

Unified model with random penalty entropy loss for robust nasogastric tube placement analysis in X-ray[☆]

GwiSeong Moon^{a,b,1}, Kyoung Min Moon^{e,1}, Inseo Park^b, Kanghee Lee^{a,b}, Doohee Lee^{a,b}, Woo Jin Kim^c, Yoon Kim^{a,b}, Ji Young Hong^d, Hyun-Soo Choi^{b,f,1,*}

^a Department of Computer Science and Engineering, Kangwon National University, Chuncheon, 24341, Kangwon, Republic of Korea

^b ZIOVISION Co., Ltd., Chuncheon, 24341, Kangwon, Republic of Korea

^c Department of Biomedical Informatics, Kangwon National University, Chuncheon, 24341, Kangwon, Republic of Korea

^d Division of Pulmonary and Critical Care Medicine, Department of Medicine, Chuncheon Sacred Heart Hospital, Hallym University Medical Center, Chuncheon, 24253, Kangwon, Republic of Korea

^e Division of Pulmonary and Allergy Medicine, Department of Internal Medicine, Chung-Ang University Hospital, Chung-Ang University College of Medicine, Seoul, 06973, Republic of Korea

^f Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, 01811, Republic of Korea

ARTICLE INFO

Keywords:

NG tube placement

Chest X-ray

Unified model

Random Penalty Entropy Loss

Model calibration

ABSTRACT

Background and objective: An accurate nasogastric (NG) tube placement assessment is essential to prevent serious complications. However, manual chest X-ray verification is prone to human error and variability. We propose a unified deep learning model that jointly performs segmentation and classification to improve the generalization and reliability of automated NG tube placement assessment.

Methods: We developed a unified architecture based on nnUNet, which was optimized simultaneously for segmentation and classification. To enhance robustness and reduce overconfidence, we introduce Random Penalty Entropy Loss, which dynamically scales entropy penalties during training. The model was evaluated on internal datasets (5674 chest X-rays from three South Korean hospitals) and an external dataset from MIMIC-CXR.

Results: On the internal test set, the proposed model outperformed the Wang 2-Stage method (F1: 93.94% vs. 87.39%), particularly in ambiguous cases. Baseline models using Focal Loss or Label Smoothing performed well internally but showed substantial performance drops and miscalibration externally. In contrast, our model with Random Penalty Entropy Loss achieved the highest external classification accuracy (F1: 66.34%, AUROC: 84.82%) and superior calibration (MCE: 0.429, ECE: 0.274).

Conclusion: The proposed unified model surpasses existing two-stage approaches in classification and calibration. Incorporating Random Penalty Entropy Loss improves robustness and generalization across diverse clinical settings. These results highlight the model's potential to reduce diagnostic errors and enhance patient safety in NG tube placement assessment.

1. Introduction

Nasogastric (NG) tubes are essential for enteral feeding in patients with swallowing impairments or critical illnesses. These tubes allow for direct delivery of nutrition and medication into the stomach. Each year, approximately 10 million NG tubes are inserted in Europe and 1.2 million in the United States (Energias Market Research, 2018), with the global market projected to grow at a CAGR of 6.4% from 2018 to 2024.

Despite their ubiquity, the blind insertion of NG tubes remains risky. Misplacement into the respiratory tract or causing pneumothorax can lead to severe, even fatal, complications. Thus, reliable verification of tube positioning is imperative (Boeykens et al., 2023). Chest X-rays (CXRs) are the gold standard for confirming NG tube placement (Metheny and Meert, 2004; Araujo-Preza et al., 2002; Ellett, 2004; Elpern et al., 2007; Kearns and Donna, 2001), but the interpretation depends on radiologist availability, making it prone to delays

[☆] This article is part of a Special issue entitled: 'AI-based CDSS' published in Computerized Medical Imaging and Graphics.

* Corresponding author at: ZIOVISION Co., Ltd., Chuncheon, 24341, Kangwon, Republic of Korea.

E-mail address: choi.hyunsoo@seoultech.ac.kr (H.-S. Choi).

¹ These authors contributed equally to this work.

or errors—especially when assessments are made by inexperienced personnel (NHS National Patient Safety Agency, 2011).

Deep learning-based diagnostic tools have shown significant promise in supporting medical image interpretation by reducing workload and minimizing human error. Initial models primarily focused on classification to distinguish correct from incorrect placements (Tang et al., 2023; Singh et al., 2019; Henderson et al., 2021). However, these models lack visual interpretability, which limits clinical trust and usability. More recent systems present segmentation alongside classification outputs (Elaanba et al., 2023), providing visual context that improves clinician confidence.

Two-stage models, such as those proposed by Wang et al. (2024), improve classification accuracy by feeding segmentation outputs into a secondary classification model. While effective, these cascaded systems depend heavily on segmentation quality, which may degrade under complex conditions. Moreover, prior studies have rarely addressed model calibration and robustness under severe class imbalance—critical factors for safety-critical applications like NG tube placement.

Transformer-based models have recently advanced medical imaging—especially in data-rich histopathology—by leveraging global context with attention-based interpretability, as demonstrated by LA-ViT (Huang et al., 2024b), ViT-AMC (Huang et al., 2023), FDTs (Huang and Luo, 2025), and MamlFormer (Huang et al., 2024a). However, ViT-style architectures typically achieve their strongest gains with large-scale pretraining or distillation and can be less data-efficient than convolutional networks when trained from scratch on moderate-sized, class-imbalanced radiographic datasets (see Table 1). Consistent with this observation, the recent large-scale benchmark study “nnU-Net Revisited” reports (Isensee et al., 2024) that nnU-Net (org.) attains a higher mean Dice score across six medical segmentation datasets (BTCV, ACDC, LiTS, BraTS, KiTS, AMOS) than representative Transformer-based alternatives, achieving 86.8% versus 83.0% for nnFormer and 85.3% for SwinUNETRv2. Moreover, SwinUNETRv2 requires substantially higher computational resources (13.4 GB VRAM and 15 GPU-hours runtime) than nnU-Net (org.) (7.7 GB and 9 GPU-hours), yet without surpassing its average accuracy. Given our safety-critical NG-tube setting and limited positive cases, we therefore adopt a CNN-based design.

To overcome these limitations, we propose a unified CNN-based deep learning framework that simultaneously performs segmentation and classification. Our approach surpasses state-of-the-art two-stage methods, particularly in ambiguous or degraded scenarios. In addition, we introduce Random Penalty Entropy Loss. This novel dynamic regularization technique adjusts entropy penalties during training to improve robustness, calibration, and generalization beyond existing methods like Focal Loss or Label Smoothing.

We conduct comprehensive evaluations using an internal multi-hospital dataset ($n = 5674$) and an external validation set curated from the MIMIC-CXR database. Our model demonstrates strong improvements across multiple metrics (F1, AUROC, AUPRC) and well-calibrated prediction distributions. Kernel Density Estimation (KDE) and calibration error analysis (ECE, MCE) further support the reliability of our method under varying clinical conditions.

The key contributions of this work are summarized as follows:

- **Unified Framework:** We propose a robust model integrating segmentation and classification into a single network. This unified design outperforms state-of-the-art two-stage pipelines, especially under challenging conditions.
- **Random Penalty Entropy Loss:** We introduce a dynamic entropy-based regularization method that enhances model calibration and generalization by adapting penalty strength during training.
- **NGT Generalization Dataset:** We curate and utilize a benchmark dataset derived from MIMIC-CXR to evaluate generalization performance under diverse and unseen clinical conditions.
- **Robust Generalization and Calibration:** Our model consistently achieves superior accuracy and reliability on internal and external datasets, validating its clinical applicability and technical advantage.

2. Background

2.1. Deep learning approaches for NG tube placement verification

Accurate verification of nasogastric (NG) tube placement is essential to avoid potentially life-threatening complications from tube misplacement. Although chest X-ray interpretation remains the clinical gold standard, manual assessment is time-consuming and subject to variability—especially in resource-limited environments—prompting the need for automated diagnostic support.

Recent studies have applied deep learning models to assist in NG tube placement verification. Early efforts focused primarily on classification models that distinguish correct from incorrect placements. For example, Singh et al. (2019) and Henderson et al. (2021) demonstrated promising diagnostic accuracy using CNN-based classifiers, highlighting AI’s potential utility in clinical workflows.

Subsequent approaches have combined classification with segmentation to improve interpretability and prediction accuracy. Segmentation provides spatial context, enabling clinicians to better assess tube positioning. Elaanba et al. (2023), for instance, presented a system that jointly visualizes segmentation and classification results, offering more comprehensive insights.

More recent models have employed two-stage pipelines that segment key anatomical structures (e.g., tube, lung, diaphragm) before classification (Wang et al., 2024; Jung et al., 2022). Wang et al. (2024) showed that incorporating spatial features from segmentation improved classification performance. However, the approach is sensitive to segmentation quality—especially in complex regions such as the diaphragm, which exhibited low Dice scores. These limitations highlight the vulnerability of cascaded systems under challenging conditions.

Moreover, existing studies have rarely addressed the effects of class imbalance and calibration on classification reliability despite their critical importance in high-stakes clinical applications such as NG tube verification.

2.2. Techniques for mitigating overfitting and enhancing generalization

Overfitting remains a persistent challenge in medical AI due to limited annotated data availability. Models trained on such datasets tend to memorize training samples rather than generalize to new, unseen data, undermining their clinical applicability.

Regularization techniques like Label Smoothing (Szegedy et al., 2016) and Focal Loss (Lin et al., 2017b) have been widely adopted to address this. Label Smoothing prevents overconfidence by softening the ground-truth distribution, improving model calibration and robustness to noisy labels (Müller et al., 2019). Initially designed for object detection, Focal Loss helps focus learning on difficult examples and alleviates the effects of class imbalance.

Advanced variants of these techniques—such as adaptive Focal Loss (Mukhoti et al., 2020), repeatability-optimized Focal Loss for clinical applications (Ahmed et al., 2021), margin-aware smoothing (Liu et al., 2023), and Bayesian regularization methods (Gal and Ghahramani, 2016)—have been explored to enhance generalization further. However, they often introduce complex hyperparameters that are hard to optimize, particularly in medical tasks with no established validation criteria and heterogeneous image characteristics.

Given these constraints, we used the base versions of Focal Loss and Label Smoothing in our experiments as comparative baselines. These loss functions provide robust, interpretable benchmarks, enabling us to assess the effectiveness of our proposed method without introducing additional task-specific complexity.

Table 1

Overview of the Internal and External datasets used for model training and evaluation. The Internal dataset is divided into Train, Validation, and Test subsets, while the External dataset consists of a single Test subset derived from the MIMIC-CXR database.

Dataset	Subset	Correct class	Misplaced class	Class ratio	Year
Internal dataset	Train	3423	230	15:1	2017~2023
Internal dataset	Validation	849	57	15:1	2017~2023
Internal dataset	Test	1054	61	17:1	2011~2020
External dataset	Test	1240	114	11:1	2011~2016

2.3. Enhancing reliability through model calibration

Model calibration is essential in medical AI, as it determines how well predicted probabilities reflect actual outcomes. Poorly calibrated models may exhibit high accuracy yet produce unreliable probability estimates—posing risks when predictions are used for clinical decision-making.

