Jurnal Sains dan Kesehatan
Volume. 9 Nomor. 1 Juni 2025
E-ISSN : 2962-1569; P-ISSN : 2580-8362, Hal 55-78
DOI: https://doi.org/10.57214/jusika.v9i1.1126
Tersedia: https://ejournal.unimman.ac.id/index.php/jusika

Enhancing Clinical Decision Support through Cost-Sensitive CNN and Reliability-Calibrated Pneumonia Classification

Danang[1*], Toni Wijanarko Adi Putra[2]
[1-2]Universitas Sains dan Teknologi Komputer, Indonesia
Email: danang150787@gmail.com [1], toni.wijanarko@stekom.ac.id [2]
**Corresponding Author: danang150787@gmail.com*

Abstract. Pneumonia detection from chest X-ray images is widely used in computer-aided diagnostic systems. However, effective clinical decision support requires not only accurate classification performance but also consideration of unequal error costs, since false negative predictions may lead to more severe consequences than false positives. In addition, prediction probabilities must be well calibrated to support threshold-based medical decisions such as triage and patient escalation. This research investigates asymmetric misclassification costs and probability calibration for binary classification (PNEUMONIA vs. NORMAL) using the Hugging Face dataset hf-vision/chest-xray-pneumonia. The proposed framework utilizes a ResNet-18 architecture integrated with cost-sensitive learning through weighted cross-entropy loss (FN:FP = 5:1), threshold optimization based on validation data to reduce expected cost, and post-hoc temperature scaling for improving probability calibration. Experimental results on the independent test set indicate that the cost-sensitive approach enhances specificity and decreases expected cost compared to the conventional cross-entropy baseline. Furthermore, temperature scaling improves the reliability of probabilistic predictions, as demonstrated by better negative log-likelihood and Brier score values. The study also explores selective prediction strategies to balance prediction coverage and risk reduction, complemented by Grad-CAM visualizations and structured failure-case analysis for qualitative assessment. Overall, the findings demonstrate that incorporating cost-aware decision thresholds and calibrated probability estimates can serve as lightweight yet effective enhancements for chest X-ray classification systems in clinical decision-support applications.

Keywords: Chest X-Ray; Cost-Sensitive Learning; Pneumonia; Probability Calibration; Temperature Scaling.

1. INTRODUCTION

Chest radiography remains a first-line imaging modality for patients with suspected respiratory infection, including pneumonia, because it is fast, inexpensive, and widely available. Over the past several years, deep learning systems have demonstrated strong performance on chest X-ray classification and related tasks, supported by large-scale datasets and benchmark studies such as ChestX-ray8, CheXpert, and MIMIC-CXR (Irvin et al., 2019; Johnson et al., 2019; Wang et al., 2017), as well as high-profile pneumonia detection results exemplified by CheXNet (Rajpurkar et al., 2017). However, translating high-performing classifiers into clinical decision support (CDS) is not simply a matter of maximizing accuracy or AUROC. In real workflows, predictions are often consumed through threshold-based actions (e.g., triage, escalation, additional testing, or watchful waiting), where the type of error matters and the numerical probability must be trustworthy. In particular, false negatives (missed pneumonia) may lead to delayed treatment and worse outcomes, while false positives may primarily incur additional imaging, monitoring, or clinician time. This asymmetry motivates learning and decision rules that reflect explicit costs rather than implicitly treating errors as equally harmful. At the same time, deep neural networks are well known to produce

Naskah Masuk: 15 April 2025; Revisi: 19 Mei 2025; Diterima: 17 Juni 2025; Terbit: 30 Juni 2025

miscalibrated confidence estimates, meaning that a "0.9" probability may not correspond to a 90% empirical correctness rate (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). Such unreliability is especially problematic in CDS settings where probabilities are used to trigger downstream actions, to allocate resources, or to support risk communication (Kelly et al., 2019; Topol, 2019).

