-aware formulations have been proposed specifically because pneumothorax contours on chest radiographs are thin, blurred, and difficult to delineate accurately [5]. Recent works also investigate two-stage solutions, anatomical priors, and segmentation-based quantification of pneumothorax extent [6; 7].

At the same time, previous studies and challenge reports indicate that the transition from probability map to final binary mask is a substantial part of the system rather than a minor technical detail. Confidence thresholding, area suppression, and multistep post-processing have repeatedly been used to reduce clinically implausible false positives and to adapt model behavior to the evaluation metric [2; 3]. However, the contribution of these operating-point choices is not always analyzed separately from the contribution of the backbone architecture or the loss function. This leaves insufficiently studied the question of how architectural selection, threshold-only control, and threshold-plus-area control interact within one unified experimental design.

This observation reveals an important methodological gap. Architectural comparison alone does not fully explain final performance when the evaluation metric itself couples segmentation quality with image-level false-positive control. In practical terms, two models with similar probability maps may behave differently after binarization, and one of them may obtain a better final Dice score simply because it produces fewer weak positive activations on negative studies. For pneumothorax segmentation, this means that the experimental logic should include not only the network backbone and loss design, but also the mechanism by which continuous outputs are converted into final binary masks [1; 2].

Research purpose. The purpose of this study is to substantiate a practical method for pneumothorax segmentation on chest radiographs and to determine how its final quality is influenced by three successive design choices: encoder backbone selection, confidence-threshold selection, and joint confidence-threshold and minimum-mask-size selection. The study aims to identify the strongest backbone under a common training framework, to establish the best threshold-only operating point for the selected model, and to determine the best final configuration after adding positive-pixel suppression.

Basic material presentation. The experiments are performed on the SIIM-ACR Pneumothorax

Segmentation dataset [1; 2]. The original run-length encoded annotations are grouped at the image level and transformed into binary masks aligned with 1024 × 1024 chest radiographs. A stratified train/validation/test split is used so that positive and negative studies remain reasonably balanced across subsets. Specifically, the training subset comprises 8643 images (6719 negative and 1924 positive cases). The validation subset, utilized for hyperparameter tuning and operating-point selection, contains 964 images (747 negative and 217 positive cases). Finally, the held-out test subset, reserved strictly for final performance reporting, includes 1068 images (830 negative and 238 positive cases). This explicit stratification guarantees a consistent proportion of positive to negative examinations across all experimental phases, preventing disproportionate concentrations of either subtle pneumothoraces or obvious empty cases in any single subset. All model selection steps, including backbone selection and operating-point tuning, are performed using the training and validation data, whereas the final numerical results are reported on the held-out test split.

Preprocessing and augmentation. Input images are normalized using dataset-specific channel statistics. Although the underlying data are chest radiographs, the network operates on three-channel inputs so that pretrained convolutional encoders can be used without changing their first-layer structure. Training-time augmentation includes geometric perturbations based on translation, scaling, and rotation, together with low-probability intensity transformations such as contrast, gamma, and brightness adjustment. Such augmentation is deliberately moderate and is intended to improve robustness without introducing unrealistic radiographic patterns.

Although the underlying modality is grayscale, representing the radiographs as three-channel inputs is a practical compromise that allows direct use of pretrained natural-image encoders without reinitializing the first convolutional layer. In this setting, the gain does not come from color information, which is absent, but from transfer of low-level edge, contrast, and texture detectors into the medical domain. The augmentation policy is intentionally conservative. Horizontal flipping is avoided because left–right reversal may distort clinically meaningful asymmetries, and overly aggressive photometric augmentation could alter the delicate pleural edge contrast that distinguishes subtle pneumothorax from normal lung boundary. Instead, the augmentation strategy focuses on moderate geometric perturbation and restrained intensity variation to improve robustness while preserving radiographic plausibility.

Network architecture. The base segmentation model is an encoder–decoder network with a U-Net-style decoder and skip connections from intermediate encoder stages. The decoder combines upsampling, lateral feature fusion, and convolutional refinement, while the network outputs a single logit map converted to probabilities by the sigmoid function. Within this common design, three encoder families are investigated: ResNet34, EfficientNet, and SE-ResNeXt-50. Their comparison is intended to determine which feature extractor provides the best balance between localization quality on positive studies and stability on negative studies [3; 4].