Recent studies in both general AI (Rajaraman et al., 2022; Guo et al., 2017; Wang et al., 2022, 2021) and medical imaging (Carse et al., 2022; Sambyal et al., 2023; Park et al., 2024) highlight the prevalence of calibration errors, particularly in imbalanced datasets. Overconfident, incorrect predictions are perilous in high-risk clinical settings.

Several strategies have been proposed to mitigate these issues. Temperature Scaling (Guo et al., 2017) is a post-hoc method that rescales logits to better align predicted probabilities with observed outcomes. Label Smoothing and Focal Loss integrate regularization into the training process to directly prevent overconfidence.

In our preliminary experiments, Temperature Scaling had minimal impact, as our model's predictions were already well-calibrated. This underscores the need for an integrated solution that improves calibration during training itself.

To address this, we propose Random Penalty Entropy Loss—a dynamic regularization technique that adaptively scales entropy penalties throughout training. Unlike fixed regularization approaches, our method improves calibration and generalization without relying on post-hoc adjustments or hyperparameter tuning.

3. Materials and methods

3.1. Data composition

3.1.1. Internal dataset

The internal dataset comprises 5674 chest X-ray (CXR) images collected from three South Korean hospitals between 2011 and 2023: Hallym University Chuncheon Sacred Heart Hospital (263 images), Gangneung Asan Hospital (1986 images), and Kangwon National University Hospital (3451 images). All images were acquired to assess the placement of a nasogastric (NG) tube.

Tube placement labels were obtained from three board-certified radiologists. Each case was independently annotated by two of the three readers, resulting in two labels per examination. Disagreement between the two radiologists was rare (79 out of 5674 examinations; 1.4%), indicating high inter-rater consistency of the internal labels. All discrepant cases were subsequently re-reviewed and, by consensus among the three radiologists. Classification labels were defined as either correct placement or misplaced placement. Additionally, segmentation masks were created for three regions of interest: the NG tube, the NG tube tip, and the background.

Of the total 5674 images, 5326 were labeled as having the correct placement and 348 as having a misplaced placement, reflecting a class

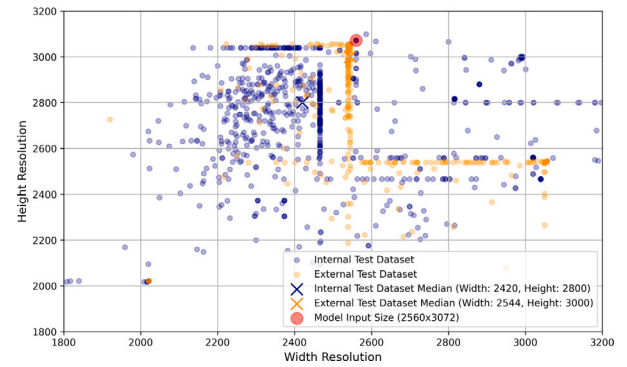


Fig. 1. Resolution distribution of Internal and External datasets with median values and model input size (2560 × 3072) highlighted.

imbalance ratio of approximately 15:1. This natural imbalance was preserved in the training, validation, and test splits to reflect realistic clinical distributions and evaluate performance under skewed label conditions.

For model evaluation, 20% of the dataset (1115 images) was designated as the test set. Importantly, all test images were collected from 2011 to 2020, providing a temporal separation from the training and validation sets. This strategy enables a rigorous assessment of the model's generalization performance over time. The remaining 80% (4559 images) was further divided into a training set (3653 images) and a validation set (906 images) using an 8:2 ratio. A summary of the dataset composition is provided in Table 1.

3.1.2. External dataset

The external dataset was curated from the publicly available MIMIC-CXR database (Johnson et al., 2019) through a three-stage pipeline designed to ensure clinical relevance and label quality:

- **Report-based Pre-selection:** Radiology reports were filtered using NG tube-related keywords (e.g., “NG tube”, “NGT”, and “nasogastric tube”), resulting in approximately 17,000 candidate studies.
- **Automatic Report Classification:** We utilized the Llama3-70B-Instruct model (Meta AI Research, 2024) with a customized system prompt to classify each report into one of five categories: normal, abnormal, borderline, irrelevant, and uncertain. Only studies categorized as normal or abnormal were retained for downstream processing.
- **Manual Annotation:** Initial tube placement labels were created by trained nursing students. A radiologist subsequently reviewed all of these labels, and cases with unclear or ambiguous findings were re-evaluated and adjudicated by two radiologists in consensus to obtain the final reference label. Segmentation of the NG tube tip was excluded due to limited visibility in a substantial number of cases.

Following this process, we obtained 1354 chest X-ray images, comprising 1240 with correct placement and 114 with misplaced placement, yielding an approximate class ratio of 11:1. This external dataset was used to evaluate the model's generalization performance on unseen data acquired from different clinical settings.

3.1.3. Dataset comparison and analysis

Figs. 1 and 2 present the distributions of image resolution and signal-to-noise ratio (SNR) for both internal and external datasets. The internal dataset has a median resolution of 2420 × 2800 pixels, whereas the external dataset exhibits a slightly higher median resolution of 2544 × 3000 pixels. Both distributions fall within the input size of

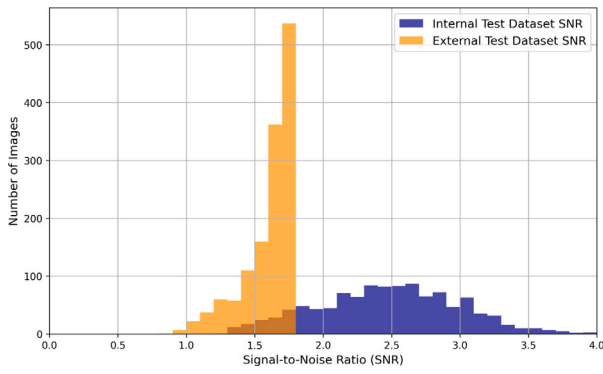
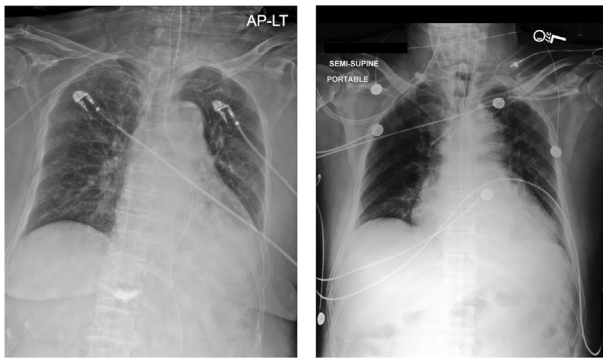


Fig. 2. Signal-to-Noise Ratio (SNR) distribution for the Internal and External datasets.



(a) Internal Data

(b) External Data

Fig. 3. Comparison of Internal (SNR: 2.94) and External (SNR: 1.54) dataset images.

2560 × 3072 pixels used in our model, ensuring compatibility without resizing artifacts.

The SNR analysis reveals notable differences between the datasets. The internal dataset demonstrates consistently higher SNR values, reflecting controlled imaging protocols across the contributing hospitals. In contrast, the external dataset, sourced from the publicly available MIMIC-CXR database, shows greater variability and generally lower SNR levels, indicative of diverse imaging conditions and equipment variability across institutions.

Fig. 3 illustrates representative samples from each dataset, highlighting the visual differences in noise characteristics. For instance, a sample from the internal dataset exhibits a signal-to-noise ratio (SNR) of 2.94. In contrast, a counterpart from the external dataset presents an SNR of 1.54. These variations underscore the necessity of evaluating model performance under heterogeneous imaging conditions to ensure robust generalization and clinical reliability.

3.2. Model architecture

Our model is built upon the nnUNet framework (Isensee et al., 2021), a widely adopted and high-performing baseline for medical image segmentation. While nnUNet demonstrates robust segmentation performance across diverse tasks, we extend it into a unified architecture that simultaneously performs NG tube segmentation and placement classification. This design allows the model to provide both interpretable segmentation maps and diagnostic classifications within a single framework, enhancing its clinical utility.

As illustrated in Fig. 4, the architecture consists of a 10-layer encoder and a 9-layer decoder. The encoder employs stacked convolutional blocks to extract spatial features from the input chest X-ray hierarchically. These features are then decoded into NG tube segmentation masks using transposed convolutions and skip connections, enabling fine-grained anatomical reconstruction.

To support accurate classification under challenging anatomical variations and subtle tube placements, we incorporate a Feature Aggregation Network (FAN) branch. Unlike traditional Feature Pyramid Networks (FPNs) (Lin et al., 2017a), which use complex top-down and lateral fusion paths, FAN offers a lightweight and practical alternative by aggregating multi-scale features from the deeper encoder layers via Global Average Pooling (GAP) (Lin et al., 2014). This design efficiently captures both fine-grained tube structures and broader anatomical context without introducing substantial architectural overhead.

Another key advantage of our model is that it processes high-resolution image patches (2560 × 3072 pixels), allowing analysis to be performed at the original resolution of chest X-rays without any downscaling. Unlike prior methods (Ronneberger et al., 2015) that rely on resized inputs or divide images into smaller patches, our approach preserves the anatomical continuity of the entire NG tube. This design enables the model to maintain complete contextual information from the insertion site to the tip, which is critical for accurate and reliable placement classification.

The unified model seamlessly integrates the segmentation and classification branches. The segmentation branch not only generates interpretable masks but also guides the classification network by acting as a spatial attention mechanism, emphasizing regions relevant to tube placement. This integrated structure enhances performance across varying image qualities and improves model explainability, supporting more reliable clinical deployment.

3.3. Training pipeline

The training pipeline is designed to optimize the unified model for simultaneous NG tube segmentation and placement classification. It consists of preprocessing, extensive data augmentation, and a multi-task loss function that collectively enhances robustness and generalization. An overview of the pipeline is provided in Algorithm 1.

3.3.1. Preprocessing and augmentation

All input chest X-ray images are center-cropped to a fixed resolution of 2560 × 3072 pixels. This size is selected to fully capture the entire NG tube trajectory while maintaining its fine structural details. Preserving the tube's full anatomical course, from nasal entry to gastric tip, is crucial for both accurate segmentation and placement classification.

To further improve generalization, we apply a diverse set of data augmentation techniques that simulate realistic clinical variations:

- **Geometric variations:** With probability 0.2, we apply random rotations around the image center, where the rotation angle θ is sampled uniformly from $[-15^\circ, 15^\circ]$. In addition, isotropic scaling is applied with probability 0.2, with the scale factor sampled from $[0.7, 1.4]$. These transforms simulate differences in patient posture and projection angles.
- **Imaging artifacts:** To emulate realistic acquisition artifacts, we add zero-mean Gaussian noise with probability 0.1, where the noise variance is sampled uniformly from $[0, 0.1]$ in normalized intensity units. Gaussian blur is applied with probability 0.2 using a blur kernel whose standard deviation σ is sampled from $[0.5, 1.0]$ pixels. To simulate reduced spatial resolution, we further use a low-resolution transform with probability 0.25, in which the image is downsampled by a scale factor sampled from $[0.5, 1.0]$ and then upsampled back to the original patch size. Together, these operations mimic noise, motion blur, and resolution loss typically observed in real-world images.