Most supervised pipelines for chest X-ray classification optimize symmetric objectives (e.g., standard cross-entropy) and report AUROC as a primary summary metric. While AUROC is valuable, it can be insensitive to clinically relevant operating points: a model may achieve a high AUROCyetstill yield an undesirable sensitivity–specificity tradeoff at plausible thresholds, particularly under class imbalance where precision–recall behavior becomes critical (Saito & Rehmsmeier, 2015). Moreover, even when ranking is strong, probability miscalibration can cause threshold selection to behave unpredictably across splits, sites, or prevalence regimes, undermining the reliability needed for decision thresholds and escalation logic (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). These issues are not merely academic; they are part of the broader set of challenges that limit real clinical impact of AI systems when model outputs are not aligned with operational constraints and safety expectations (Kelly et al., 2019; Topol, 2019).

Motivated by these practical requirements, this work focuses on an end-to-end pipeline that aligns model training and decision-making with two CDS-oriented goals. First, we explicitly incorporate an FN:FP cost ratio into learning and operating-point selection, combining cost aware training with validation-based threshold optimization to minimize expected decision cost (Elkan, 2001). Second, we improve the reliability of probability outputs via post-hoc calibration, so that predicted probabilities better match observed outcome frequencies and can support stable threshold-based actions (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). In addition, because a CDS system may benefit from deferring uncertain cases rather than forcing a potentially high-risk decision, we include an optional selective prediction mechanism that trades coverage for reduced risk/cost by abstaining on low-confidence examples (Chow, 1970; Geifman & El-Yaniv, 2017).

Objectives. Our objective is to build a fully reproducible binary classifier for PNEUMONIA vs. NORMAL using the Hugging Face dataset hf-vision/chest-xray-pneumonia, accessed through the datasets ecosystem (Lhoest et al., 2021). We evaluate standard classification metrics (Ac curacy, Precision, Recall, F1, AUROC, AUPRC, Sensitivity, Specificity) together with decision support metrics that directly reflect deployment concerns, including Expected Cost under an explicit FN:FP cost ratio of 5:1 and calibration quality via

ECE and the Brier score (Brier, 1950; Guo et al., 2017; Saito & Rehmsmeier, 2015). Where relevant, we emphasize that "good ranking" does not guarantee "good decisions" unless operating points and probability reliability are explicitly addressed.

Contributions. This paper makes five concrete contributions. First, we establish a strong and transparent baseline for pneumonia classification using a ResNet-18 backbone (He et al., 2016), trained and evaluated under a reproducible experimental protocol. Second, we imple ment a cost-sensitive variant that aligns learning with asymmetric clinical consequences by using weighted cross-entropy and by selecting decision thresholds on the validation set to min imize expected cost, following foundational principles of cost-sensitive learning (Elkan, 2001) and drawing on practical weighting strategies commonly used for imbalance and hard examples (Cui et al., 2019; Lin et al., 2017). Third, we apply post-hoc temperature scaling to calibrate predictive probabilities and report reliability diagnostics, including ECE, reliability diagrams, and proper scoring via the Brier score (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Fourth, we provide an optional selective prediction (abstention) analysis that quanti f ies the coverage–risk tradeoff, allowing the system to defer uncertain cases in exchange for lower expected cost/risk (Chow, 1970; Geifman & El-Yaniv, 2017). Finally, to support inter pretability and structured error analysis, we include Grad-CAM visualizations and a curated failure-case panel to highlight representative false negatives/false positives and potential failure modes (Ribeiro et al., 2016; Selvaraju et al., 2017). Together, these components form a practi cal CDS-oriented workflow that prioritizes operationally meaningful decisions and trustworthy probabilities rather than relying solely on aggregate AUROC. Figure 1 summarizes the end-to-end pipeline.

2. LITERATURE REVIEW

Deep Learning for Chest X-Ray Classification

Large-scale chest X-ray datasets and benchmark efforts have been a major catalyst for au tomated radiograph interpretation. ChestX-ray8 provided one of the earliest hospital-scale collections for weakly supervised disease classification and localization, enabling broad experi mentation on multi-label thoracic findings (Wang et al., 2017). Subsequent datasets emphasized different aspects of realism and clinical relevance, including label uncertainty and expert com parison (CheXpert) (Irvin et al., 2019) and the pairing of images with free-text reports at scale (MIMIC-CXR) (Johnson et al., 2019). These resources shaped the common experimental fram ing of chest X-ray learning: models are trained as image classifiers, evaluated on held-out splits, and compared primarily through ranking metrics such as AUROC.

