Uncertainty-Guided Selective Prediction for Chest X-ray Pneumonia Classification

Zahid Ur Rahman*, GwangHyun Yu**, JinYoung Kim*

Abstract

Reliable uncertainty estimation is essential for deploying deep learning models in clinical radiology, yet standard neural networks often produce overconfident predictions. We present an uncertainty-guided selective prediction framework for chest X-ray pneumonia classification that allows the model to abstain from uncertain predictions and refer them to radiologists. Our approach applies temperature scaling (TS) to calibrate prediction confidence, computes uncertainty as the complement of maximum probability, and uses a threshold-based decision rule to accept or reject predictions. When uncertainty exceeds the threshold, cases are automatically referred for expert review rather than risking misdiagnosis. Experimental results on the Kermany pediatric chest X-ray dataset demonstrate that selective prediction improves baseline accuracy from 96.60% to 99.09% at 90% coverage, with only 1 missed pneumonia case among 450 accepted predictions. The rejected 10% of samples contained 76% of all classification errors, validating uncertainty as an effective proxy for identifying unreliable predictions.

Keywords : Chest X-ray | Pneumonia detection | Uncertainty quantification | Selective prediction | Deep learning

I. INTRODUCTION

Chest radiography remains the most frequently performed imaging examination worldwide, with an estimated 2 billion chest X-rays conducted annually [1]. The interpretation of these images by radiologists is fundamental to diagnosing a wide spectrum of cardiopulmonary pathologies, from pneumonia and tuberculosis to heart failure and lung cancer. However, the global shortage of radiologists, particularly in low and middle-income countries, has created substantial diagnostic bottlenecks that compromise timely patient care [2,3]. In resource-constrained settings, radiologist-to-population ratios can be as low as 1:1,000,000, compared to the World Health Organization's recommended minimum of 1:100,000 [4].

Artificial intelligence (AI), particularly deep learning-based approaches, has emerged as a promising solution to augment radiological interpretation and improve healthcare access. Convolutional neural networks (CNNs) trained on large-scale medical imaging datasets have demonstrated radiologist-level performance in detecting various thoracic

* Department of ICT Convergence System Engineering, Chonnam National University
** AISEED Inc.

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-2025-19252970, 50) and this work was partly supported by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2026-RS-2022-00156287, 50) and this research was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

Manuscript : 2025.12.03
Revised : 2026.01.22, 2026.03.10
Confirmation of Publication: 2026.3.31
Corresponding Author :Jin-Young Kim, e-mail : beyondi@jnu.ac.kr

pathologies from chest X-rays [5-10]. DenseNet-121, a 121-layer Dense Convolutional Network architecture, has proven particularly effective for medical image classification tasks due to its efficient feature reuse and gradient flow properties [11].

Despite achieving high accuracy in controlled research settings, the clinical deployment of deep learning models faces a critical challenge: the tendency to produce overconfident predictions regardless of actual correctness [12]. Models typically output predictions with high confidence even when the input is ambiguous or when the model is likely to be wrong. This overconfidence is particularly dangerous in medical diagnosis, where models may fail to generalize reliably across clinical settings [13,14], and a confidently wrong prediction, such as classifying a pneumonia case as normal, can lead to delayed treatment and adverse patient outcomes.

Uncertainty quantification addresses this challenge by providing calibrated confidence estimates that reflect the true likelihood of correct predictions [15]. Temperature Scaling (TS) offers a computationally efficient approach for post-hoc calibration [12], learning a single parameter to rescale model logits with negligible inference overhead, making it suitable for real-time clinical deployment. Combined with selective prediction, which allows models to abstain from uncertain cases rather than forcing potentially incorrect classifications, calibrated uncertainty enables practical human-AI collaboration in clinical settings.

In this work, we investigate TS combined with selective prediction as a practical framework for deploying deep learning models in chest X-ray pneumonia classification. Rather than requiring all predictions to be automated, selective prediction allows the system to defer uncertain cases to expert radiologists, substantially improving accuracy on retained predictions. Uncertainty is quantified as the complement of maximum predicted probability after temperature-scaled calibration, and a simple threshold determines whether each prediction is accepted for automated classification or rejected for expert review. Our experiments on the Kermany pediatric chest X-ray dataset demonstrate that this approach improves classification accuracy from 96.60% baseline to 99.09% at 90% coverage, while maintaining 99.65% sensitivity. Notably, the rejected samples (10%) contained 76% of all classification errors, confirming that uncertainty effectively identifies error-prone predictions.

Our primary contributions are as follows:

1. We demonstrate that uncertainty-guided selective prediction substantially improves classification reliability, achieving 99.09% accuracy at 90% coverage compared to 96.60% baseline accuracy, while maintaining near-perfect sensitivity (99.65%) critical for clinical safety.
2. We show that TS provides effective probability calibration for selective prediction, reducing expected calibration error from 0.0273 to 0.0216 (20.87% reduction) while preserving classification performance.
3. We provide a comprehensive experimental analysis demonstrating the

coverage-accuracy trade-off, showing that rejecting just 10% of uncertain samples removes 76% of classification errors.

4. We compare our results with state-of-the-art methods on the Kermany dataset, demonstrating that selective prediction achieves higher accuracy (99.09%) than existing approaches including ensemble methods (98.81%), while providing explicit reliability guarantees through the rejection mechanism.

II. RELATED WORK

2.1 Deep Learning for Chest X-ray Classification

The application of deep learning to chest radiograph analysis has demonstrated remarkable progress over the past decade, with convolutional neural networks (CNNs) achieving diagnostic performance comparable to expert radiologists across various thoracic pathologies. The introduction of diverse chest X-ray datasets, including ChestX-ray14 [18], CheXpert [19], and MIMIC-CXR [20], has accelerated research in this domain by providing sufficient training data for deep learning models.