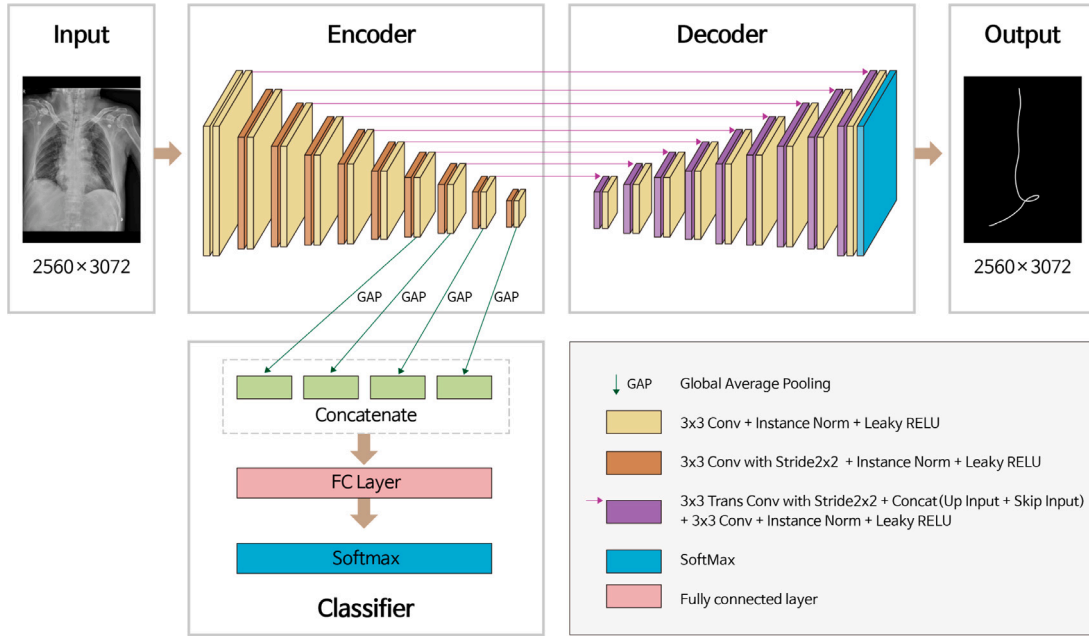


Fig. 4. The proposed unified model architecture for NG tube placement analysis, combining segmentation and classification tasks. The model uses a 10-layer encoder and a 9-layer decoder, with Global Average Pooling (GAP) layers aggregating multi-scale spatial features for classification. High-resolution input patches (2560×3072) ensure accurate NG tube segmentation and classification.

- **Intensity variations:** Multiplicative brightness augmentation is applied with probability 0.15, with the intensity multiplier sampled from [0.75, 1.25]. Contrast is perturbed independently with probability 0.15, using a contrast factor sampled from [0.75, 1.25] while preserving the overall intensity range. Gamma correction is applied in two variants with γ sampled from [0.7, 1.5]: one variant (apply probability 0.1) includes intensity inversion before and after gamma correction, and the other (apply probability 0.3) applies gamma without inversion. These intensity variations mimic the variability introduced by different image acquisition settings and hardware.
- **Symmetry variations:** Random horizontal flipping is enabled along the left-right axis with probability 0.5, reflecting anatomical symmetry and aligning with the test-time augmentation protocol.

These augmentations enable the model to learn more invariant and robust representations, improving performance across both internal and external datasets.

3.3.2. Unified model loss function

The unified model is optimized using a multi-task loss function that jointly considers segmentation performance, classification performance, and regularization via entropy penalization. The overall objective function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{penalty}} \quad (1)$$

Each component contributes to a distinct aspect of model training:

- **Segmentation Loss ($\mathcal{L}_{\text{seg}}$):** A hybrid of Dice loss ($\mathcal{L}_{\text{dice}}$) and pixel-wise cross-entropy loss ($\mathcal{L}_{\text{ce}}$) supervises the segmentation branch. Multi-scale outputs from all decoder layers are aggregated to promote hierarchical feature learning and training stability:

$$\mathcal{L}_{\text{seg}}(S) = \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}(S) + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}}(S), \quad (2)$$

where S denotes the set of segmentation outputs at different decoder depths.

Algorithm 1: Training pipeline for NG tube classification and segmentation

Input: Original X-ray image I

Output: Trained model parameters θ

Preprocess: Center-crop I to resolution 2560×3072 .

foreach training epoch do

Augment: Apply the following augmentations to I :

- Random rotation, scaling
- Gaussian noise, blur
- Brightness, Contrast adjustment
- Low-resolution simulation
- Gamma correction
- Random mirroring

Forward: Input augmented image I_a to the model $\mathcal{M}$:

 Compute output $S, C = \mathcal{M}(I_a)$

Loss Computation:

- Segmentation loss:
 $\mathcal{L}_{\text{seg}}(S) = \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}}$
- Classification loss:
 $\mathcal{L}_{\text{cls}}(C) = \mathcal{L}_{\text{ce}}$
- Penalty Entropy loss:
 $\mathcal{L}_{\text{penalty}}(C) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{\text{class}} C_{ij} \log(C_{ij})$
- Sample penalty scaling factor:
 $\lambda \sim \mathcal{U}[a, b]$
- Total loss:
 $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}(S) + \mathcal{L}_{\text{cls}}(C) + \lambda \mathcal{L}_{\text{penalty}}(C)$

Backward: Update model parameters θ :

$\theta \leftarrow \theta - \alpha \nabla \mathcal{L}_{\text{total}}$

end

return θ

- **Classification Loss ($\mathcal{L}_{\text{cls}}$):** A standard cross-entropy loss is used to supervise the classification branch, guiding the model to distinguish between correct and misplaced NG tube placements.

- **Penalty Entropy Loss** ($\mathcal{L}_{\text{penalty}}$) : To discourage overconfident predictions and enhance generalization, an entropy-based regularization term is incorporated. It is defined as:

$$\mathcal{L}_{\text{penalty}}(C) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K C_{ij} \log(C_{ij}), \quad (3)$$

where C_{ij} is the predicted probability for class j in sample i , and K is the number of classes.

3.3.3. Random Penalty Entropy Loss

We propose the Random Penalty Entropy (RPE) loss, which varies the strength of entropy regularization across training epochs. Instead of using a fixed coefficient λ for the penalty entropy term, RPE introduces an epoch-dependent coefficient λ_t . At the beginning of each epoch t , λ_t is resampled from a uniform distribution. Substituting this into the unified objective in Eq. (1), the epoch-wise total loss can be written as

$$L_{\text{total}}^{(t)} = L_{\text{seg}} + L_{\text{cls}} + \lambda_t L_{\text{penalty}}, \quad (4)$$

where $\lambda_t \sim \mathcal{U}[a, b]$ is drawn from a uniform distribution at the beginning of each epoch.

Intuitively, the RPE loss controls how strongly the entropy regularizer influences the gradients of the classification head at each epoch. When λ_t is close to the lower bound a , the contribution of L_{penalty} to the classification branch is small, and the standard cross-entropy term dominates the gradients of the logits z in the classification head.

This *weak-regularization* regime behaves similarly to a conventional classifier and allows the network to make very confident predictions once its features become reliable.

In contrast, when λ_t is close to the upper bound b , the entropy term is strongly amplified: the gradient component

$$\lambda_t \frac{\partial L_{\text{penalty}}}{\partial z} \quad (5)$$

becomes large and explicitly pushes the logits toward higher-entropy probability distributions, suppressing extremely peaked outputs such as $[0.99, 0.01]$ unless the input evidence is apparent. Training with randomly sampled λ_t therefore alternates the classifier between weak- and strong-regularization regimes during optimization.

Compared with using a single fixed coefficient λ , this stochastic modulation has two advantages. A small fixed λ effectively behaves like a nearly unregularized classifier and may not sufficiently discourage overconfident predictions under severe class imbalance. Conversely, a large fixed λ enforces strong entropy regularization at all times, which can over-smooth the predicted probabilities, hinder convergence, and ultimately degrade final accuracy. By resampling $\lambda_t \sim \mathcal{U}[a, b]$ at every epoch, the proposed RPE loss exposes the model to a continuum of regularization strengths instead of committing to a single value, without requiring a hand-crafted scheduling policy. This stochastic averaging over λ acts as a simple form of noise injection in probability space and leads to better-calibrated and more robust solutions than both the non-regularized baseline and the fixed- λ counterpart.

3.3.4. Training mechanism

During each training epoch, augmented input images are passed through the unified model to produce segmentation outputs (S) and classification logits (C). The total loss $\mathcal{L}_{\text{total}}$ is then minimized via gradient-based optimization:

$$\theta \leftarrow \theta - \alpha \nabla \mathcal{L}_{\text{total}}, \quad (6)$$

where θ denotes the model parameters and α is the learning rate.

The training process leverages the complementary nature of the two tasks: spatial localization via segmentation and global placement assessment via classification. By jointly optimizing these objectives within a single framework, the model learns both fine-grained anatomical details and holistic placement patterns. This integrated training mechanism enhances robustness and generalization, particularly across datasets with varying image quality and clinical characteristics.

Algorithm 2: Inference pipeline for NG tube segmentation and classification

input: Original X-ray image I
output: Segmentation map S and classification result C
Test-time Augmentation: Generate augmented set $\mathcal{I} = \{I, I_v, I_h, I_{vh}\}$, where I_v , I_h , and I_{vh} denote vertical, horizontal, and combined flips.
foreach $I_a \in \mathcal{I}$ **do**
 Divide I_a into overlapping patches:
 $\mathcal{P}_a = \{P \mid P \subset I_a, \text{resolution } 2560 \times 3072\}$
 foreach $P_k \in \mathcal{P}_a$ **do**
 Compute segmentation output: $S_k = \text{Segment}(P_k)$
 Compute classification logits: $C_k = \text{Classify}(P_k)$
 end
 Patch-wise Aggregation:
 • Segmentation : $S_a = \sum_k w_k S_k$ (Gaussian-weighted sum)
 • Classification: $\bar{C}_a = \frac{1}{|\mathcal{P}_a|} \sum_k C_k$
end
Augmentation-level Aggregation $\mathcal{I}$:
 • Final segmentation map: $S = \frac{1}{|\mathcal{I}|} \sum_a S_a$
 • Final classification logits: $\bar{C} = \frac{1}{|\mathcal{I}|} \sum_a \bar{C}_a$
 • Final classification result: $C = \text{Softmax}(\bar{C})$
return S, C

3.4. Inference pipeline

The inference pipeline is designed to ensure accurate and robust NG tube segmentation and classification. As outlined in Algorithm 2, it integrates test-time augmentations (TTA) and patch-based analysis.