Onthe modeling side, convolutional neural networks (CNNs) remain the dominant backbone family for chest X-ray classification because they are computationally efficient and empirically strong. ResNet-style architectures are widely used as reliable baselines for medical image tasks due to their depth scalability and optimization stability (He et al., 2016). CheXNet, for example, used a DenseNet variant to demonstrate strong pneumonia detection on chest X-rays and helped popularize end-to-end CNN pipelines for screening-oriented tasks (Rajpurkar et al., 2017). In practice, these backbones are often trained with standard cross-entropy objectives and modern optimizers (e.g., Adam) within mainstream deep learning frameworks such as PyTorch (Kingma & Ba, 2015; Paszke et al., 2019).

More recent work explores transformer-based vision models and hybrid CNN–transformer designs, frequently reporting incremental AUROC gains. However, much of the deployment discussion in the broader clinical AI literature stresses that achieving clinical impact requires more than high headline metrics: models must integrate with workflow constraints, provide actionable and trustworthy outputs, and be evaluated in a way that matches the intended use (Kelly et al., 2019; Topol, 2019). In particular, thresholded decisions (triage, escalation, or follow-up testing) require careful selection of operating points and reliable probabilities, topics that are often under-emphasized relative to backbone comparisons and AUROC improvements. Related work in other AI application domains likewise shows continuing interest in adaptive frameworks and hybrid deep learning designs when operational deployment requirements remain central (Danang et al., 2025).

Imbalance and Cost-Sensitive Learning

Two practical issues commonly arise in medical screening formulations: class imbalance and asymmetric error costs. Even when a dataset split is moderately balanced, real-world prevalence of disease can differ substantially by site, population, and time, affecting both precision and the downstream burden of false alarms. Under imbalance, AUROC can remain superficially strong while precision may degrade, making the precision–recall curve and AUPRC particularly informative for screening-oriented assessment (Saito & Rehmsmeier, 2015). This matters in CDS settings because clinicians typically experience model performance through thresholded decisions, where false positives translate into additional workload and false negatives represent safety risks.

At the training level, a standard approach to imbalance is to modify the loss to empha size under-represented or harder examples. Focal loss down-weights easy negatives and focuses optimization on difficult samples, which can improve minority-class learning in highly skewed settings (Lin et al., 2017). Class-balanced reweighting based on the effective number

of sam ples provides a principled alternative to naive inverse-frequency weighting, aiming for stable gradients when class counts differ substantially (Cui et al., 2019). These techniques improve learning dynamics, but they do not, by themselves, encode the idea that a false negative may be intrinsically more harmful than a false positive.

Cost-sensitive learning formalizes asymmetric risk by assigning different utilities or costs to different types of errors. A foundational perspective is that the classifier should minimize expected cost (or maximize expected utility) under an explicit cost matrix, rather than minimiz ing symmetric error (Elkan, 2001). Practically, cost sensitivity can be implemented through (i) reweighting (loss weighting), (ii) threshold moving (choosing an operating point that minimizes expected cost), or (iii) transforming scores into calibrated probabilities and then applying a cost based decision rule (Zadrozny & Elkan, 2002). In screening contexts, threshold optimization is often the most direct bridge to deployment: once costs are specified, the optimal threshold follows from minimizing expected cost on representative validation data (Elkan, 2001). This work adopts this pragmatic view by combining a weighted loss with validation-based threshold selection, explicitly aligning the operating point with an FN:FP cost ratio rather than treating thresholding as an afterthought.

Probability Calibration and Reliability

Modern neural networks frequently exhibit miscalibration: predicted probabilities can be sys tematically too confident or too uncertain, even when ranking metrics such as AUROC are high. This disconnect is important for CDS because many downstream actions use a probabil ity threshold (e.g., “if p(pneumonia) > t then escalate”), and miscalibration can cause unstable behavior when prevalence or data composition changes. Calibration quality is therefore a dis tinct axis from discrimination: two models can have similar AUROC yet differ substantially in whether a “0.8” score corresponds to approximately 80% observed positives (Niculescu-Mizil & Caruana, 2005).