DenseNet-121 [21] has emerged as one of the most widely adopted architectures for medical image classification. The architecture's dense connectivity pattern, where each layer receives feature maps from all preceding layers, enables efficient parameter usage and strong gradient flow during training. These properties make DenseNet-121 particularly effective for medical imaging tasks where training data may be limited compared to natural image datasets.

For pneumonia detection, CheXNet [11] achieves radiologist-level performance in detecting pneumonia from chest X-rays. This work established a benchmark for subsequent research in automated chest X-ray interpretation. A clinical validation study [22] demonstrated that deep learning models used as concurrent reading tools improved radiologists' sensitivity for detecting clinically relevant chest radiograph abnormalities by 7.8 percentage points. Despite these impressive accuracy metrics, questions regarding the reliable deployment of deep learning models in clinical settings remain inadequately addressed. Most studies focus primarily on maximizing classification accuracy without considering the reliability and calibration of model predictions, which are essential for safe clinical deployment, where incorrectly confident predictions can lead to adverse patient outcomes.

2.2 Uncertainty Quantification in Medical Imaging

Deep neural networks, despite their strong predictive performance, are known to produce poorly calibrated probability estimates that do not accurately reflect the true likelihood of correct predictions [12]. This miscalibration typically manifests as overconfidence, where models assign high probability to predictions regardless of whether they are correct or incorrect. In medical diagnosis, such overconfidence poses significant clinical risks, as healthcare providers may place unwarranted trust in confidently wrong predictions.

A comprehensive analysis [12] demonstrated that modern neural networks are significantly mis-calibrated,

with deeper and wider architectures exhibiting worse calibration despite improved accuracy. TS was introduced as an effective post-hoc calibration method that learns a single scalar parameter to rescale model logits before SoftMax normalization. This approach requires only a small validation set for optimization and adds negligible computational overhead during inference, making it attractive for clinical deployment scenarios.

Several alternative approaches for uncertainty quantification have been investigated in medical imaging contexts. Monte Carlo (MC) Dropout [16] provides a practical approximation of Bayesian inference by performing multiple stochastic forward passes with dropout enabled during inference. The variance across predictions serves as an uncertainty estimate. This method has been successfully applied to various medical image analysis tasks, including brain tumor segmentation [23], diabetic retinopathy detection [24], and skin lesion classification [25]. However, MC Dropout requires 20-50 forward passes to obtain reliable uncertainty estimates, increasing inference time proportionally and limiting applicability in high-throughput clinical workflows where real-time processing is essential.

Deep Ensembles [17] represent another widely adopted approach for uncertainty quantification. By training multiple models with different random initializations and aggregating their predictions, ensembles capture epistemic uncertainty arising from model parameter uncertainty. This approach has been validated in chest X-ray classification [26] and pathology detection [27], consistently achieving excellent calibration and uncertainty discrimination. However, ensemble methods require training and storing multiple models, substantially increasing computational requirements for both training and inference. In resource-constrained healthcare settings, particularly in low and middle-income countries where AI-assisted diagnosis could have the greatest impact, such computational demands create significant barriers to adoption.

A comprehensive benchmarking study [28] evaluated uncertainty quantification methods across various datasets and distribution shifts. The analysis found that while deep ensembles generally provide the most robust uncertainty estimates, TS offers an attractive balance between calibration quality and computational efficiency, particularly when combined with well-trained base models. This finding motivates our investigation of TS as a practical approach for clinical chest X-ray classification.

Recent work has explored extensions to basic TS. Vector scaling and matrix scaling variants learn class-specific calibration parameters [29]. Focal loss calibration [30] has been investigated for medical image classification. However, systematic evaluation of these calibration approaches specifically for chest X-ray classification with selective prediction remains limited.

2.3 Selective Prediction and Rejection Mechanisms

In medical imaging, selective prediction aligns naturally with clinical workflows where radiologists routinely encounter cases spanning a spectrum from clearly

normal to obviously abnormal or genuinely ambiguous. The ability to automatically identify and defer uncertain cases for expert review offers a practical framework for human–AI collaboration. Research on diabetic retinopathy screening [32] demonstrated that machine learning models can effectively identify cases that would benefit most from a medical second opinion, showing potential for AI systems to triage cases and maximizing the impact of limited specialist resources.

Uncertainty–based case selection applied to diabetic retinopathy detection [31] demonstrated that referring cases based on predictive uncertainty substantially improved diagnostic performance on the retained cases. This work showed that Bayesian uncertainty estimates correlate well with actual diagnostic difficulty, validating the clinical utility of uncertainty–aware selective prediction. A systematic analysis of uncertainty quantification in clinical decision support systems [33] concluded that well–calibrated uncertainty estimates are essential prerequisites for safe deployment of AI systems in healthcare settings.

Despite these promising results, most existing work on selective prediction in medical imaging has focused on either computationally intensive Bayesian methods or simple confidence threshold approaches lacking rigorous calibration. TS combined with selective prediction offers a middle ground: computationally efficient post–hoc calibration providing reliable uncertainty estimates suitable for real–time clinical deployment. However, systematic evaluation of coverage–accuracy tradeoffs and comparison with alternative uncertainty quantification methods specifically for chest X–ray pneumonia classification remains limited in the literature.