Within this architecture, the encoder and decoder fulfill different functional roles. The pretrained backbone acts as a hierarchical feature extractor that progressively aggregates large-scale contextual information, while the decoder reconstructs spatial detail from semantically enriched feature maps. This separation is particularly useful in chest radiography, where the model must distinguish true pleural separation from a wide range of confounders, including rib edges, overlapping soft tissue, acquisition artifacts, and chest tubes. The skip connections allow local boundary cues to remain accessible during upsampling, while the deeper encoder features help suppress anatomically implausible responses. The comparison among ResNet34, EfficientNet, and SE-ResNeXt-50 therefore reflects more than computational preference; it also tests different trade-offs between feature diversity, representation depth, and robustness of transferred pretrained filters.

Optimization strategy and 512 × 512 warm start. Training is organized as a progressive fine-tuning procedure. The first stage is performed at 512 × 512 resolution and serves as a computationally efficient warm start. At this stage, the model can be optimized with larger effective batch sizes and greater stability, while already learning the large-scale distinction between empty and non-empty studies and the approximate localization of pleural abnormalities. The 512 × 512 stage is therefore not merely a technical simplification, but a regularized initialization phase that prepares both decoder and upper encoder blocks for subsequent high-resolution refinement.

The first 512 × 512 stage can be viewed as a task-specific pretraining phase carried out within the target dataset. Its purpose is not to produce the final operating model, but to place the parameters into a stable basin before full-resolution optimization. At this scale, the model learns coarse localization of pleural abnormalities, adapts the decoder to the radiographic domain, and begins to calibrate the relationship between positive and

negative studies. The reduced computational burden makes it possible to train with a larger effective batch size and more frequent parameter updates, which is especially helpful when the backbone is being progressively unfrozen. In this sense, the 512 × 512 stage provides both optimization stability and a regularizing effect: the model first learns where the disease is likely to appear before being asked to resolve the exact contour at 1024 × 1024.

After this warm start, the model is fine-tuned at full 1024 × 1024 resolution. The higher resolution is necessary because pneumothorax boundaries may be represented as thin peripheral structures that are under-resolved at smaller scales. Encoder unfreezing is staged: first, only a limited subset of layers is trainable, and then progressively deeper encoder blocks are released. This reduces instability and allows the decoder to adapt before all backbone parameters are updated jointly.

The transition to 1024 × 1024 is essential for the final stage because clinically relevant pneumothorax contours are often only a few pixels wide after downsampling. A model trained only at reduced resolution may identify the general region of abnormality, yet still blur the boundary or produce fragmented low-confidence regions near the pleura. Full-resolution training restores the spatial granularity needed for more realistic contour formation. It also changes the optimization problem: at high resolution, the network must simultaneously maintain global contextual understanding and refine a sparse, elongated target embedded in a very large background. This is one reason why the earlier warm-start phase is useful; by the time full-resolution refinement begins, the model has already learned a stable global representation of the task and can dedicate more capacity to spatial precision.

Optimization uses Adam together with a cyclic learning-rate schedule of triangular type and, where required, gradient accumulation. In this schedule, the learning rate increases from a base value to a maximum value and then decreases again over a fixed epoch cycle. During early adaptation, such oscillation helps newly trainable layers escape poor local minima without relying on an overly aggressive constant learning rate. During later stages, it enables controlled refinement in a low-learning-rate regime. In practice, wider cycles and larger learning-rate amplitudes are used earlier, whereas narrower, gentler cycles are used during full-resolution fine-tuning.

In practical optimization terms, the triangular cycle avoids two opposite failure modes. A uniformly low learning rate can make later-stage fine-tuning too conservative, especially after additional encoder blocks are unfrozen, whereas a uniformly high learning rate may destabilize already useful pretrained representations. By repeatedly increasing and decreasing the learning rate within controlled bounds, the schedule encourages periodic exploration without abandoning the fine-grained structures already learned. This behavior is particularly helpful when different loss functions are introduced across stages, because the objective landscape changes as the model moves from pixelwise calibration to overlap-oriented refinement. The cycle therefore acts as a transition mechanism between stages of the curriculum rather than only as a generic optimization heuristic.

Loss functions. The optimization objective is also staged. Weighted binary cross-entropy is used early to compensate for the strong foreground–background imbalance inherent in pneumothorax segmentation. By assigning more influence to positive pixels, this loss helps the model learn stable coarse localization before boundary quality becomes reliable. Soft Dice is introduced in later stages to emphasize overlap quality at the object level and to reduce domination by the large number of easy background pixels. In the final phase, symmetric Lovasz loss is used to align optimization more closely with set-based overlap behavior. Together, these losses form a curriculum: first the network learns to distinguish foreground from background, then it improves mask overlap, and finally it refines difficult pixel-ordering errors affecting the contour [4; 5].