Temperature scaling is a widely used post-hoc calibration method that rescales logits by a single scalar parameter fit on a validation set to minimize negative log-likelihood (Guo et al., 2017). A key practical benefit is that it does not alter rank ordering and therefore preserves AU ROCand AUPRC,while improving probability quality as measured by likelihood-based criteria (Guo et al., 2017). Calibration is commonly evaluated visually through reliability diagrams and quantitatively through summary statistics such as Expected Calibration Error (ECE), alongside proper scoring rules that reward accurate probabilities (Niculescu-Mizil & Caruana, 2005). The Brier score is a classic proper scoring rule that measures mean squared

error between predicted probabilities and outcomes, making it interpretable and useful for comparing probabilistic fore casts (Brier, 1950).

In medical classification, reliability is not merely a "nice-to-have" metric; it directly affects safety and trust in automated support. Clinical AI discussions repeatedly highlight that model outputs must be interpretable, dependable, and aligned with clinical decision-making under uncertainty (Kelly et al., 2019; Topol, 2019). For this reason, a CDS-oriented pipeline benefits from reporting both discrimination (AUROC/AUPRC) and reliability (ECE/Brier/NLL) rather than treating probability values as automatically meaningful.

Selective Prediction and Interpretability

A practical approach to increasing safety is to allow the model to abstain on uncertain cases rather than forcing a hard decision for every input. The reject option has a long history in pattern recognition, where an abstaining classifier trades coverage for reduced error by deferring difficult samples (Chow, 1970). Modern selective classification extends this idea to deep neural networks by learning or defining a selection function that decides when a prediction should be issued, often achieving lower risk on the accepted subset of cases (Geifman & El-Yaniv, 2017). In CDS terms, selective prediction can be interpreted as "automate when confident, defer to clinician or additional testing when uncertain," which is operationally compatible with safety-first deployment.

Interpretability complements selectivity by providing qualitative insight into what evidence the model may be using. Gradient-based attribution methods such as Grad-CAM generate coarse localization maps indicating which image regions most influenced a prediction (Selvaraju et al., 2017). While not definitive explanations, these visualizations can help practitioners identify potential spurious correlations or shortcut learning (e.g., text markers, borders, devices) and support structured error analysis. Model-agnostic explanation approaches such as LIME provide another lens by approximating local decision boundaries and highlighting influential features (Ribeiro et al., 2016). In the clinical AI context, interpretability is often framed as part of the broader requirement for trust, accountability, and workflow integration (Kelly et al., 2019; Topol, 2019).

Finally, it is worth noting that many of these pipelines rely on strong but simple imple mentation practices: standardized training recipes, consistent preprocessing, and careful data augmentation. Data augmentation is commonly used to improve generalization in vision mod els, and systematic reviews emphasize its importance for robustness when training data are limited or distribution shifts occur (Shorten & Khoshgoftaar, 2019). Accordingly, transparent

reporting of augmentation and preprocessing choices is important for reproducibility and fair comparison.

Gap and Positioning of this Work

Prior chest X-ray classification studies frequently emphasize backbone choices and AUROC gains, yet a CDS-oriented deployment requires more than discrimination. First, operating points must reflect clinical asymmetry: a false negative can be far more costly than a false positive, so threshold selection should be tied to an explicit cost ratio rather than defaulting to 0.5 or reporting only threshold-free metrics (Elkan, 2001; Saito & Rehmsmeier, 2015). Second, probability estimates must be reliable if they are to support threshold-based actions and resource allocation; miscalibration can undermine trust even when AUROC is high (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Third, safety-aware behavior can benefit from deferring uncertain cases instead of making brittle automated decisions for every input (Chow, 1970; Geifman & El-Yaniv, 2017), and interpretability tools can surface systematic failure modes (Ribeiro et al., 2016; Selvaraju et al., 2017).

This work positions itself as a lightweight, reproducible, and deployment-aware pipeline built on a standard CNN backbone. Rather than proposing a new architecture, we focus on aligning training, thresholding, and probability outputs with practical CDS needs: explicit cost-aware optimization, validation-driven operating points, post-hoc calibration, and optional selective prediction, complemented by qualitative interpretability. This emphasis is consistent with broader clinical AI guidance that stresses operational alignment and trustworthy outputs as prerequisites for real-world impact (Kelly et al., 2019; Topol, 2019). Where applicable, we also follow standard experimental practice by reporting both discrimination and reliability metrics, and by structuring evaluation to support transparent replication using modern tooling (Kingma & Ba, 2015; Lhoest et al., 2021; Paszke et al., 2019).