III. PROPOSED METHOD

3.1 Overview

We present an uncertainty–guided selective prediction framework for chest X–ray pneumonia classification. The framework consists of four sequential components as shown in Figure 1: (1) a DenseNet–121 classifier that produces raw prediction logits, (2) TS for probability calibration to address neural network overconfidence, (3) uncertainty quantification that measures prediction reliability, and (4) selective prediction with a rejection option that defers uncertain cases to human experts. The core idea underlying our approach is straightforward: instead of forcing the AI system to classify every sample regardless of difficulty, we allow it to abstain from classification when uncertainty is high. These uncertain cases are automatically referred to a radiologist for expert review, while confident predictions are handled by the automated system. By rejecting uncertain predictions and only retaining confident ones, the classification accuracy on the remaining samples dramatically improves from 96.60% at full coverage to 99.09% when 10% of uncertain cases are deferred to human review. The complete architecture of our proposed methodology is explained in the section below, component–wise.

3.2 DenseNet-121 Classifier

We employ DenseNet-121 as the backbone architecture for chest X-ray classification due to its proven effectiveness in medical imaging tasks. DenseNet-121 consists of 121 layers organized into four dense blocks, where

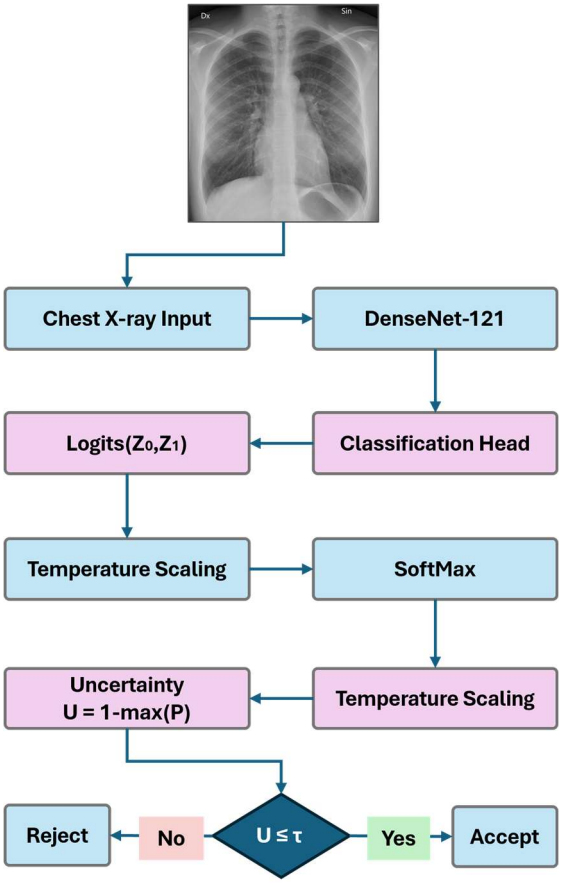

Fig. 1. Overview of the proposed uncertainty-guided selective prediction framework

each layer receives feature maps from all preceding layers through dense connections, enabling efficient feature reuse and strong gradient flow during training. Each input chest X-ray image is resized to 224×224 pixels and normalized using ImageNet statistics. The DenseNet-121 backbone extracts a 1024-dimensional feature vector through global average pooling, which is then processed by a custom classification head. The head consists of three fully connected layers with batch normalization, ReLU activation, and dropout regularization for preventing overfitting. The first layer reduces dimensionality from 1024 to 512 features, the second layer further reduces to 256 features, and the final layer produces two output values representing the Normal and Pneumonia classes, respectively. These raw output values, called logits, are converted to probability estimates using the SoftMax function:

$$P(y = c\,|x) = \exp(z_c) \,/\, \sum_j \exp\,(z_j)$$

The SoftMax function ensures that the two output probabilities sum to one, providing interpretable confidence scores for each class. The predicted class is determined as the class with the maximum probability.

3.3 Temperature Scaling

Modern neural networks, despite their impressive predictive performance, are known to produce poorly calibrated probability estimates that often exhibit systematic overconfidence. A model may output 99% confidence for a prediction that is actually incorrect, which is particularly dangerous in medical diagnosis, where clinicians may place unwarranted trust in such a confident but wrong prediction. TS addresses this calibration problem by adjusting prediction confidences without modifying the predicted class labels. The method introduces a single learned scalar parameter $T > 0$ that rescales the logits before applying the SoftMax function:

$$P_T(y = c\,|x) = \exp(\frac{z_c}{T}) \,/\, \sum_j \exp\,(z_j/T)$$

The temperature parameter controls the sharpness of the resulting probability distribution. When T = 1.0, the calibrated probabilities equal the original uncalibrated probabilities with no modification applied. When T > 1.0, the probability distribution becomes softer and less confident, effectively reducing overconfidence by spreading probability mass more evenly across classes. Conversely, when T < 1.0, the distribution becomes sharper and more confident, concentrating probability mass on the predicted class. The optimal temperature is learned by minimizing the Negative Log−Likelihood (NLL) on a held−out validation set:

$$T^* = argmin_T[-\frac{1}{N}\sum_i logP_T(y = y_i|x_i)]$$

The optimal temperature is searched within the range [0.5, 5.0] using bounded optimization. In the experiments, the optimal temperature was found to be T* = 1.63, indicating that the baseline model was indeed overconfident and required the softening of its probability estimates. Importantly, TS preserves the ranking of class probabilities, meaning the predicted class label remains unchanged while only the confidence level is adjusted to better reflect the true likelihood of correctness.