For each input X-ray image I , the TTA set $\mathcal{I} = \{I, I_v, I_h, I_{vh}\}$ is generated, where I_v , I_h , and I_{vh} represent vertical, horizontal, and combined flips, respectively. This TTA strategy enhances robustness by exposing the model to multiple spatial variants of the input.

Each augmented image $I_a \in \mathcal{I}$ is divided into overlapping patches $\mathcal{P}_a = \{P \mid P \subset I_a; \text{resolution } 2560 \times 3072\}$, ensuring the NG Tube is entirely encompassed within at one patch. The segmentation and classification tasks are then performed concurrently for each patch $P_k \in \mathcal{P}_a$.

- **Segmentation:** The model generates patch-wise segmentation outputs $S_k = \text{Segment}(P_k)$, identifying the NG Tube region within each patch.
- **Classification:** The model calculates patch-wise classification logits $C_k = \text{Classify}(P_k)$ providing confidence scores for correct placement or misplaced placement of the NG tube.

Subsequently, patch-level outputs are aggregated within each augmented image:

- **Segmentation aggregation:** The patch-wise segmentation maps S_k are combined using a Gaussian-weighted sum:

$$S_a = \sum_k w_k S_k, \quad (7)$$

where w_k denotes the Gaussian weight for patch P_k , assigning higher weights to the central region of each patch to reduce boundary artifacts.

- **Classification aggregation:** The patch-wise classification logits C_k are averaged to obtain the aggregated logits for the augmented image:

$$\bar{C}_a = \frac{1}{|\mathcal{P}_a|} \sum_k C_k. \quad (8)$$

Finally, the outputs across all augmentations are aggregated to produce the final results:

- **Segmentation:** Segmentation maps S_a are averaged across all augmentations:

$$S = \frac{1}{|I|} \sum_a S_a. \quad (9)$$

- **Classification:** Aggregated classification logits $\bar{C}_a$ are averaged to compute the final classification logits:

$$\bar{C} = \frac{1}{|I|} \sum_a \bar{C}_a. \quad (10)$$

A softmax function is applied to $\bar{C}$ to obtain the final predicted label:

$$C = \text{Softmax}(\bar{C}). \quad (11)$$

By combining patch-based inference with augmentation ensembling, the proposed pipeline enhances robustness and maintains high performance across varied image conditions and anatomical presentations.

4. Results and discussion

4.1. Experiment setting

4.1.1. Training details

Model training was conducted using eight NVIDIA A6000 GPUs, each equipped with 48 GB of memory, allowing efficient processing of high-resolution chest X-ray images. Inference was performed on a single NVIDIA RTX 3090 GPU, demonstrating compatibility with widely accessible hardware configurations. The end-to-end processing time for a single chest radiograph was 0.558 s on average with a peak GPU memory usage of 7208 MB, indicating that near real-time analysis is feasible in typical clinical settings.

The key training hyperparameters were as follows:

- **Optimizer:** Stochastic Gradient Descent (SGD) with a weight decay of 3×10^{-5} , momentum of 0.99, and Nesterov acceleration.
- **Learning Rate:** An initial learning rate of 1×10^{-2} was used and decayed over time using a polynomial learning rate scheduler.
- **Batch Size:** A batch size of 16 was employed, balancing computational efficiency and training stability.
- **Epochs:** The model was trained for 1000 epochs, with 250 mini-batches per epoch.
- **Entropy regularization:** For the fixed penalty entropy loss, the coefficient was set to $\lambda = 1$ whereas for the RPE loss λ_i was sampled from $\mathcal{U}[0.5, 1.2]$ at each epoch.

4.1.2. Threshold selection

Classification thresholds for each model were determined on the internal validation set by selecting the probability cutoff that maximized the F1 score, and the resulting model-specific thresholds were then kept fixed for all subsequent experiments. Accordingly, the same predefined threshold was used for each model to report performance on both the internal test set and the external MIMIC-CXR dataset. In clinical deployment, a fixed threshold would likewise be applied consistently. For completeness, we also examined the clinical utility of this fixed operating point using decision-curve analysis on the validation set (Fig. 5). the F1-optimized threshold ($p_i = 0.459$) lies within a region of positive benefit, indicating that the chosen fixed threshold is clinically reasonable and was therefore used unchanged for the internal test and external MIMIC-CXR evaluations.

For segmentation, a fixed probability threshold of 0.5 was applied, consistent with the default configuration in nnUNet (Isensee et al., 2021).

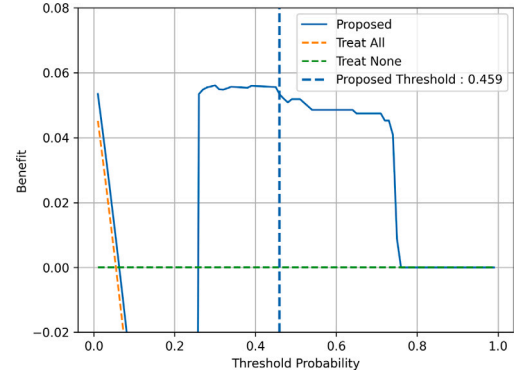


Fig. 5. Decision-curve analysis on the validation dataset.

4.1.3. Evaluation metrics and clinical interpretation

We report AUROC and AUPRC to assess global discrimination, and F1 score, balanced accuracy, sensitivity, and specificity at a fixed decision threshold. AUROC summarizes how well the model separates misplaced from correctly placed tubes over all possible thresholds. In contrast, AUPRC emphasizes performance on the clinically critical but rare “misplaced” class in this imbalanced setting. Because clinical use typically relies on a prespecified decision threshold, these threshold-based metrics are more directly relevant to practice. Among them, we use the F1 score as the primary evaluation metric, as it jointly reflects the risk of missing truly misplaced tubes (false negatives, with potential patient harm) and the burden of unnecessary alarms for correctly placed tubes (false positives). Sensitivity quantifies the proportion of misplacements that are correctly detected, specificity quantifies the proportion of correctly positioned tubes, and balanced accuracy provides a single summary that gives equal weight to both classes. Across internal and external cohorts, all point estimates are accompanied by two-sided 95% confidence intervals computed via non-parametric, study-level bootstrap.

For tube localization, we additionally report the Dice similarity coefficient to measure the spatial overlap between the predicted tube region and the reference annotation.

4.1.4. Comparison models and ablation design

To jointly serve as an ablation study, all models were trained and evaluated under *identical* settings (datasets, tube-only masks, preprocessing, backbone/segmentation branch, training schedule, and the fixed F1-optimized threshold). We vary **one factor at a time**: (i) the *pipeline architecture* (two-stage vs. unified single-stage), and (ii) the *classification loss/regularization* within the unified architecture. This design lets us attribute performance differences in Table 2 to the ablated component rather than to confounders.

Wang 2-Stage (architectural ablation: two-stage). A modified two-stage framework inspired by the NG tube placement model of Wang et al. (2024). The segmentation network used in this model is architecturally identical to the segmentation branch of our proposed unified model. Its output mask is concatenated with the original X-ray before being passed to a separate ResNet-50 classifier (Henderson et al., 2021) for placement classification. Unlike Wang et al. who also segmented the diaphragm and lung, our implementation focuses solely on the NG tube due to limitations in the dataset. While this design benefits from spatial information, it lacks joint optimization and is vulnerable to error propagation from segmentation inaccuracies.

Unified Baseline (loss ablation: cross-entropy). A unified single-stage model trained with standard cross-entropy on the classification branch. This serves as the reference variant for loss-function ablations.

Unified Focal (loss ablation: focal, $\gamma = 2$). Replaces cross-entropy with focal loss to emphasize hard/rare (misplaced) examples and mitigate class imbalance.

Table 2

Classification and segmentation performance on the internal (top) and external (bottom) test datasets. Values are point estimates with 95% confidence intervals in brackets. Metrics include F1 score, AUROC, AUPRC, Balanced Accuracy (Bal-Acc), Sensitivity, Specificity, Tube DICE, and Tip DICE. Best point estimates are bolded.

Internal test dataset								
Model	F1	AUROC	AUPRC	Bal-Acc	Sensitivity	Specificity	Tube DICE	Tip DICE
Wang	87.39	97.46	93.51	90.07	82.06	98.09	85.48	62.45
2-Stage	[82.94, 91.49]	[94.33, 99.31]	[89.95, 96.61]	[85.53, 94.55]	[73.02, 90.77]	[97.27, 98.86]		
Unified	93.91	99.00	96.84	90.89	81.98	99.81	86.14	58.30
Baseline	[90.45, 96.89]	[97.80, 99.79]	[94.46, 98.81]	[86.06, 95.32]	[72.22, 90.77]	[99.52, 1.00]		
Unified	93.81	99.28	97.00	93.88	88.44	99.32	86.65	60.82
Focal	[90.52, 96.74]	[98.46, 99.81]	[94.57, 98.86]	[89.50, 97.52]	[79.56, 95.60]	[98.78, 99.81]		
Unified	93.25	98.26	96.56	89.29	78.68	99.90	86.41	62.06
Smoothing	[89.55, 96.44]	[95.03, 99.78]	[93.57, 98.62]	[84.06, 94.27]	[68.37, 88.71]	[99.71, 1.00]		
Unified	94.52	94.96	95.33	92.55	85.39	99.71	86.41	58.83
Entropy	[91.14, 97.32]	[89.07, 99.45]	[91.53, 98.39]	[87.78, 96.70]	[75.86, 93.76]	[99.33, 1.00]		
Proposed	93.94	97.20	95.60	90.96	82.11	99.81	86.56	60.50
	[90.46, 96.87]	[94.18, 99.59]	[92.34, 98.20]	[86.13, 95.47]	[72.41, 91.23]	[99.52, 1.00]		
External test dataset								
Model	F1	AUROC	AUPRC	Bal-Acc	Sensitivity	Specificity	Tube DICE	Tip DICE
Wang	59.85	82.82	67.32	76.01	76.33	75.69	79.04	–
2-Stage	[56.76, 63.03]	[78.55, 86.95]	[63.03, 72.23]	[71.91, 79.84]	[68.53, 83.77]	[73.44, 78.04]		
Unified	49.22	74.54	60.99	67.63	75.49	59.77	77.80	–
Baseline	[46.70, 51.88]	[70.03, 78.91]	[57.52, 65.11]	[63.36, 71.66]	[67.79, 82.91]	[57.00, 62.46]		
Unified	51.46	79.58	69.18	69.52	75.58	63.45	79.37	–
Focal	[48.80, 54.27]	[74.86, 83.89]	[64.35, 74.07]	[65.43, 73.47]	[67.35, 83.19]	[60.78, 66.07]		
Unified	52.17	78.85	62.18	71.91	80.72	63.10	77.87	–
Smoothing	[49.33, 54.77]	[74.49, 83.01]	[58.76, 66.27]	[68.06, 75.57]	[73.50, 87.50]	[60.40, 65.73]		
Unified	59.54	81.92	65.46	75.78	76.31	75.26	77.22	–
Entropy	[56.69, 62.56]	[77.72, 85.83]	[61.43, 69.94]	[71.82, 79.67]	[68.75, 83.93]	[72.82, 77.52]		
Proposed	66.34	84.82	69.84	79.28	74.67	83.89	79.37	–
	[62.79, 69.70]	[80.70, 88.70]	[65.38, 74.58]	[74.80, 83.46]	[66.02, 82.54]	[81.88, 85.85]		

Unified Smoothing (loss ablation: label smoothing, $\alpha = 0.1$). Applies label smoothing to reduce overconfident predictions and improve calibration under label noise.