3. PROPOSED METHOD

Overview

We propose a practical, decision-support–oriented pipeline that extends a standard CNN classi f ier with explicit mechanisms for asymmetric risk, threshold selection, and probability reliability. As summarized in Figure 1, the pipeline consists of (i) image preprocessing and augmenta tion, (ii) a compact CNN backbone (ResNet-18) fine-tuned via transfer learning, (iii) cost sensitive learning through weighted cross-entropy to reflect asymmetric clinical consequences, (iv) validation-based threshold optimization to obtain an operating point that minimizes ex pected cost, (v) post-hoc probability calibration via

temperature scaling to improve reliability, and (vi) optional safety and transparency modules: selective prediction (abstention on low con f idence) and interpretability via Grad-CAM and structured failure-case analysis. The overall design is intentionally lightweight: rather than proposing a new architecture, we emphasize modifications that can be layered onto widely used baselines to produce outputs that are more appropriate for clinical decision support (Kelly et al., 2019; Topol, 2019).

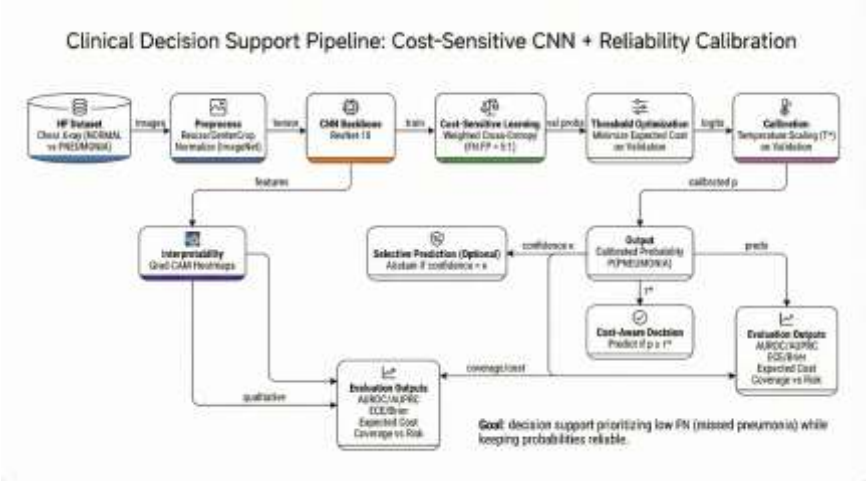

Figure 1. Clinical decision support pipeline: cost-sensitive CNN training and reliability calibration, with optional selective prediction and interpretability outputs**.**

Preprocessing

Each radiograph is converted to three-channel RGB to match the expected input format of ImageNet-pretrained CNN backbones. For evaluation, images are resized and center-cropped to 224×224, then normalized using standard ImageNet statistics, which is a common and ef fective transfer-learning recipe for ResNet-style models (He et al., 2016). During training, we additionally apply lightweight data augmentation (random resized crops and horizontal flips) to improve generalization and reduce overfitting, consistent with prior work showing that augmen tation is an important component of robust deep learning pipelines (Shorten & Khoshgoftaar, 2019). We keep augmentation simple and conservative to avoid introducing unrealistic artifacts that could distort medically relevant signals. All preprocessing operations are deterministic at evaluation time to ensure consistent metrics and to support reproducibility across runs.

Backbone Model

We use ResNet-18 as the backbone due to its strong empirical performance, computational efficiency, and frequent use as a reference architecture in vision research and medical imaging baselines (He et al., 2016). Given a preprocessed input image x, the network produces a two dimensional logit vector $z = f\theta(x)$ corresponding to the NORMAL and

PNEUMONIA classes. We convert these logits into class probabilities using the standard softmax normalization, yielding a probability distribution over the two classes.