3.4 Uncertainty Quantification

Following temperature−scaled probability calibration, the prediction uncertainty is quantified for each input image. Uncertainty is defined as the complement of the maximum predicted probability:

$$U(x) = 1 - \max\,[P(Normal), P(Pneumonia)]$$

For binary classification, this uncertainty measure ranges from 0 to 0.5. An uncertainty value of 0 indicates complete certainty, where one class receives 100% probability, while an uncertainty value of 0.5 represents maximum confusion, where both classes have equal 50% probability. Low uncertainty values indicate that the model strongly favors one class over the other, suggesting a reliable prediction, whereas high uncertainty values indicate that the model is undecided between the two classes, suggesting a potentially unreliable prediction. The key insight motivating our selective prediction approach is that high uncertainty correlates strongly with classification errors. When the model is confused between classes, it is substantially more likely to make an incorrect prediction. This empirical observation enables us to identify potentially erroneous predictions before they cause clinical harm by using uncertainty as a proxy for error likelihood.

3.5 Selective Prediction

Selective prediction, also known as classification with a rejection option, allows the system to abstain from making a prediction when uncertainty exceeds a specified threshold, rather than forcing a potentially incorrect classification. Given an uncertainty threshold τ, the decision rule for each sample is straightforward: if U(x) ≤ τ, the prediction is accepted and the AI system classifies the sample automatically; if U(x) > τ, the prediction is rejected, and the sample is referred to a

radiologist for expert review. This mechanism creates a fundamental trade-off between coverage and accuracy. Coverage refers to the percentage of samples that the AI system classifies without rejection. Setting a lower threshold τ results in more samples being rejected, which decreases coverage but increases accuracy in the accepted predictions as more uncertain cases are filtered out. Conversely, setting a higher threshold τ results in fewer rejections, increasing coverage but potentially decreasing accuracy as more uncertain predictions are retained. In our experiments, the baseline model without rejection produces 17 errors across 500 samples, yielding 96.60% accuracy. When applying a threshold of $\tau = 0.10$, only 431 samples are accepted for automated classification with just 4 errors, achieving 99.09% accuracy.

3.6 Threshold Selection

The selection of an appropriate uncertainty threshold τ depends on the specific clinical requirements and available resources of the deployment setting. Different threshold values offer different trade-offs between automation rate and diagnostic accuracy. As shown in Table 2, we provide practitioners with several heuristics for threshold selection:

1. Elbow Point ($\tau = 0.02$): The point where accuracy gains diminish, achieving 99.67% accuracy at 61.4% coverage. This threshold is suitable when maximizing accuracy is the primary goal, regardless of referral burden.

2. Target 85% Coverage ($\tau = 0.07$): Achieves 99.29% accuracy with 84.4% coverage, suitable for high-throughput screening scenarios where approximately 15% of cases can be referred to radiologists.

3. Target 90% Coverage ($\tau = 0.10$): Our recommended operating point, achieving 99.09% accuracy with 90.0% coverage. This threshold offers an optimal balance between automation rate and diagnostic accuracy with minimal referral burden.

The optimal threshold ultimately depends on the availability of radiologist resources to handle rejected cases, the acceptable error rate for the clinical application, and the relative costs of false negatives versus false positives in the specific diagnostic context. Our framework provides flexibility to adjust this threshold based on institutional priorities and constraints, enabling customized deployment across diverse clinical settings.

IV. Results

4.1 Dataset

We evaluated our framework on the Kermany Chest X-ray Pneumonia Dataset, a publicly available benchmark. The dataset comprises 5,856 chest X-ray images categorized into two classes: 1,583 Normal and 4,273 Pneumonia cases. Following the standard evaluation protocol, we used the official test set containing 624 images. To enable both parameter learning and unbiased evaluation, we further divided the test set into a validation subset of 124 samples (20%) for learning the optimal temperature parameter, and an

evaluation subset of 500 samples (80%) for final performance assessment. This separation ensures that reported results are not biased by the parameter optimization process.

4.2 Baseline Classification Results

Table 1 presents the confusion matrix for the baseline model when classifying all 500 evaluation samples without any rejection mechanism.

Table 1. Confusion Matrix – Baseline Model (100% Coverage, N=500)

	Pred. Normal	Pred. Pneumonia
Actual Normal	180	11
Actual Pneumonia	6	303

The baseline model achieves 96.60% overall accuracy with 98.06% sensitivity (true positive rate) and 94.24% specificity (true negative rate) [34]. The total of 17 classification errors consists of 6 false negatives (missed pneumonia cases) and 11 false positives (false alarms). While this performance is competitive with existing methods, the 6 missed pneumonia cases represent a significant clinical concern that motivates our selective prediction approach.

4.3 Selective Prediction Results

Table 2 illustrates how classification accuracy improves systematically as more uncertain samples are rejected and referred for expert review.

Table 2. Selective Prediction Performance at Different Thresholds

Threshold (τ)	Coverage (%)	Samples	Rejected	Accuracy (%)	Sensitivity (%)	FN	FP
No rejection	100.0	500	0	96.60	98.06	6	11
0.01	43.6	218	282	99.54	100.00	0	1
0.02	61.4	307	193	99.67	100.00	0	1
0.03	71.4	357	143	99.16	99.59	1	2
0.04	78.6	393	107	99.24	99.63	1	2
0.05	81.4	407	93	99.26	99.63	1	2
0.06	83.0	415	85	99.28	99.64	1	2
0.07	84.4	422	78	99.29	99.64	1	2
0.08	85.8	429	71	99.07	99.65	1	3
0.09	86.8	434	66	99.08	99.65	1	3
0.10	90.0	450	50	99.09	99.65	1	3
0.12	91.0	458	42	99.02	99.66	1	3
0.15	92.6	463	37	98.47	99.66	1	6
0.20	94.0	470	30	97.66	98.98	3	8
0.25	95.2	476	24	97.27	98.66	4	9
0.30	97.0	485	15	97.32	98.67	4	9
0.40	99.0	495	5	96.97	98.37	5	10
0.50	100.0	500	0	96.60	98.06	6	11