The interaction among the three losses is complementary rather than redundant. Weighted binary cross-entropy provides dense pixelwise supervision and encourages clear separation between foreground and background logits, which is especially important at the beginning of training when predictions are noisy. Soft Dice then shifts emphasis toward whole-mask agreement and reduces the dominance of the background class in the gradient signal. Symmetric Lovasz is introduced only after masks become reasonably stable, because its ranking-oriented behavior is more informative when the model already distinguishes foreground from background in a consistent way. Under this curriculum, the optimization proceeds from coarse discrimination to overlap improvement and finally to fine-grained correction of hard errors near the decision boundary [4; 5].

Implementation details. All experiments were conducted utilizing an NVIDIA Tesla P100 GPU. To accommodate the computational demands of high-resolution image processing, the training procedure was explicitly partitioned into two stages with distinct dataloader configurations. During the

initial 512 × 512 warm-start phase, the network was trained with a batch size of 14 and 4 data loader workers. For the subsequent 1024 × 1024 refinement phase, the batch size was reduced to 3 with 2 data loader workers to adhere to GPU memory limits. Optimization was governed by a cyclic learning-rate schedule of triangular type, where the learning rate oscillated between defined base and maximum values across specified epoch spans. The training curriculum comprised a total of 50 epochs (40 epochs at 512 × 512 resolution, followed by 10 epochs at 1024 × 1024 resolution) and incorporated staged unfreezing of the pre-trained backbone blocks alongside a progressive loss function transition. The precise hyperparameters for each mini-stage are detailed in Table 1.

Inference and operating-point control. At inference time, the network produces a continuous probability map, which must be converted into a binary mask. Instead of fixing the confidence threshold at 0.50, the present work evaluates a range of candidate thresholds on the validation set. In the final step, the thresholded mask may also be reset to empty if the total number of positive pixels is below a selected minimum value. This rule is intended to suppress small scattered false-positive regions that are unlikely to correspond to a plausible pneumothorax extent. Such a strategy is especially relevant under the SIIM-ACR scoring protocol, where an incorrect non-empty mask on a negative study is heavily penalized [1; 2].

The operating-point analysis is therefore more than a post-processing convenience. It is the stage at which the continuous statistical output of the network is translated into a discrete clinical decision. Confidence thresholding determines how much evidence is required before a pixel is declared positive, whereas the positive-pixel threshold determines how much spatial support is required before an image is treated as containing a meaningful lesion. These two controls influence different error types. Confidence thresholding mainly changes the aggressiveness of the mask itself, while minimum-mask-size suppression removes isolated fragments that may survive thresholding but remain too small to be clinically plausible. Their joint analysis is thus necessary to understand how the model behaves under the SIIM scoring protocol.

Experimental results. The experimental evaluation is organized in three successive stages. First, encoder backbones are compared under a common decoder and training scheme. Second, the best backbone is evaluated across different confidence thresholds. Third, the same model is analyzed under joint confidence-threshold and minimum-mask-size control. This design makes it possible to separate the contribution of architecture selection from the contribution of operating-point selection.

In the following tabular results, the metric termed "Positive accuracy" represents image-level sensitivity for the positive class; it is calculated as the proportion of ground-truth positive studies in which the model correctly predicted a non-empty mask.

As shown in Table 2, SE-ResNeXt-50 provides the strongest overall backbone among the tested candidates. It achieves the highest mean Dice and the highest negative-case Dice, and therefore serves as the base model for the remaining experiments.

Although EfficientNet demonstrated a higher Positive-case Dice (0.4259 compared to 0.4072 for SE-ResNeXt-50), the selection of SE-ResNeXt-50 as the optimal backbone was driven purely by the mathematical formulation of the global evaluation metric. Under the SIIM-ACR scoring protocol, the final Mean Dice score heavily penalizes false-positive predictions on negative (empty) studies. SE-ResNeXt-50 achieved a noticeably higher