We train the backbone using transfer learning from ImageNet-pretrained weights and fine tune all layers end-to-end. Optimization is performed with Adam due to its strong practical performance and widespread adoption (Kingma & Ba, 2015), and the implementation follows standard PyTorch conventions to support reproducible replication (Paszke et al., 2019).

Cost-Sensitive Learning

Clinical decision support often implies asymmetric consequences: missing a true pneumonia case (false negative) can be more harmful than triggering an unnecessary follow-up (false positive). To encode this asymmetry at training time, we adopt a cost-sensitive objective implemented through weighted cross-entropy.

Hrough weighted cross-entropy. Concretely, we assign a larger training weight to the positive (pneumonia) class so that mistakes on pneumonia cases contribute more strongly to the optimization objective. These weights serve two purposes simultaneously: they mitigate class imbalance (so the majority class does not dominate gradient updates) and they reflect asymmetric clinical cost by penalizing false negatives more heavily than false positives (Elkan, 2001). We adopt an explicit FN:FP cost ratio of 5:1 and incorporate this ratio into the positive-class weighting strategy, following the common view that loss reweighting can approximate cost-sensitive risk minimization when the relative costs are known and consistent (Elkan, 2001; Zadrozny & Elkan, 2002).

We note that alternative imbalance-oriented objectives exist, including focal loss to em phasize hard examples (Lin et al., 2017) and class-balanced reweighting based on the effective number of samples (Cui et al., 2019). In this work, we use standard weighted cross-entropy because it is simple, stable in optimization, and straightforward to interpret under an explicit cost-ratio framing.

Threshold Optimization Under Expected Cost

Even with cost-sensitive training, a deployment policy ultimately requires a decision threshold. Rather than defaulting to $\tau = 0.5$, we explicitly select an operating point that minimizes expected cost on a validation split. Given predicted scores or probabilities ˆpi for validation instances and a threshold τ, we define a cost model with false-negative cost cFN and false positive cost cFP. The empirical expected cost is:

$$E[\text{cost}(\tau)] = \frac{c_{\text{FN}} \cdot \text{FN}(\tau) + c_{\text{FP}} \cdot \text{FP}(\tau)}{N} \quad (1)$$

where FN(τ) and FP(τ) are the counts under threshold τ and N is the number of validation examples. We set cFN = 5 and cFP = 1 to match the target FN:FP ratio and search across candidate thresholds to find:

$$\tau^* = \arg\min_{\tau} E\,[\text{cost}(\tau)] \quad (2)$$

This procedure directly links the operating threshold to decision priorities, providing a clear justification for the chosen operating point in CDS settings (Elkan, 2001; Zadrozny & Elkan, 2002). Importantly, threshold optimization is performed on the validation split and then applied to the held-out test split to avoid optimistic bias.

Temperature Scaling Calibration

Because neural networks can be miscalibrated even when discrimination is strong, we apply post-hoc calibration via temperature scaling (Guo et al., 2017). Let z = fθ(x) denote logits. Temperature scaling rescales logits by a scalar T > 0 and produces calibrated probabilities:

$$p_T(y = k \mid x) = \frac{\exp(z_k/T)}{\sum_{j\in\{0,1\}} \exp(z_j/T)} \quad (3)$$

We fit T∗ on the validation split by minimizing negative log-likelihood (equivalently, cross entropy) with respect to T, holding model parameters θ fixed (Guo et al., 2017). This approach is attractive because it preserves the ordering of predictions (and thus typically preserves AU ROC/AUPRC) while improving probability quality. Calibration is assessed using both visual and quantitative measures: reliability diagrams and Expected Calibration Error (ECE), and a proper scoring rule via the Brier score (Brier, 1950; Niculescu-Mizil & Caruana, 2005). Be cause calibration aims to make probabilities meaningful, we treat reliability metrics as first-class outputs alongside discrimination metrics.

Selective Prediction (Optional)

In safety-critical settings, it can be preferable to abstain on uncertain cases rather than forcing an automated decision. We include an optional selective prediction module based on confidence thresholds, grounded in the classical reject-option literature and its modern deep-learning exten sions (Chow, 1970; Geifman & El-Yaniv, 2017). For a calibrated probability pT of the positive class, we define confidence as:

$$\text{conf}(x) = \max(p_T, 1 - p_T). \quad (4)$$

Given a confidence threshold κ, the system predicts when conf(x) ≥ κ and abstains otherwise. We report coverage (the fraction of inputs not abstained) together with the expected cost computed on the accepted subset, yielding a coverage–risk/cost tradeoff curve. In practice, abstention is not free; thus, this analysis should be interpreted as a mechanism for selecting

policies that balance automation with safety and workload constraints (Chow, 1970; Geifman & El-Yaniv, 2017).