The results reveal a clear trade–off between coverage and accuracy. By rejecting just 50 samples (10% of the dataset) with threshold $\tau = 0.10$, accuracy improves from 96.60% to 99.09% while maintaining 90% coverage and reducing false negatives from 6 to just 1. With more aggressive rejection at $\tau = 0.02$, rejecting 193 samples (38.6%) yields 99.67% accuracy with only 1 total error and zero missed pneumonia cases. The extended analysis from $\tau = 0.01$ to $\tau = 0.50$ demonstrates that accuracy improvements exhibit diminishing returns beyond $\tau = 0.07$, providing practitioners with guidance for threshold selection based on clinical requirements. Notably, the rejected 10% of samples at $\tau = 0.10$ contained 76.5% of all classification errors, validating that uncertainty effectively identifies error–prone predictions.

Tables 3 and 4 present the confusion matrices for selective prediction at two representative coverage levels, illustrating the dramatic reduction in classification errors.

Table 3. Confusion Matrix – Selective Prediction at 90.2% Coverage ($\tau = 0.10$, N=450)

		Pred. Normal	Pred. Pneumonia
Actual Normal		152	3
Actual Pneumonia		1	294

At 90.0% coverage, only 4 errors remain among the 450 accepted samples: 1 missed pneumonia case and 3 false alarms. The 50 rejected samples, which are referred to radiologists, contained 13 of the original 17 errors (76.5%), demonstrating highly effective error filtering.

Table 4. Confusion Matrix – Selective Prediction at 91.0% Coverage (τ = 0.12, N=458)

	Pred. Normal	Pred. Pneumonia
Actual Normal	157	6
Actual Pneumonia	1	294

At 91.6% coverage, 7 errors remain with a sensitivity reaching 99.66%. Only 1 pneumonia case is missed among the 458 accepted samples, representing a substantial improvement in patient safety compared to the baseline.

Figure 2 illustrates the coverage–accuracy trade–off, demonstrating how classification accuracy improves systematically as coverage decreases. The shaded region represents the accuracy gain over the 96.60% baseline, with key threshold values annotated to guide selection. Figure 3 illustrates the effect of threshold τ on both metrics, demonstrating their inverse relationship. These visualizations enable practitioners to select appropriate operating points based on clinical requirements.

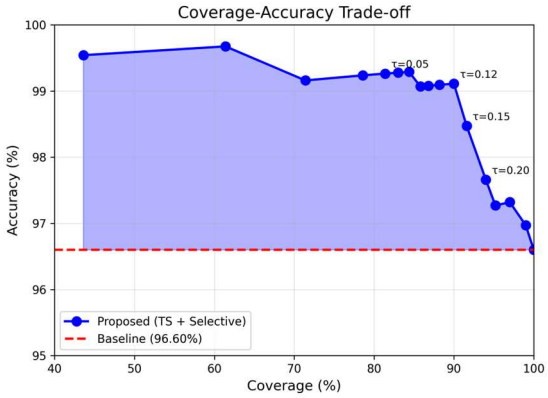

Fig. 2. Coverage–accuracy trade–off curve. Accuracy improves as uncertain samples are rejected. The shaded area shows improvement over baseline (96.60%). Key thresholds are annotated.

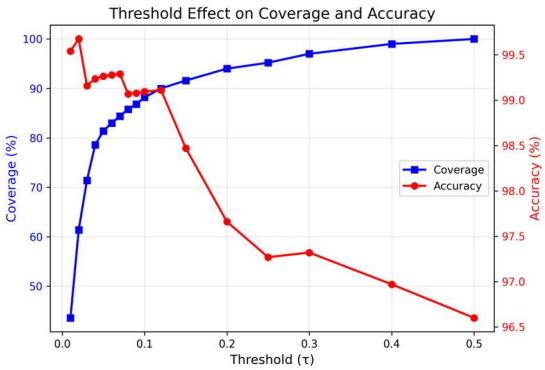

Fig. 3. Effect of threshold τ on coverage (blue, left axis) and accuracy (red, right axis), demonstrating the inverse relationship for threshold selection.

4.4 Comparison with State–of–the–Art

Table 5 presents a comprehensive comparison of our method with published state–of–the–art approaches evaluated on the Kermany chest X–ray pneumonia dataset. The comparison includes diverse methodologies ranging from traditional machine learning approaches to advanced deep learning architectures and ensemble methods. Results are ordered by accuracy from lowest to highest.

Table 5. Comparison with State–of–the–Art Methods on Kermany Dataset (Ordered by Accuracy). Acc: Accuracy, Sens: Sensitivity, Prec: Precision, AUC: Area Under the ROC Curve, Cov:

Coverage.