Table 1

Progressive training schedule, staged unfreezing, and cyclic learning rates

Resolution	Backbone Unfreezing Phase	Loss Function	Epochs	Base LR	Max LR
512x512	Decoder only (Backbone frozen)	Weighted BCE	12	1.00E-04	1.00E-03
	Unfreeze Backbone Block 4	Weighted BCE	2	5.00E-05	5.00E-04
	Unfreeze Backbone Block 3	Weighted BCE	2	5.00E-05	5.00E-04
	Unfreeze Backbone Block 2	Weighted BCE	2	5.00E-05	5.00E-04
	Unfreeze Backbone Block 1	Weighted BCE	2	5.00E-05	5.00E-04
	All layers unfrozen	Weighted Soft Dice	10	1.00E-05	1.00E-04
	All layers unfrozen	Symmetric Lovasz	10	1.00E-05	1.00E-04
1024x1024	All layers unfrozen	Weighted Soft Dice	4	1.00E-05	1.00E-04
	All layers unfrozen	Symmetric Lovasz	6	3.00E-06	6.00E-05

Negative-case Dice (0.9747 versus 0.9614 for EfficientNet), which mathematically outweighed its slightly lower positive-case performance. By minimizing weak positive activations outside plausible pleural regions, SE-ResNeXt-50 ultimately yielded the highest Global Mean Dice (0.8482). The selection is therefore justified by a strict alignment with the primary benchmark metric, explicitly trading a marginal loss in positive-case sensitivity for a critical reduction in false-positive penalties on healthy cases.

Table 3 shows the performance of the selected SE-ResNeXt-50 model across confidence thresholds. The results show that threshold tuning materially changes the trade-off between positive-case sensitivity and suppression of false-positive masks. Among threshold-only settings, the best operating point is achieved at a confidence threshold of 0.70.

The threshold-only analysis demonstrates that the default value of 0.50 should not be interpreted as an inherently meaningful operating point. It is merely a conventional midpoint on the sigmoid scale. The experiments show that the model reaches its best threshold-only result at 0.70, which implies that the raw probability map is better used in a more conservative manner. This observation is consistent with the class imbalance of the task and with the challenge metric itself. When false-positive fragments on negative studies are expensive, a threshold that is too permissive can lower final performance even if it preserves additional low-confidence positives.

The threshold-only analysis indicates that a more conservative decision rule improves overall performance mainly through stronger control of false-positive non-empty predictions on negative studies. However, confidence thresholding alone does not fully eliminate weak scattered activations.

Table 4 summarizes the results of the joint confidence-threshold and minimum positive-pixel analysis. The combination of a 0.55 confidence threshold and a 2000-pixel minimum mask size yields the highest overall Mean Dice score (0.8549). While this configuration results in a slight decrease in Positive-case Dice compared to lower pixel thresholds (e.g., 0.4019 at 1000 pixels versus 0.3995 at 2000 pixels), the decision to adopt the 0.55 + 2000 operating point was driven entirely by the SIIM-ACR benchmark metric. The global Mean Dice heavily penalizes false-positive predictions on empty studies. Suppressing masks smaller than 2000 pixels aggressively eliminates spurious, low-volume false positives, thereby maximizing the Negative-case Dice (0.9855). Consequently, the chosen operating point represents a strict mathematical optimization for the challenge evaluation protocol. It deliberately sacrifices marginal sensitivity to exceedingly small true-positive regions in exchange for a substantial reduction in false-positive penalties across the dataset.

Taken together, the staged evaluation supports a concrete final method. The comparison of backbones identifies SE-ResNeXt-50 as the strongest encoder under the common optimization scheme. The threshold-only study identifies 0.70 as the best confidence threshold when no additional suppression is used. The final joint analysis shows that the strongest overall configuration is obtained at confidence threshold 0.55 with a minimum positive-pixel threshold of 2000. Thus, the final system

Table 2

Comparison of backbone encoders under the common training setup

Backbone	Mean Dice	Positive-case Dice	Negative-case Dice	Positive accuracy
ResNet34	0.8229	0.4027	0.9433	0.8403
EfficientNet	0.8421	0.4259	0.9614	0.7479
SE-ResNeXt-50	0.8482	0.4072	0.9747	0.7017

Table 3

SE-ResNeXt-50 performance across confidence thresholds

Confidence threshold	Mean Dice	Positive-case Dice	Negative-case Dice	Positive accuracy
0.1	0.697654	0.449978	0.768675	0.953782
0.2	0.806575	0.417738	0.918072	0.857143
0.3	0.836135	0.407528	0.959036	0.739496
0.4	0.843214	0.409886	0.96747	0.722689
0.5	0.848228	0.407172	0.974699	0.701681
0.6	0.851421	0.400495	0.980723	0.659664
0.7	0.851857	0.385646	0.985542	0.642857
0.8	0.847855	0.359283	0.987952	0.592437
0.9	0.835525	0.303954	0.987952	0.542017