Unified Entropy (loss ablation: fixed entropy penalty, $\lambda = 1$). Augments cross-entropy with an entropy-penalty term of fixed strength, encouraging less peaked outputs without introducing stochasticity.

Proposed (loss ablation: random entropy penalty, $\lambda \sim \mathcal{U}[0.5, 1.2]$). Combines the unified architecture with *Random Penalty Entropy* (RPE): an entropy-based regularizer whose coefficient is resampled once per epoch. This stochastic scaling acts as probability-space noise injection, improving calibration and robustness under distribution shift compared with the fixed- λ counterpart.

With all non-ablated factors held constant, this comparative framework quantifies how unified modeling, regularization strategy, and stochastic scaling each contribute to the robustness and reliability of NG-tube placement analysis.

4.2. Quantitative performance analysis

Table 2 summarizes the classification and segmentation results of all models on the internal (top) and external (bottom) test datasets. Values are reported as point estimates with 95% confidence intervals in brackets.

Internal test dataset results. All unified model variants significantly outperformed the Wang 2-Stage model across classification metrics. The Unified Baseline model achieved an F1 score of 93.91% [90.45, 96.89], AUROC of 99.00% [97.80, 99.79], and AUPRC of 96.84% [94.46, 98.81], demonstrating the effectiveness of joint optimization of segmentation and classification tasks.

The Unified Focal model improved sensitivity to 88.44% [79.56, 95.60], while maintaining high AUROC and AUPRC, validating the benefit of emphasizing difficult examples in imbalanced datasets. The Unified Smoothing model achieved the highest specificity (99.90% [99.71, 100]), effectively reducing false positives. The Unified Entropy model attained the best overall F1 score (94.52%, [91.14, 97.32]), indicating enhanced robustness against misclassification.

The proposed model, which integrates Random Penalty Entropy Loss, achieved performance comparable to the Unified Baseline (F1: 93.94% [90.46, 96.87]), suggesting that stochastic regularization does not degrade classification accuracy on familiar (seen) data.

External test dataset results. On the external test set, performance trends changed significantly. Unified Baseline, Focal, and Smoothing models exhibited notable degradation, likely due to overfitting to the internal data distribution. Interestingly, the Wang 2-Stage model, although weaker on internal data, maintained stable generalization (F1: 59.85% [56.76, 63.03], AUROC: 82.82% [78.55, 86.95]). Given the Wang 2-Stage model's high performance on the internal dataset, its performance on the external dataset is likely attributable to domain shift rather than to architectural exclusion of auxiliary anatomical segmentation (e.g., lungs or diaphragm).

In contrast, the Proposed model consistently outperformed all others across all external evaluation metrics. It achieved the best F1 score (66.34% [62.79, 69.70]), AUROC (84.72% [80.70, 88.70]), AUPRC (69.84% [65.38, 74.58]), and specificity (83.89% [81.88, 85.85]) while preserving high sensitivity (74.67% [66.02, 82.54]). These results highlight the effectiveness of entropy-based regularization with random scaling in enhancing model robustness under distributional shift.

In summary, while the unified architecture improves overall performance in controlled environments, its ability to generalize depends heavily on the design of the loss function. The Proposed model balances accuracy, calibration, and robustness across internal and external domains, clearly demonstrating its superiority over both previous unified variants and the state-of-the-art Wang 2-Stage model.

4.3. Visual and qualitative analysis

This section presents a visual and qualitative analysis to complement the quantitative evaluation. We examine representative cases from both internal and external test datasets to assess how the proposed model and its variants perform under challenging or ambiguous conditions. The analysis focuses on three key aspects: (1) comparison with the Wang 2-Stage baseline, (2) behavioral differences across unified

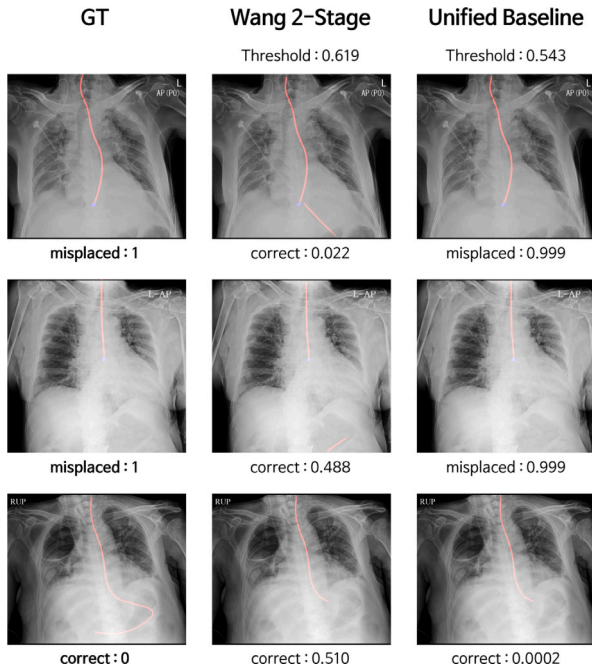


Fig. 6. Comparison of the Wang 2-stage and Unified Baseline on internal test data. The classification threshold used by each model is indicated below the model name. The first column shows the ground truth segmentation and classification, while the second and third columns display the predictions from the Two-Stage and Basic Models, respectively. Note that the classification result and confidence score are given below each image.

model variants, (3) failure case analysis to clarify the model's clinical reliability boundaries, and (4) generalization under domain shift. These comparisons elucidate the impact of architectural design and loss function selection on prediction reliability, confidence, and segmentation quality in real-world scenarios.

4.3.1. Internal test analysis – Comparison with Wang 2-Stage

Fig. 6 shows internal test examples comparing the Wang 2-Stage model and the Unified Baseline. In the first and second rows (misplaced placement), the Wang 2-Stage model produced inaccurate segmentation masks, resulting in classification errors or low-confidence predictions. In contrast, the Unified Baseline yielded more accurate segmentation and confidently predicted the correct label, highlighting the advantages of joint optimization.

In the third row (correct placement), both models exhibited difficulty segmenting the lower portion of the NG tube due to poor visibility. However, the Unified Baseline still produced a confident and correct classification. In contrast, the Wang 2-Stage model's prediction was close to the decision threshold, reflecting greater uncertainty.

4.3.2. Internal test analysis – Unified model variants

Fig. 7 illustrates additional internal cases evaluated across all Unified model variants to analyze classification behavior under ambiguous conditions. Notably, the Unified Baseline model was the only variant to misclassify both challenging cases. In contrast, all other models correctly predicted at least one instance.

The Unified Focal model demonstrated a tendency to classify uncertain cases as misplaced placement, thereby favoring sensitivity. Conversely, the Unified Smoothing, Unified Entropy, and Proposed models frequently predicted correct placement under ambiguity.

A key observation is that most models predicted overly confident probability scores far from their decision thresholds, even when the predictions were incorrect. However, the Unified Entropy and Proposed

models predicted confident probability scores closer to their respective thresholds. This behavior suggests improved calibration and a reduction in overconfidence under uncertainty.

These findings underscore the impact of loss function design on classification behavior. In particular, entropy-based regularization promotes more calibrated decision-making, which is especially critical in safety-sensitive tasks such as NG tube placement verification, where overconfident errors may lead to clinical harm.

4.3.3. Internal test analysis – Failure case analysis

To clarify the clinical reliability boundaries of the proposed model, we analyzed representative misclassified cases (Fig. 8). As shown in Fig. 8(a) and (b), two false-positive cases (correct placement predicted as misplaced) occurred when the nasogastric tube was poorly visible on X-ray, leading to incomplete segmentation and subsequent classification errors. Fig. 8(c) and (d) illustrate two false-negative cases (misplaced placement predicted as correct) involving complex tube trajectories in which the tube initially descended below the diaphragm, satisfying typical criteria for correct placement, but subsequently formed a loop and re-ascended into the thoracic cavity. Because the tube segment before looping appeared normal, the model was likely misled by these locally correct features despite the globally abnormal trajectory.

These findings indicate that low tube visibility and non-monotonic tube paths remain challenging scenarios and should be interpreted with caution in clinical deployment.

4.3.4. External test analysis – Generalization under domain shift

Fig. 9 presents segmentation and classification predictions from the Wang 2-Stage and Unified variant models on several challenging external test cases.

The image in the first row contains an unfamiliar overlap artifact not seen during training. All models mis-segmented the artifact as an NG tube. Despite the artifact's ambiguity, the Wang 2-Stage, Unified Baseline, and Unified Smoothing models classified the case as misplaced with unjustifiably high confidence. In contrast, the Unified Focal and Unified Entropy models also predicted “misplaced placement” but with lower confidence, demonstrating reduced overconfidence in ambiguous scenarios. The proposed model also misclassified the case but produced a probability near the threshold, reflecting calibrated uncertainty.

The second row shows a coiled NG tube under low image quality, which caused segmentation difficulties across all models. While several models correctly classified the case as misplaced, they did so with high confidence despite the uncertainty. It again showed misclassification with restrained confidence, suggesting conservative behavior.

In the third and fourth rows (correct placement cases), segmentation was successful across all models. However, the Unified Baseline, Focal, Smoothing, and Entropy models misclassified these cases as misplaced, often with overconfident predictions. Only the proposed model correctly classified both, demonstrating improved robustness under domain shift.

These findings suggest that the Proposed model maintains robust performance and calibrated behavior in the presence of unfamiliar or degraded inputs. Such caution under domain shift highlights its potential for safer and more reliable deployment in real-world clinical settings.

4.4. Probabilistic behavior and confidence evaluation

This section evaluates the probabilistic outputs and calibration performance of all models across the validation, internal, and external datasets.

We employ Kernel Density Estimation (KDE) to visualize how closely the predicted probability distributions match the ground truth (GT) class distributions, offering insight into class-wise prediction confidence.