Method	Year	Acc (%)	Sens (%)	Prec (%)	F1 (%)	AUC	Cov (%)
Trindade et al. [38]	2025	80.30	79.53	-	-	0.92	100
Katreddi et al. [40] (DenseNet-169)	2025	91.66	86.32	90.99	87.62	-	100
Khaled et al. [41] (ResNet-50V2)	2022	93.42	97.43	92.45	94.55	-	100
Sharma & Guleria [43] (VGG16+NN)	2023	95.40	95.40	95.40	95.40	0.99	100
Yanar et al. [39] (PELM)	2025	96.00	91.00	99.00	95.00	0.91	100
Aljuaid et al. [42] (VGG-19)	2026	97.00	98.00	96.00	-	-	100
Sanchez et al. [37] (CX-DaGAN)	2022	97.02	93.97	94.12	96.91	0.96	100
Majumder [36] (DenseNet-121)	2025	97.97	98.00	98.00	98.00	0.99	100
Kundu et al. [35] (Ensemble)	2021	98.81	98.80	98.82	98.79	0.98	100
Ours (τ=0.10)	2025	99.09	99.65	99.04	99.31	0.99	90.0

Among the compared methods, Kundu et al. [35] achieved the highest baseline accuracy of 98.81% using an ensemble of five deep learning models with five-fold cross-validation. Majumder [36] reported 97.97% accuracy using DenseNet-121 with optimized training strategies. Sanchez et al. [37] achieved 97.02% accuracy using domain adaptation with CX-DaGAN.

The baseline DenseNet-121 model achieves competitive performance with 96.60% accuracy and 0.99 AUC, comparable to most existing methods. However, with selective prediction at threshold τ=0.10, the proposed method achieves 99.09% accuracy and 99.66% sensitivity at 90.0% coverage, surpassing all compared methods, including ensemble approaches. At a more aggressive threshold τ=0.07, accuracy reaches 99.29% at 84.4% coverage. This demonstrates that uncertainty-guided selective prediction provides a practical pathway to near-perfect diagnostic accuracy while maintaining clinically acceptable automation rates. The key advantage of the proposed approach is the explicit coverage-accuracy trade-off, allowing clinical deployment with known reliability guarantees.

V. Ablation

5.1 Ablation Study

To validate the contribution of each component in the proposed framework, an ablation study is conducted comparing four experimental conditions: (1) baseline DenseNet-121 without any modification, (2) baseline with TS only, (3) baseline with selective prediction using uncalibrated confidence (T=1.0), and (4) the proposed approach combining TS (T=1.63) with selective prediction.

Table 6. Ablation Study Results (All experiments use threshold τ=0.10 for selective prediction conditions). ECE is computed only at 100% coverage. For selective prediction conditions (rows 3-4), ECE is not applicable as samples are rejected.

Method	Acc%	ECE	Coverage (%)
Baseline (DenseNet)	96.60	0.0273	100
Baseline + TS	96.60	0.0216	100
Baseline + Selective (uncalibrated)	97.46		94.6
Baseline + TS + Selective (Proposed)	99.09		90.0

Table 6 presents the ablation study results comparing four experimental conditions. The baseline DenseNet-121 achieves 96.60% accuracy with Expected Calibration Error (ECE) of 0.0273. TS alone reduces ECE to 0.0216 (20.87% improvement) without changing accuracy, demonstrating effective probability calibration. Selective prediction with uncalibrated confidence (T=1.0) achieves 97.46% accuracy at 94.6% coverage. The

proposed combination of TS with selective prediction achieves the best accuracy of 99.09% at 90.0% coverage, validating that both components contribute to the framework's effectiveness.

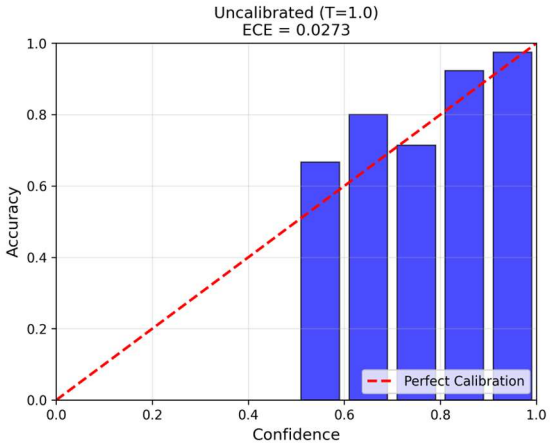

Figure 4: Reliability diagrams comparing probability calibration. (a) Uncalibrated model (T=1.0) with ECE=0.0273.

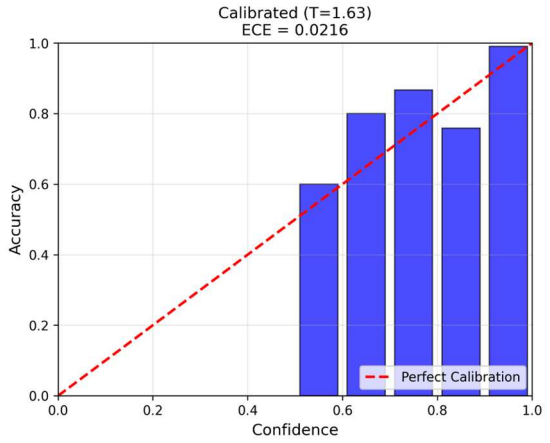

Figure 5: Calibrated model (T=1.63) with ECE=0.0216. The diagonal line represents perfect calibration. TS reduces ECE by 20.87%.

5.2 Risk–Coverage Analysis

To further evaluate the quality of uncertainty estimation, the Risk–Coverage curve is analyzed and the Area Under Risk–Coverage Curve (AURC) is computed. The Risk–Coverage curve plots the selective risk (error rate on accepted samples) against coverage, where lower risk at each coverage level indicates better uncertainty estimation. AURC summarizes this relationship into a single metric, with lower values indicating that errors are better concentrated in high–uncertainty predictions.