Table 4

SE-ResNeXt-50 performance under joint confidence-threshold and positive-pixel-threshold selection

Confidence threshold	Pixel threshold	Mean Dice	Positive-case Dice	Negative-case Dice	Positive accuracy
0.6	250	0.8542	0.4003	0.9843	0.6512
0.6	500	0.8541	0.4003	0.9843	0.6512
0.55	1000	0.8545	0.4019	0.9843	0.6428
0.55	1500	0.8540	0.3995	0.9843	0.6386
0.55	2000	0.8549	0.3995	0.9855	0.6386
0.55	2500	0.8537	0.3942	0.9855	0.6260

is defined jointly by its backbone, training schedule, and inference decision rule.

Model robustness and variance. To verify the stability of the proposed approach and ensure the results are not heavily biased by a single data partition, an additional 10-fold cross-validation experiment was conducted. The selected SE-ResNeXt-50 architecture was evaluated across all ten folds utilizing the confidence-threshold-only operating condition. Across the 10 independent test sets, the model achieved a mean Dice score of 0.8495 with a standard deviation of only 0.0038 (alongside a validation mean of 0.8452 ± 0.0044). This exceptionally low variance demonstrates that the model's representational capabilities are highly robust and generalize consistently across different subsets of the data. Although the final combined post-processing rule (confidence threshold plus minimum positive pixel mask) was specifically tuned and evaluated on the primary stratified split to establish the absolute optimal operating point for the SIIM-ACR protocol, the underlying stability of the backbone and training methodology is clearly confirmed by the cross-validation results.

Context within state-of-the-art methods. While top-performing solutions in the SIIM-ACR challenge often rely on heavy ensembling of multiple architectures to maximize absolute metric scores, the primary objective of this study is methodological transparency. Complex ensembles obscure the specific impact of individual design choices. By analyzing a single, strong baseline model (SE-ResNeXt-50), this research explicitly isolates how the staged training curriculum and post-processing decision rules (confidence thresholding and minimum mask size) independently contribute to final performance. Therefore, rather than proposing a highly parameterized ensemble to compete for marginal state-of-the-art leaderboard gains, this study offers a clear, reproducible framework demonstrating how systematic operating-point selection bridges the gap between raw probability maps and clinically applicable, metric-optimized predictions.

Conclusions. The conducted study substantiates a practical method for pneumothorax segmentation on chest radiographs based on a U-Net-style decoder, pretrained convolutional encoders, progressive 512 × 512 to 1024 × 1024 fine-tuning, cyclic learning-rate scheduling, and a staged BCE-Dice-Lovasz loss curriculum. The three-stage evaluation demonstrates that the final quality of the method is determined by a consistent chain of design choices. SE-ResNeXt-50 proved to be the strongest backbone among the evaluated candidates. Within threshold-only analysis, the best operating point was achieved at confidence threshold 0.70. The strongest final result was obtained when a moderate confidence threshold of 0.55 was combined with suppression of masks containing fewer than 2000 positive pixels.

An additional practical advantage of this methodology is its transparency. Each major performance gain can be attributed to a clearly identifiable design choice: the backbone comparison establishes the representation capacity of the encoder, the 512 × 512 warm start improves optimization stability, the 1024 × 1024 refinement restores boundary detail, and the final operating-point selection aligns inference with the evaluation metric. Such transparency is useful both scientifically and operationally. Scientifically, it allows the contribution of architecture, training schedule, and decision rule to be separated. Operationally, it suggests that meaningful gains can still be obtained through disciplined experimental design, even when no architectural novelty is introduced.

These findings show that the final operating point must be treated as an integral element of a pneumothorax segmentation method rather than a secondary afterthought. Under the SIIM-ACR metric, the conversion of a continuous probability map into a clinically meaningful binary mask strongly influences overall performance, particularly through the control of false-positive masks

on negative studies. Consequently, architecture, training schedule, loss design, and inference decision rule should be reported together as components of one system.

Limitations and prospects for further research. A primary limitation of the current study is that the proposed model has been trained and evaluated exclusively on the SIIM-ACR dataset. Consequently, its direct applicability to diverse, real-world clinical environments remains an open question. Radiographic characteristics vary significantly across medical institutions due to differences in X-ray acquisition protocols and input data characteristics. For instance, the visual presentation of a pneumothorax differs substantially between erect (standing) and supine (lying down) patient positioning, the latter being very common in emergency and intensive care units. Furthermore, variations in hardware (portable versus fixed-room scanners), exposure parameters, and the presence of overlapping clinical features (such as pleural effusions, pneumonia, chest tubes, or external medical lines) can alter the delicate contrast of the pleural line.