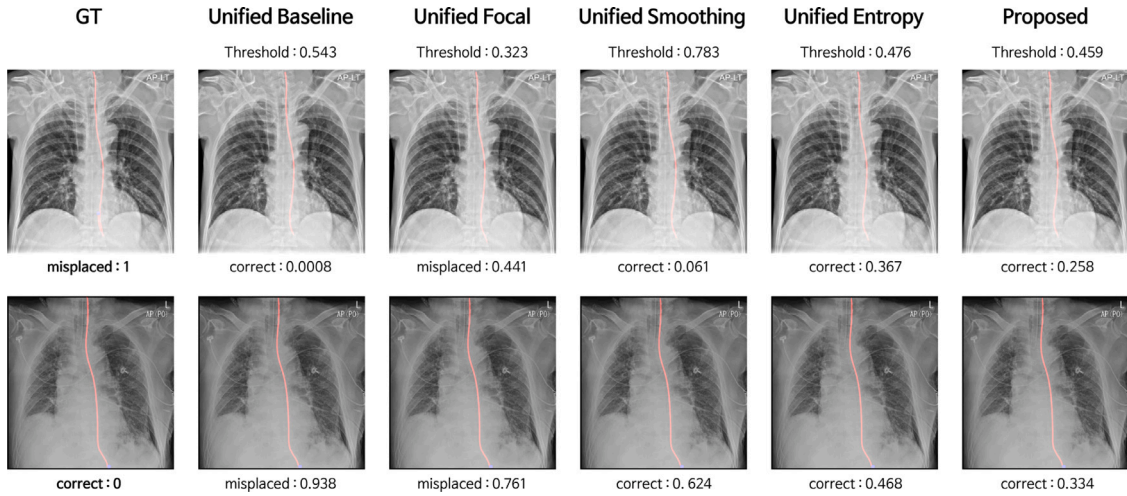


Fig. 7. Comparison of different, unified models on internal test data. The classification threshold used by each model is indicated below the model name. The first column represents the ground truth, while subsequent columns show outputs from the Unified Baseline, Unified Focal, Unified Smoothing, Unified Entropy, and Proposed Models, illustrating their varying segmentation and classification tasks capabilities.

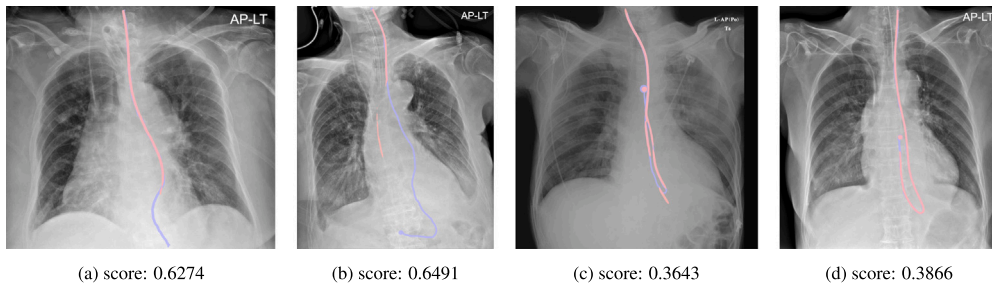


Fig. 8. Failure cases illustrating the clinical reliability boundaries of the proposed model. Blue contours indicate ground-truth NG tube segmentation, and red contours denote model predictions. (a) and (b) show false-positive cases involving correct placement with poor tube visibility. (c) and (d) show false-negative cases involving misplaced tubes with looping above the diaphragm.

Additionally, Reliability Diagrams are used to assess the alignment between predicted confidence scores and actual classification accuracy. These plots provide a quantitative and visual assessment of model calibration, highlighting the generalization capability and identifying potential overconfidence issues that may arise under distribution shifts.

4.4.1. Kernel Density Estimation (KDE)

KDE plots (Fig. 10) visualize the distribution of predicted classification probabilities for correct and misplaced placement cases across the validation, internal, and external datasets. Gray histograms indicate the ground truth (GT) class distributions for comparison. Dashed vertical lines denote classification thresholds determined from the validation dataset.

The Wang 2-Stage model exhibits well-separated peaks for correct placement and misplaced placement near 0 and 1 in the validation dataset, with class proportions closely aligning with the GT distribution. In the internal dataset, although the peaks remain well-separated, the proportion of correct placement predictions appears slightly underestimated, indicating potential difficulty in maintaining accurate class proportions. In the external dataset, the alignment with the GT class ratio further deteriorates, with the misplaced placement peak being overly emphasized and indicating a deviation from the expected class distribution (approximately 11:1). These observations suggest that, although the Wang 2-Stage model achieves strong performance on familiar data, it struggles to maintain robustness when faced with unseen or shifted data distributions.

The Unified Baseline, Unified Focal, and Unified Smoothing models also show highly concentrated peaks near 0 and 1 in the validation

and internal datasets, reflecting overconfident predictions. Although Unified Focal and Unified Smoothing models slightly alleviate this overconfidence by producing less concentrated peaks, their distributions still fail to capture the correct GT ratio when applied to the external dataset. This suggests that standard regularization methods such as Focal Loss and Label Smoothing are insufficient for maintaining calibration under domain shift.

In contrast, Unified Entropy and Proposed models exhibit notably different probabilistic behaviors. Both models produce smoother, less sharply peaked distributions, often centered around moderate probabilities (e.g., 0.25 and 0.75), indicating reduced overconfidence. Importantly, their predicted probability distributions closely reflect the GT class ratios across all datasets, including the external dataset with its domain shift and severe imbalance. This behavior demonstrates the effectiveness of entropy-based regularization in promoting calibrated confidence outputs in achieving calibrated confidence distributions without sacrificing class discrimination.

The Proposed model showed the most consistent and well-calibrated probability distributions across all datasets. This suggests that introducing stochasticity into the entropy-based regularization discourages overconfident predictions, leading to more reliable behavior under distribution shifts.

These results confirm that the proposed penalty entropy-based regularization, especially when combined with random scaling, effectively mitigates overconfidence, improves calibration, and enhances generalization across internal and external datasets. This property makes the proposed approach especially suitable for clinical deployment, where calibrated and cautious predictions are critical for patient safety.

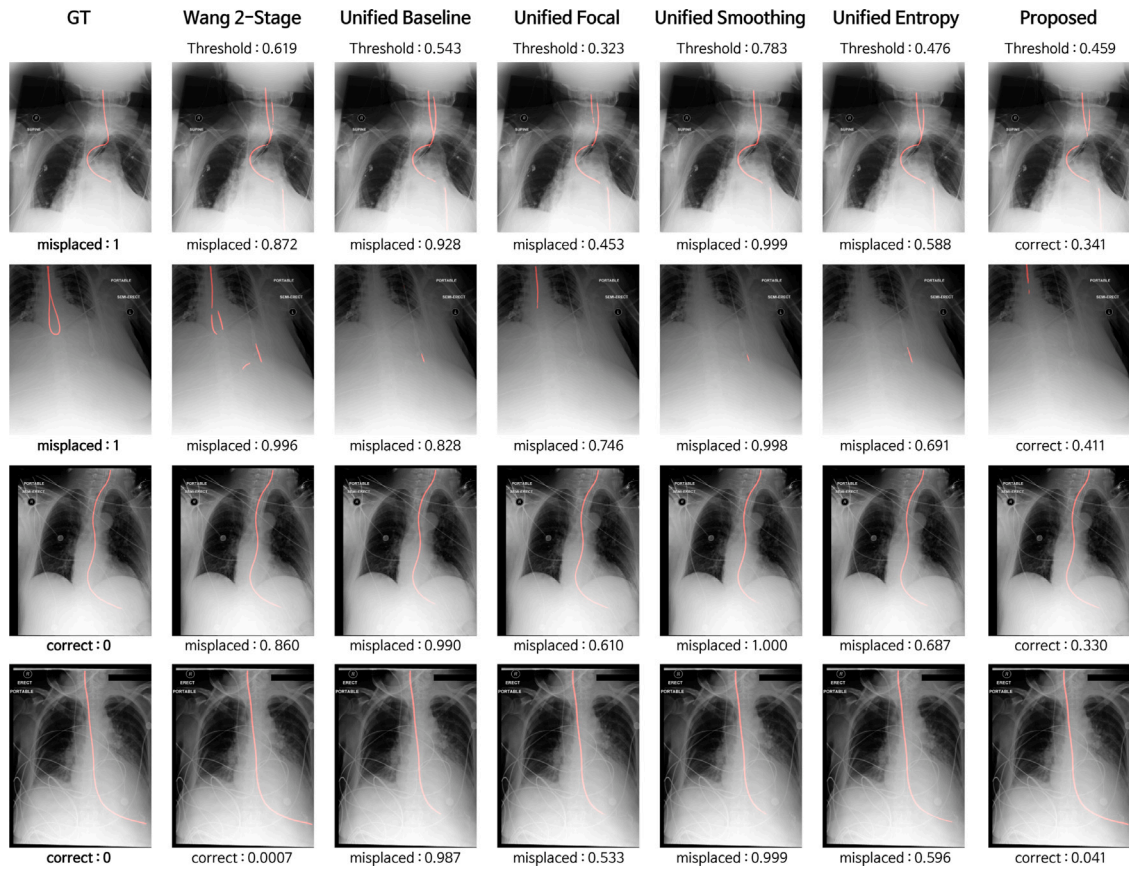


Fig. 9. Segmentation and classification predictions on the external test dataset for the Two-Stage, Basic, Focal, Smoothing, and Proposed Models (with and without Uniform Scaling). The classification threshold used by each model is indicated below the model name. Each row represents challenging NG tube placement scenarios, including misplaced (Rows 1–2) and correct placement (Rows 3–4) cases. The ground truth (GT) segmentation and labels are in the first column.

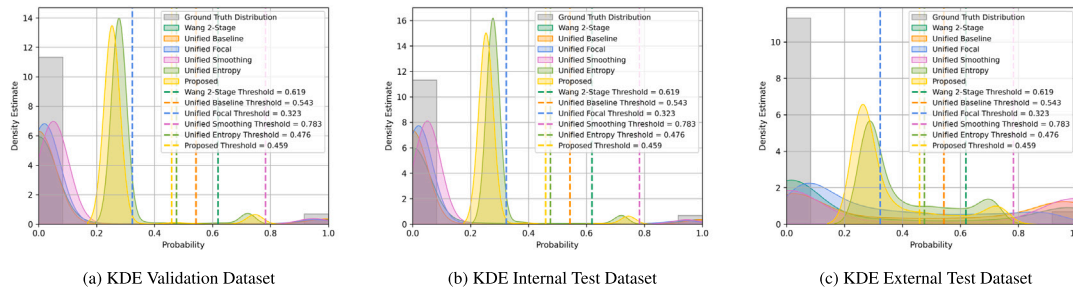


Fig. 10. Kernel Density Estimation (KDE) plots illustrating predicted probability distributions for (a) validation, (b) internal test, and (c) external test datasets. The gray histogram represents the Ground Truth (GT) distribution, highlighting the proportions of correct placement and misplaced placement cases. Dashed vertical lines indicate the optimal classification thresholds for each model, determined from the validation dataset based on the F1 score.

4.4.2. Reliability diagrams

We used reliability diagrams to evaluate calibration performance by comparing predicted confidence levels with actual accuracies across prediction bins. A well-calibrated model produces a reliability curve that closely follows the diagonal, indicating that predicted probabilities correspond well with observed accuracies. Fig. 11 presents the reliability diagrams across the validation, internal, and external datasets, alongside quantitative calibration metrics, as well as two quantitative calibration metrics: Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

The Wang 2-Stage model exhibited inconsistent calibration across datasets. Although the model achieved a low ECE (0.072) on the validation set, its high MCE (0.467) and a notable drop in accuracy at high-confidence bins indicated miscalibration. Calibration further

deteriorated on the internal dataset (ECE: 0.087, MCE: 0.779) and significantly worsened on the external dataset (ECE: 0.374, MCE: 0.959), where the reliability curve showed severe deviations, highlighting its limited adaptability to unseen or heterogeneous data distributions.