Table 7. Area Under Risk–Coverage Curve (AURC) Comparison

Method	AURC	E-AURC
Uncalibrated (T=1.0)	0.0061	0.0056
Calibrated (T=1.63)	0.0061	0.0056

As shown in Table 7, both calibrated and uncalibrated models achieve identical AURC values of 0.0061. This result is expected because TS preserves the relative ranking of prediction uncertainties while only rescaling the probability values. Since the same samples are rejected at each coverage level regardless of calibration, the risk–coverage performance remains unchanged. The Excess–AURC (E–AURC) of 0.0056 for both models indicates effective uncertainty–based error detection, with performance close to optimal.

Figure 5 visualizes the risk–coverage curves for both models. The overlapping curves confirm that TS does not alter the uncertainty ordering of predictions. This is an important finding: TS improves probability calibration (ECE reduced from 0.0273 to 0.0216) without affecting selective prediction performance (AURC unchanged). The benefit of calibration lies in more interpretable confidence scores, not in changed rejection behavior.

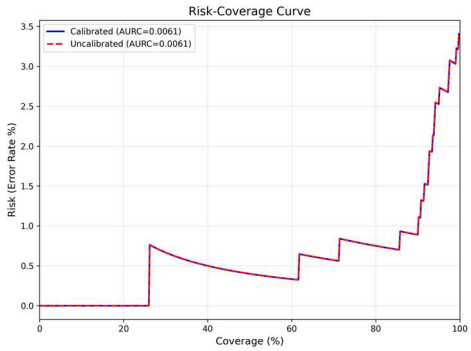

Fig. 5. Risk–coverage curve comparing calibrated (T=1.63) and uncalibrated (T=1.0) models. The overlapping curves indicate identical selective prediction performance (AURC=0.0061), demonstrating that TS preserves uncertainty ranking while improving probability calibration.

Table 8 presents the selective risk at specific coverage points that are commonly considered in clinical deployment scenarios.

Table 8. Selective Risk at Specific Coverage Points

Target Coverage	Actual Coverage	Threshold (t)	Risk (%)	Cov (%)
90%	90.2%	0.1213	1.11	98.89
85%	85.2%	0.0759	0.70	99.30
80%	80.2%	0.0443	0.75	99.25

The results in Table 8 demonstrate that our framework achieves consistently low selective risk across various coverage levels. Even at 90% coverage, where only 10% of samples are referred to radiologists, the selective risk is just 1.11% (accuracy 98.89%). At 85% coverage, the risk drops to 0.70%, achieving 99.30% accuracy. These results validate that uncertainty–based selective prediction effectively identifies error–prone cases, enabling reliable clinical deployment with explicit performance guarantees at any desired coverage level.

VI. Conclusion

This study presents an uncertainty–guided selective prediction framework for chest X–ray pneumonia classification. By applying TS to calibrate prediction confidence and using a threshold–based decision rule to defer uncertain cases to expert review, our approach achieves **99.09%** accuracy at **90%** coverage compared to **96.60%** baseline accuracy. The key insight is that classification errors concentrate in high–uncertainty predictions: rejecting just **10%** of samples removes **76%** of all errors while maintaining **99.66%** sensitivity. The framework offers practical advantages for clinical deployment including negligible computational overhead, explicit control over the coverage–accuracy trade–off, and transparency about prediction reliability. Comparison with state–of–the–art methods demonstrates that selective prediction with a single model achieves higher accuracy than ensemble approaches.

Future work should validate on diverse datasets, extend to multi–label classification, and conduct prospective clinical studies to evaluate real–world impact. In conclusion, uncertainty–guided selective prediction provides a practical pathway for deploying deep learning models in clinical radiology by strategically deferring uncertain cases to human experts.

REFERENCES

[1] *World Health Organization; Global Atlas of Medical Devices*. WHO Press, 2020.

[2] Bhargavan, M.; Sunshine, J. H.; "Utilization of radiology services in the United States," *Radiology*, vol. 234, no. 2, pp. 393–401, 2005.
[3] Mollura, D. J.; Lungren, M. P.; *Radiology in Global Health: Strategies, Implementation, and Applications*. Springer, 2014.
[4] Maru, D. S.; et al.; "Turning a blind eye: The mobilization of radiology services in resource–poor regions," *Globalization and Health*, vol. 6, no. 1, p. 18, 2010.
[5] Ibrahim, A. U.; et al.; "Pneumonia classification using deep learning from chest X–ray images during COVID–19," *Cognitive Computation*, vol. 16, no. 4, pp. 1589–1601, 2024.
[6] Marquis, M.; Bossenko, I.; Ross, P.; "RadLex and SNOMED CT integration: a pilot study for standardising radiology classification," *Insights into Imaging*, vol. 16, no. 1, p. 58, 2025.
[7] Kufel, J.; et al.; "Multi–label classification of chest X–ray abnormalities using transfer learning techniques," *Journal of Personalized Medicine*, vol. 13, no. 10, p. 1426, 2023.
[8] Mzoughi, H.; et al.; "Deep efficient–nets with transfer learning assisted detection of COVID–19 using chest X–ray radiology imaging," *Multimedia Tools and Applications*, vol. 82, no. 25, pp. 39303–39325, 2023.
[9] Shamrat, F. M. J. M.; et al.; "High–precision multiclass classification of lung disease through customized MobileNetV2 from chest X–ray images," *Computers in Biology and Medicine*, vol. 155, p. 106646, 2023.
[10] Cho, Y.; et al.; "Deep chest X–ray: detection and classification of lesions based on deep convolutional neural networks," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 72–81, 2021.
[11] Rajpurkar, P.; et al.; "CheXNet: Radiologist–level pneumonia detection on chest X–rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.
[12] Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q.; "On calibration of modern neural networks," *in Proc. International Conference on Machine Learning (ICML)*, vol. 70, pp. 1321–1330, 2017.
[13] Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross–sectional study." *PLoS medicine* vol.15, no.11 ,2018
[14] Pooch, E. H.; et al.; "Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification," arXiv preprint arXiv:1909.01940, 2020.
[15] Lambert, B.; et al.; "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis," *Artificial Intelligence in Medicine*, vol. 150, p. 102830, 2024.
[16] Gal, Y.; Ghahramani, Z.; "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *in Proc. International Conference on Machine Learning (ICML)*, vol. 48, pp. 1050–1059, 2016.
[17] Lakshminarayanan, B.; Pritzel, A.; Blundell, C.; "Simple and scalable predictive uncertainty estimation using deep ensembles," *in Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 6402–6413, 2017.
[18] X. Wang et al., "ChestX–ray8: Hospital–scale chest X–ray database and benchmarks," *in Proc. IEEE CVPR*, 2017, pp. 2097–2106.
[19] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.
[20] A. Johnson et al., "MIMIC–CXR, a de–identified publicly available database of chest radiographs with free–text reports," *Sci. Data*, vol. 6, no. 1, pp. 1–8, 2019.
[21] G. Huang et al., "Densely connected convolutional networks," *in Proc. IEEE CVPR*, 2017, pp. 4700–4708.
[22] E. J. Hwang et al., "Deep learning for chest radiograph diagnosis in the