Because the proposed method relies heavily on a tuned operating point (0.55 confidence and 2000-pixel minimum area) specifically optimized for the SIIM-ACR dataset's unique imaging distribution, these static thresholds may not generalize out-of-the-box to datasets with different image characteristics. Therefore, future research must prioritize external validation on independent, multi-institutional datasets to evaluate the cross-domain transferability of the model and its post-processing rules. Additionally, further investigation should explore dynamic, calibration-aware operating-point selection, and compare simple positive-pixel suppression against more complex largest-connected-component analysis to better adapt to varying clinical inputs.

BIBLIOGRAPHY:

1. SIIM-ACR Pneumothorax Segmentation Challenge [Electronic resource]. Kaggle. URL: https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation (accessed: 25.03.2026).
2. Pneumothorax Kaggle Challenge overview and winning teams [Electronic resource]. SIIM. URL: https://siim.org/research-journal/siim-machine-learning-challenges/pneumothorax-kaggle-challenge/ (accessed: 25.03.2026).
3. Sneddy. SIIM-ACR Pneumothorax first-place solution [Electronic resource]. GitHub. URL: https://github.com/sneddy/pneumothorax-segmentation (accessed: 25.03.2026).
4. Abedalla A., Abdullah M., Al-Ayyoub M., Benkhelifa E. Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. PeerJ Computer Science. 2021. Vol. 7. Art. e607. DOI: 10.7717/peerj-cs.607.
5. Wang Y., Wang K., Peng X. et al. DeepSDM: Boundary-aware pneumothorax segmentation in chest X-ray images. *Neurocomputing*. 2021. Vol. 454. P. 201–211. DOI: 10.1016/j.neucom.2021.05.029.
6. Dumbrique J. I. S., Hernandez R. B., Cruz J. M. L. et al. Pneumothorax detection and segmentation from chest X-ray radiographs using a patch-based fully convolutional encoder-decoder network. *Frontiers in Radiology*. 2024. Vol. 4. Art. 1424065. DOI: 10.3389/fradi.2024.1424065.
7. Sae-Lim W., Wettayaprasit W., Suwannanon R., Cheewatanakornkul S., Aiyarak P. Automated pneumothorax segmentation and quantification algorithm based on deep learning. *Intelligent Systems with Applications*. 2024. Vol. 22. Art. 200383. DOI: 10.1016/j.iswa.2024.200383.

REFERENCES:

1. SIIM-ACR Pneumothorax Segmentation Challenge. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation
2. SIIM. (n.d.). Pneumothorax Kaggle Challenge overview and winning teams. Retrieved from https://siim.org/research-journal/siim-machine-learning-challenges/pneumothorax-kaggle-challenge/
3. Sneddy. (n.d.). SIIM-ACR pneumothorax first-place solution. GitHub. Retrieved from https://github.com/sneddy/pneumothorax-segmentation
4. Abedalla, A., Abdullah, M., Al-Ayyoub, M., & Benkhelifa, E. (2021). Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. PeerJ Computer Science, 7, e607. https://doi.org/10.7717/peerj-cs.607
5. Wang, Y., Wang, K., Peng, X., Shi, L., Sun, J., Zheng, S., Shan, F., Shi, W., & Liu, L. (2021). DeepSDM: Boundary-aware pneumothorax segmentation in chest X-ray images. *Neurocomputing*, 454, 201–211. https://doi.org/10.1016/j.neucom.2021.05.029
6. Dumbrique, J. I. S., Hernandez, R. B., Cruz, J. M. L., et al. (2024). Pneumothorax detection and segmentation from chest X-ray radiographs using a patch-based fully convolutional encoder-decoder network. *Frontiers in Radiology*, 4, Article 1424065. https://doi.org/10.3389/fradi.2024.1424065
7. Sae-Lim, W., Wettayaprasit, W., Suwannanon, R., Cheewatanakornkul, S., & Aiyarak, P. (2024). Automated pneumothorax segmentation and quantification algorithm based on deep learning. *Intelligent Systems with Applications*, 22, Article 200383. https://doi.org/10.1016/j.iswa.2024.200383

Дата першого надходження статті до видання: 14.04.2026
Дата прийняття статті до друку після рецензування: 11.05.2026
Дата публікації (оприлюднення) статті: 30.05.2026