The Unified Baseline model exhibited near-perfect calibration on the validation and internal datasets in both the validation (ECE: 0.006, MCE: 0.083) and internal datasets (ECE: 0.010, MCE: 0.115). However, under external conditions, calibration degraded sharply (ECE: 0.373, MCE: 0.900), suggesting overfitting to the training distribution and reduced confidence reliability under domain shift.

The Unified Focal model exhibited improved calibration compared to the baseline. Its ECE and MCE were consistently low in both validation (ECE: 0.020, MCE: 0.065) and internal (ECE: 0.020, MCE: 0.069) evaluations. Notably, it achieved the lowest ECE (0.245) among all models on the external dataset. However, its relatively high MCE

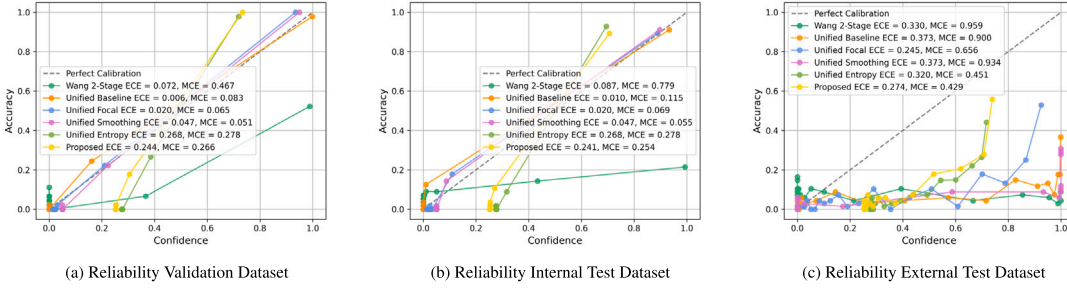


Fig. 11. Reliability diagrams before post-hoc calibration. Panels show (a) validation, (b) internal test, and (c) external test datasets. The dashed diagonal indicates perfect calibration (confidence = accuracy). Legends report each model's Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

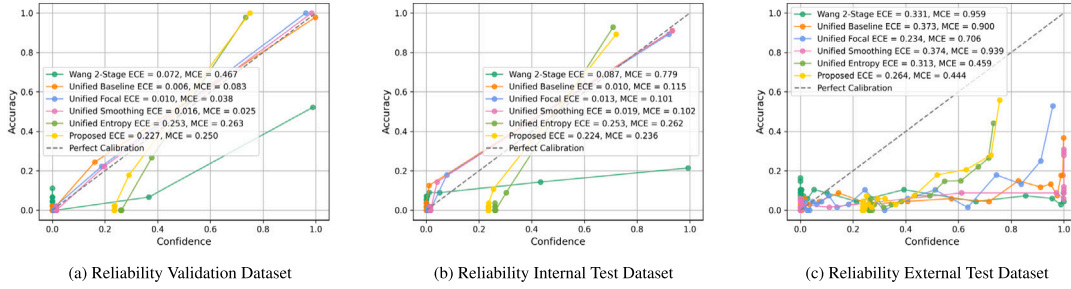


Fig. 12. Reliability diagrams after post-hoc calibration with Temperature Scaling. Panels show (a) validation, (b) internal test, and (c) external test datasets. The dashed diagonal indicates perfect calibration (confidence = accuracy). Legends report each model's Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

(0.656) and fluctuation in high-confidence regions suggested local miscalibrations. The reliability curve exhibited fluctuations—particularly declines in accuracy at high confidence levels—suggesting instability in confidence estimates when the distribution shifts. These findings indicate that while Focal Loss contributes to improved average calibration, it remains vulnerable to localized miscalibrations in high-confidence regions when applied to unseen data.

The Unified Smoothing model exhibited similar trends to the baseline, with marginally higher ECE and MCE values across validation and internal datasets. Its degraded performance on the external dataset (ECE: 0.373, MCE: 0.934) indicates that label smoothing alone is insufficient to improve calibration robustness under distribution shifts.

The Unified Entropy and Proposed models exhibited distinct calibration. In in-distribution settings (validation and internal datasets), both models showed slightly higher ECE and MCE values than the baseline, reflecting the inherent trade-off of promoting generalization and reducing overconfidence. However, under domain shift, both models demonstrated improved calibration, with the Proposed model achieving the best overall performance. The Proposed model achieved the second-lowest ECE (0.274) and the lowest MCE (0.429) among all models. Its reliability curve displayed a smooth, gradual alignment with increasing confidence levels, avoiding the sharp deviations observed in other models.

These findings suggest that although the proposed model shows slightly higher calibration errors in in-distribution data in in-distribution settings, this behavior stems from its conservative probabilistic predictions aimed at reducing overconfidence. As a result, they produce more cautious and calibrated predictions under distribution shifts, which is particularly important in clinical applications where cautious decision-making is critical.

4.4.3. Post-hoc calibration (temperature scaling)

We applied TS to all models. A single temperature was fitted per model on the internal validation set and then kept fixed for the internal

and external evaluations. TS modestly reduced average miscalibration (ECE) without altering case ranking. Crucially, the relative calibration patterns observed pre-TS were preserved post-TS (Fig. 12). Thus, our main conclusions remain robust to post-hoc adjustments.

4.5. Sensitivity analysis of the λ_i sampling range

To further validate the proposed Random Penalty Entropy (RPE) loss, we conducted a sensitivity analysis over the penalty coefficient λ_i sampling range. In addition to the fixed setting (Fixed, $\lambda = 1.0$) and our proposed configuration (Proposed, $\lambda_i \sim \mathcal{U}[0.5, 1.2]$), we trained two additional variants with a narrower range (Narrow, $\lambda_i \sim \mathcal{U}[0.8, 1.0]$) and a wider range (Wide, $\lambda_i \sim \mathcal{U}[0.2, 1.5]$), while keeping all other training settings identical. Fig. 13 summarizes AUROC and ECE on both the internal and external test sets. Overall, Fig. 13 suggests that introducing stochasticity in λ_i can be beneficial. Still, its effect depends strongly on the range of λ_i . On the internal test set, sampling λ_i across a range consistently improves both AUROC and ECE relative to the fixed- λ baseline. However, in an external test, the behavior becomes non-monotonic. The narrow range (Narrow, $\lambda_i \sim \mathcal{U}[0.8, 1.0]$) leads to inferior external performance relative to Fixed, which is consistent with insufficient diversity in regularization strength: the variability is too small to expose the model to diverse training regimes, yet it can still introduce optimization noise that does not translate to improved generalization. In contrast, the wide range (Wide, $\lambda_i \sim \mathcal{U}[0.2, 1.5]$) provides substantial regime diversity. It yields a slight improvement over Fixed on the external test set. Nevertheless, excessively large variability in λ_i may introduce excessive stochasticity into training, limiting additional gains under distribution shift. Taken together, these results indicate that an appropriately chosen λ_i range is critical for achieving robust performance and calibration across institutions; in our experiments, the proposed range ($\lambda_i \sim \mathcal{U}[0.5, 1.2]$) provides the most favorable and consistent behavior on both internal and external test sets.

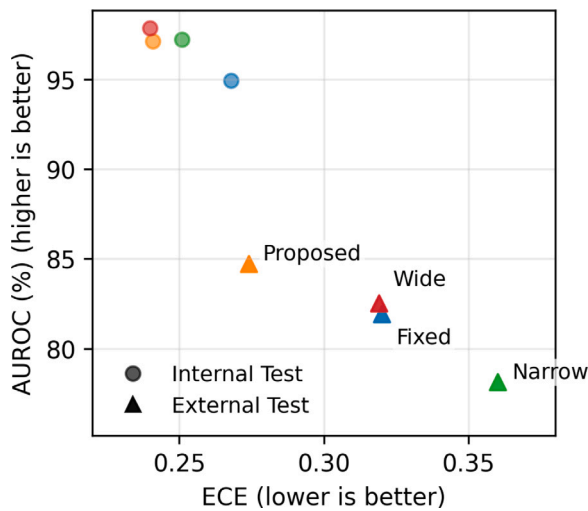


Fig. 13. Sensitivity analysis of RPE to the λ_i sampling range. AUROC versus ECE on the internal and external test sets under four λ_i settings: *Fixed* (blue, $\lambda = 1.0$), *Proposed* (orange, $\lambda_i \sim U[0.5, 1.2]$), *Narrow* (green, $\lambda_i \sim U[0.8, 1.0]$), and *Wide* (red, $\lambda_i \sim U[0.2, 1.5]$).

5. Conclusions

In this study, we proposed a unified deep-learning model capable of performing NG tube placement classification and segmentation simultaneously. The architecture enables end-to-end optimization of both tasks, allowing the classification branch to leverage spatial cues from shared encoder features directly. This integrated design reduces error propagation between tasks and improves learning efficiency. To further enhance robustness, we introduced a novel regularization strategy—Random Penalty Entropy Loss—where entropy-based regularization is dynamically scaled during training to mitigate overconfidence and promote better generalization.

The proposed unified framework outperformed the state-of-the-art Wang 2-Stage model in classification performance on the internal dataset, validating the advantage of joint modeling. While models employing Focal Loss or Label Smoothing showed slightly improved calibration on internal data, their performance degraded significantly on the external dataset, suggesting overfitting to familiar distributions.

In contrast, the proposed model with Random Penalty Entropy Loss consistently maintained both high classification accuracy and reliable calibration across all datasets. On the external test set, it achieved the best balance between performance and reliability, surpassing other unified variants in both the F1 score and maximum calibration error. The use of random scaling further reduced overconfidence, especially in ambiguous or low-quality cases.

Several directions remain for future work that will address the current limitations of this study and broaden its clinical scope. First, the relatively small number of misplaced NG tube cases limited the granularity of our sensitivity analyses. We plan to expand the dataset to include more misplacement data and a broader range of countries, which is expected further to improve the robustness and stability of the model. Second, from a clinical standpoint, ideal NG tube placement requires not only that the tip lie within the stomach but also that all side holes lie within the stomach. However, the longitudinal positions of side holes differ across NG tube products, leading to device-specific definitions of a normal placement range. We therefore plan to extend our framework to incorporate device type or product specifications, enabling the model to learn product-specific criteria for correct placement. Third, we plan to move beyond tube-only analysis by explicitly modeling the position of the NG tube relative to surrounding organs (e.g., lungs, diaphragm, and stomach) through additional anatomical

segmentation or landmark detection. Jointly reasoning about the tube trajectory and nearby structures may enable more precise characterization of malposition patterns and better support clinically meaningful risk stratification.