Interpretability with Grad-CAM

To support qualitative inspection and debugging, we apply Grad-CAM to visualize spatial re gions that contribute most strongly to the pneumonia prediction (Selvaraju et al., 2017). Grad CAM uses gradients of the target logit with respect to convolutional feature maps to produce a coarse localization heatmap, which we overlay on the input image. These overlays help assess whether the model focuses on clinically plausible lung regions or potentially relies on spurious cues such as markers, borders, or acquisition artifacts. While gradient-based attributions are not definitive causal explanations, they are widely used as practical interpretability tools and can be complemented by model-agnostic approaches such as LIME (Ribeiro et al., 2016). In addition to a qualitative grid of examples, we present a structured failure-case panel composed of high-confidence false positives and false negatives to make systematic errors explicit and to guide future improvements. This interpretability component is aligned with broader clinical AI guidance emphasizing transparency, debugging, and workflow trust when deploying machine learning systems (Kelly et al., 2019; Topol, 2019).

4. RESULT AND DISCUSSION

Experimental Setup

All experiments are conducted on the Hugging Face dataset hf-vision/chest-xray-pneumonia, which provides a binary classification setting for PNEUMONIA vs. NORMAL chest radiographs. Our implementation is written in PyTorch and follows a reproducible evaluation workflow that logs configuration, metrics, and derived artifacts (tables and figures) in a consistent directory struc ture (Paszke et al., 2019). We adopt a transfer-learning protocol using ImageNet-pretrained weights and fine-tune a ResNet-18 backbone as a compact, widely validated baseline for medical image classification (He et al., 2016). Optimization uses Adam with a fixed learning rate sched ule and early monitoring on the validation split (Kingma & Ba, 2015). In addition to standard preprocessing (resize and center crop to 224×224 for evaluation), we apply lightweight data aug mentation during training (random resized crops and horizontal flips) to improve generalization, consistent with established augmentation practice in deep learning (Shorten & Khoshgoftaar, 2019).

A key aspect of our evaluation design is that we report both threshold-free and operating point metrics. Threshold-free metrics (AUROC and AUPRC) summarize ranking performance and enable comparison independent of a particular decision threshold. However, since clini cal

decision support systems act at a threshold (or a small set of thresholds), we also report operating-point metrics including Sensitivity and Specificity, and we compute an expected de cision cost under an explicit FN:FP ratio of 5:1. This cost framing is aligned with cost-sensitive learning principles: the "best" model is not necessarily the one with the highest AUROC, but the one that minimizes expected harm or resource expenditure under a specified utility model (Elkan, 2001). Accordingly, for relevant methods we select operating thresholds using validation-based threshold optimization (threshold moving) and report the chosen threshold τ alongside test performance (Elkan, 2001; Zadrozny & Elkan, 2002). Finally, probability cali bration is performed post-hoc via temperature scaling, with the calibration parameter fit on validation logits and applied to test logits to assess out-of-sample reliability (Guo et al., 2017).

Hardware/Software. Experiments were run on an NVIDIA Tesla T4 GPU using CUDA 12.6, with CPU measurements recorded as a deployment-oriented baseline to approximate con strained settings where GPU access may be limited. All runs were executed in a fixed soft ware environment with explicit version logging and deterministic seeding to support replication (Paszke et al., 2019). The dataset pipeline is handled through the datasets ecosystem, ensuring transparent access and consistent splits across runs (Lhoest et al., 2021).

Main Quantitative Results

Table 1 summarizes the primary test-set results for five configurations: (i) a standard cross entropy baseline evaluated at the default threshold $\tau = 0.5$, (ii) the same baseline augmented with validation-based threshold optimization, (iii) a cost-sensitive model trained with weighted cross-entropy (wCE) evaluated at $\tau = 0.5$, (iv) the full pipeline that adds temperature scaling calibration to wCE, and (v) the calibrated model with threshold optimization. We report both discrimination metrics (AUROC and AUPRC) and operating-point metrics (Sensitivity, Specificity), together with the expected cost computed using FN:FP = 5:1.