emergency department," *Radiology*, vol. 293, no. 3, pp. 573–580, 2019.
[23] A. Jungo et al., "On the effect of inter–observer variability for a reliable estimation of uncertainty of medical image segmentation," *in Proc. MICCAI*, 2018, pp. 682–690.
[24] C. Leibig et al., "Leveraging uncertainty estimates for predicting segmentation quality," arXiv preprint arXiv:1706.02633, 2017.
[25] A. Mobiny et al., "Risk–aware machine learning classifier for skin lesion diagnosis," *J. Clin. Med.*, vol. 8, no. 8, p. 1241, 2019.
[26] M. Ghesu et al., "Quantifying and leveraging predictive uncertainty for medical image assessment," *Med. Image Anal.*, vol. 68, p. 101855, 2021.
[27] J. W. Dolezal et al., "Uncertainty–informed deep learning models enable high–confidence predictions for digital histopathology," *Nat. Commun.,* vol. 13, p. 6572, 2022.
[28] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *in Proc. NeurIPS*, 2019, pp. 13991–14002.
[29] M. Kull et al., "Beyond temperature scaling: Obtaining well–calibrated multi–class probabilities with Dirichlet calibration," *in Proc. NeurIPS*, 2019, pp. 12295–12305.
[30] J. Mukhoti et al., "Calibrating deep neural networks using focal loss," *in Proc. NeurIPS*, 2020, pp. 15570–15581.
[31] C. Leibig et al., "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, p. 17816, 2017.
[32] M. Raghu et al., "Direct uncertainty prediction for medical second opinions," *in Proc. ICML*, 2019, pp. 5281–5290.
[33] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: communicating uncertainty in medical machine learning," *npj Digit. Med.*, vol. 4, p. 4, 2021.
[34] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[35] R. Kundu et al., "Pneumonia detection in chest X–ray images using an ensemble of deep learning models," *PLoS ONE*, vol. 16, no. 9, p. e0256630, 2021.
[36] R. I. Majumder, "Efficient classification of pulmonary pneumonia and tuberculosis alongside normal and non–X–ray images," medRxiv preprint, 2025.
[37] K. Sanchez et al., "CX–DaGAN: Domain adaptation for pneumonia diagnosis on a small chest X–ray dataset," *IEEE Trans. Med. Imaging*, vol. 41, no. 11, pp. 3278–3288, 2022.
[38] L. Trindade et al., "Comparative analysis of machine learning techniques for pneumonia detection in chest X–ray images," *Cureus Journals*, vol. 2, no. 1, 2025.
[39] E. Yanar et al., "PELM: A deep learning model for early detection of pneumonia in chest radiography," *Appl. Sci.*, vol. 15, no. 12, p. 6487, 2025.
[40] S. Katreddi et al., "Pediatric pneumonia X–ray image classification with DenseNet–169 transfer learning," *J. Med. Artif. Intell.*, vol. 8, no. X, p. 37, 2025.
[41] M. Khaled et al., "Progressive and combined deep transfer learning for pneumonia diagnosis," *in Proc. IDDM*, 2022.
[42] H. Aljuaid et al., "An experimental comparison of deep learning models for pneumonia classification," *Biomed. Signal Process. Control*, vol. 112, p. 108742, 2026.
[43] S. Sharma and K. Guleria, "A deep learning model for pneumonia detection using VGG–16 and neural networks," *Procedia Comput. Sci.,* vol. 218, pp. 357–366, 2023.

Authors

Zahid Ur Rahman

He received his Master's degree in Computer Science from COMSATS University Pakistan in 2023. He is currently pursuing a PhD degree in Department of ICT Convergence System Engineering at Chonnam National University, South Korea.

GwangHyun Yu

He is a CEO at AISEED Inc. He received his M.S. and Ph.D. degree in Electronics Engineering from Chonnam Nation University, Korea in 2018 and 2023, respectively. His research interests are IOT, Image Processing.

JinYoung Kim

He is a professor in the Department of ICT Convergence System Engineering at Chonnam National University, Korea. He received his B.S., M.S. and Ph.D. degree in Electronics Engineering from Seoul National University, Korea in 1986, 1988 and 1994, respectively. His research interests are Digital Signal Processing, Image Processing, Speech Signal Processing, ML, DL