At the point of care, particularly in emergency and intensive care units, the model could be deployed on bedside or portable CXR systems to provide immediate feedback on NG tube placement, enabling earlier detection of high-risk misplacements even during off-hours. In such a deployment, the AI output would be used as a decision-support tool rather than as a stand-alone reader: When the model's prediction and the interpreting clinician's assessment are concordant, the analysis results are immediately utilized. Whereas discordant cases would be reviewed for additional expert or multidisciplinary discussion before nutrition or medication is administered, establishing a structured safety net around uncertain cases. Together with the model's consistently robust performance and reliable calibration across internal and external datasets, and its ability to generalize under distribution shifts while producing cautious yet accurate predictions, this workflow highlights its strong potential for safe real-world clinical deployment, where safety and reliability are paramount.

CRedit authorship contribution statement

GwiSeong Moon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kyoung Min Moon:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Inseo Park:** Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kanghee Lee:** Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization. **Doohee Lee:** Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Woo Jin Kim:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Yoon Kim:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Ji Young Hong:** Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Hyun-Soo Choi:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics statement

This study was approved by the Institutional Review Boards of Hallym University Chuncheon Sacred Heart Hospital (HUCSHH-2023-12-006-001) and Gangneung Asan Hospital (GNAH-2024-09-003) and Gangwon National University Hospital (2024-10-003). All internal CXR images were fully de-identified before analysis, in accordance with institutional policies and relevant privacy regulations. Explicit identifiers and DICOM header fields containing personal information were removed so that the data could not be traced back to individual patients by the researchers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2021-KH114109) and Chung-Ang University Research Grants in 2025.

Data availability

The data that has been used is confidential.

References

- Ahmed, S.R., Lemay, A., Hoebel, K., Kalpathy-Cramer, J., 2021. Focal loss improves repeatability of deep learning models. In: *Proceedings of the Medical Imaging with Deep Learning*. MIDL, URL: <https://openreview.net/pdf?id=7Cov6FwmOP1>.
- Araujo-Preza, C.E., Melhado, M.E., Gutierrez, F.J., Maniatis, T., Castellano, M.A., 2002. Use of capnometry to verify feeding tube placement. *Crit. Care Med.* 30 (10), 2255–2259. <http://dx.doi.org/10.1097/00003246-200210000-00013>.
- Boeykens, K., Holvoet, T., Duysburgh, I., 2023. Nasogastric tube insertion length measurement and tip verification in adults: a narrative review. *BMC* <http://dx.doi.org/10.1186/s13054-023-04611-6>, (317).
- Carse, J., Olmo, A.A., McKenna, S., 2022. Calibration of deep medical image classifiers: An empirical comparison using dermatology and histopathology datasets. In: *Proceedings of the Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. UNSURE, pp. 89–99. http://dx.doi.org/10.1007/978-3-031-16749-2_9.
- Elaanba, A., Ridouani, M., Hassouni, L., 2023. Automatic diagnosis framework for catheters and tubes semantic segmentation and placement errors detection. *IBICA* 649 (4), 176–188. http://dx.doi.org/10.1007/978-3-031-27499-2_17.
- Ellett, M.L.C., 2004. What is known about methods of correctly placing gastric tubes in adults and children. *Gastroenterol Nurs.* 27 (6), 253–259. <http://dx.doi.org/10.1097/00001610-200411000-00002>.
- Elpern, E.H., Killeen, K., Talla, E., Perez, G., Gurka, D., 2007. Capnometry and air insufflation for assessing initial placement of gastric tubes. *Am. J. Crit. Care* 16 (6), 544–549. <http://dx.doi.org/10.4037/ajcc2007.16.6.544>.
- Energias Market Research, 2018. Global nasogastric tube (NGT) market to witness a CAGR of 6.4% during 2018–2024. Available from: globenewswire.com. (Accessed 26 July 2018).
- Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning*. ICML, vol. 33, pp. 1050–1059, URL: <https://proceedings.mlr.press/v48/gall16.html>.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*. ICML, vol. 70, pp. 1321–1330, URL: <https://proceedings.mlr.press/v70/guo17a.html>.
- Henderson, R., Yi, X., Adams, S., Babyn, P., 2021. Automatic detection and classification of multiple catheters in neonatal radiographs with deep learning. *J. Digit. Imaging* 34 (4), 888–897. <http://dx.doi.org/10.1007/s10278-021-00473-y>.
- Huang, P., He, P., Tian, S., Ma, M., Feng, P., Xiao, H., Mercaldo, F., Santone, A., Qin, J., 2023. A ViT-AMC network with adaptive model fusion and multiobjective optimization for interpretable laryngeal tumor grading from histopathological images. *IEEE Trans. Med. Imaging* 42 (1), 15–28. <http://dx.doi.org/10.1109/TMI.2022.3202248>.
- Huang, P., Li, C., He, P., Xiao, H., Ping, Y., Feng, P., Tian, S., Chen, H., Mercaldo, F., Santone, A., Yeh, H.-y., Qin, J., 2024a. MamlFormer: Priori-experience guiding transformer network via manifold adversarial multi-modal learning for laryngeal histopathological grading. *Inf. Fusion* 108, 102333. <http://dx.doi.org/10.1016/j.inffus.2024.102333>.
- Huang, P., Luo, X., 2025. FDTs: A feature disentangled transformer for interpretable squamous cell carcinoma grading. *IEEE/CAA J. Autom. Sin.* 1–3. <http://dx.doi.org/10.1109/JAS.2024.125082>.
- Huang, P., Xiao, H., He, P., Li, C., Guo, X., Tian, S., Feng, P., Chen, H., Sun, Y., Mercaldo, F., Santone, A., Qin, J., 2024b. LA-ViT: A network with transformers constrained by learned-parameter-free attention for interpretable grading in a new laryngeal histopathology image dataset. *IEEE J. Biomed. Health Inform.* 28 (6), 3557–3570. <http://dx.doi.org/10.1109/JBHI.2024.3373438>.
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18 (2), 203–211. <http://dx.doi.org/10.1038/s41592-020-01008-z>.
- Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jäger, P.F., 2024. nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention*. MICCAI, vol. 15009, pp. 488–498. http://dx.doi.org/10.1007/978-3-031-72114-4_47.
- Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* <http://dx.doi.org/10.1038/s41597-019-0322-0>.
- Jung, S., Oh, J., Ryu, J., Kim, J., Lee, J., Cho, Y., Yoon, M., Jeong, J., 2022. Classification of central venous catheter tip position on chest X-ray using artificial intelligence. *J. Pers. Med.* 12 (10), <http://dx.doi.org/10.3390/jpm12101637>.
- Kearns, P.J., Donna, C., 2001. A controlled comparison of traditional feeding tube verification methods to a bedside, electromagnetic technique. *JPEN J. Parenter. Enteral Nutr.* 25 (4), 210–215. <http://dx.doi.org/10.1177/0148607101025004210>.
- Lin, M., Chen, Q., Yan, S., 2014. Network in network. In: *International Conference on Learning Representations*. ICLR, Available at: <https://arxiv.org/abs/1312.4400>.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, pp. 2117–2125. <http://dx.doi.org/10.1109/CVPR.2017.106>.
- Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. ICCV, pp. 2980–2988. <http://dx.doi.org/10.1109/ICCV.2017.324>.
- Liu, B., Rony, J., Galdran, A., Dolz, J., Ben Ayed, I., 2023. Class adaptive network calibration. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, pp. 16070–16079. <http://dx.doi.org/10.1109/CVPR52729.2023.01542>.
- Meta AI Research, 2024. Llama3: Open-source instruction-tuned language models. Available at: <https://ai.meta.com/llama>. (Accessed 2024).
- Metheny, N.A., Meert, K.L., 2004. Monitoring feeding tube placement. *Nutr. Clin. Pr.* 19 (5), 487–495. <http://dx.doi.org/10.1177/0115426504019005487>.
- Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H.S., Dokania, P.K., 2020. Calibrating deep neural networks using focal loss. In: *Advances in Neural Information Processing Systems*. NeurIPS, vol. 34, <http://dx.doi.org/10.48550/arXiv.2002.09437>.
- Müller, R., Kornblith, S., Hinton, G., 2019. When does label smoothing help? In: *Advances in Neural Information Processing Systems*. NeurIPS, vol. 33, pp. 4694–4703, URL: https://papers.nips.cc/paper_files/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.
- NHS National Patient Safety Agency, 2011. Reducing the harm caused by misplaced nasogastric feeding tubes in adults, children, and infants. Available from: cas.mhra.gov.uk. (Accessed 10 March 2011).
- Park, I., Kim, W.H., Ryu, J., 2024. Style-KD: Class-imbalanced medical image classification via style knowledge distillation. *Biomed. Signal Process. Control.* 91 (1), 105928. <http://dx.doi.org/10.1016/j.bspc.2023.105928>.
- Rajaraman, S., Ganesan, P., Antani, S., 2022. Does deep learning model calibration improve performance in class-imbalanced medical image classification? *PLoS One* 17 (1), <http://dx.doi.org/10.1371/journal.pone.0262838>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention*. MICCAI, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.
- Sambyal, A.S., Niyaz, U., Krishnan, N.C., Bathula, D.R., 2023. Understanding calibration of deep neural networks for medical image classification. *Comput. Methods Programs Biomed.* 242, 107816. <http://dx.doi.org/10.1016/j.cmpb.2023.107816>.
- Singh, V., Danda, V., Gorniak, R., Flanders, A., Lakhani, P., 2019. Assessment of critical feeding tube malpositions on radiographs using deep learning. *J. Digit. Imaging* 32 (4), 651–655. <http://dx.doi.org/10.1007/s10278-019-00229-9>.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, pp. 2818–2826. <http://dx.doi.org/10.1109/CVPR.2016.308>.
- Tang, C., Seah, J., Ahmad, H., Milne, M., Wardman, J., Buchlak, Q., Esmaili, N., Lambert, J., Jones, C., 2023. Analysis of line and tube detection performance of a chest X-ray deep learning model to evaluate hidden stratification. *Diagnostics* 13 (14), 2317. <http://dx.doi.org/10.3390/diagnostics13142317>.
- Wang, C., Balazs, J., Szarvas, G., Ernst, P., Poddar, L., Danchenko, P., 2022. Calibrating imbalanced classifiers with focal loss: An empirical study. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP, pp. 145–153. <http://dx.doi.org/10.18653/v1/2022.emnlp-industry.14>.
- Wang, D.B., Feng, L., Zhang, M.L., 2021. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In: *Advances in Neural Information Processing Systems*. NeurIPS, vol. 35, URL: <https://proceedings.neurips.cc/paper/2021/hash/61f3a6dbc9120ea78ef75544826c814e-Abstract.html>.
- Wang, C.-H., Hwang, T., Huang, Y.-S., Tay, J., Wu, C.-Y., Wu, M.-C., Roth, H.R., Yang, D., Zhao, C., Wang, W., Huang, C.-H., 2024. Deep learning-based localization and detection of malpositioned nasogastric tubes on portable supine chest X-Rays in intensive care and emergency medicine: A multi-center retrospective study. *J. Digit. Imaging* 37 (5), 1024–1032. <http://dx.doi.org/10.1007/s10278-024-01181-z>.