Table 1. Test Results and Expected Cost (FN:FP = 5:1).

Method	τ	Acc	F1	AUROC	AUPRC	Sens	Spec	Cost
Baseline (CE)	0.50	0.7115	0.8125	0.9554	0.9637	1.0000	0.2308	0.125
Baseline + thr-opt	0.726	0.7452	0.8307	0.9554	0.9637	1.0000	0.3205	0.518
Cost-sensitive (wCE)	0.50	0.7644	0.8411	0.9445	0.9534	0.9974	0.3761	0.719
FULL (wCE + calib)	0.50	0.7644	0.8411	0.9445	0.9533	0.9974	0.3761	0.2420
FULL + thr-opt	0.490	0.7612	0.8393	0.9445	0.9533	0.9974	0.3675	0.2452

Interpretation. Several observations are important for a decision-support perspective. First, the baseline model exhibits strong ranking performance (AUROC and AUPRC above 0.95) yet yields poor specificity at the default threshold $\tau = 0.5$. This is a concrete example of why threshold-free metrics alone can be misleading for deployment: despite excellent discrimination, the default operating point generates many false positives, which would translate into unneces sary follow-up actions and increased workload in a triage pipeline. Second, simply optimizing the decision threshold on the validation split (baseline + thr-opt) substantially improves ex pected cost and specificity while keeping sensitivity at 1.0. This result supports the practical utility of threshold moving in cost-sensitive decision-making (Elkan, 2001; Zadrozny & Elkan, 2002): if the user can specify a cost ratio, an operating threshold can be selected to better match that preference without changing the underlying ranking.

Third, cost-sensitive training with wCE produces a stronger operating point even before threshold optimization: specificity increases and expected cost decreases relative to the baseline. This behavior is consistent with the intent of asymmetric-risk optimization, where the training loss is shaped to reduce costly mistakes while retaining high sensitivity. Interestingly, wCE slightly reduces AUROC/AUPRC compared to the baseline, yet improves the cost metric and the sensitivity–specificity balance. This tradeoff illustrates a central theme of this paper: for CDS, a modest decrease in a threshold-free ranking metric can be acceptable (or even desirable) if the decision-relevant objective improves.

Finally, the "FULL" configuration (wCE + calibration) matches wCE on discrete classifica tion metrics at $\tau = 0.5$. This is expected: temperature scaling is a monotonic rescaling of logits that typically preserves rank ordering and therefore does not change argmax decisions at the same threshold, while still improving probability quality (Guo et al., 2017). Accordingly, the primary value of calibration should be assessed using reliability-oriented metrics rather than AUROC/AUPRC.

Calibration Analysis

We fit temperature scaling on validation logits and obtained $T* \approx 1.278$, consistent with typical values reported for modern deep neural networks (Guo et al., 2017). Calibration is evaluated using reliability diagrams, ECE, and proper scoring via the Brier score (Brier, 1950; Niculescu Mizil & Caruana, 2005). On the test split, ECE decreases slightly (raw 0.2311 → calibrated 0.2286), while likelihood-based and squared-error metrics show clearer improvements (Brier 0.2051 → 0.1971). This pattern is not unusual: ECE is sensitive to binning choices and sample size, whereas proper scoring rules such as NLL and the Brier score more directly reflect the quality of probabilistic forecasts (Brier, 1950; Niculescu-Mizil & Caruana, 2005). From a de ployment standpoint, the improvement in Brier score indicates that the calibrated probabilities are closer, on average, to empirical outcomes, supporting safer threshold-based decision rules.

Figure 2 provides illustrative reliability diagrams. For clarity and to avoid misleading com parisons, we emphasize that calibration is fit on validation and then applied to test; therefore, the most appropriate visual comparison is typically "test raw" vs "test calibrated" rather than mixing splits. Nevertheless, the diagrams collectively show the qualitative effect of scaling logits to reduce overconfidence, a known issue in deep networks (Guo et al., 2017).

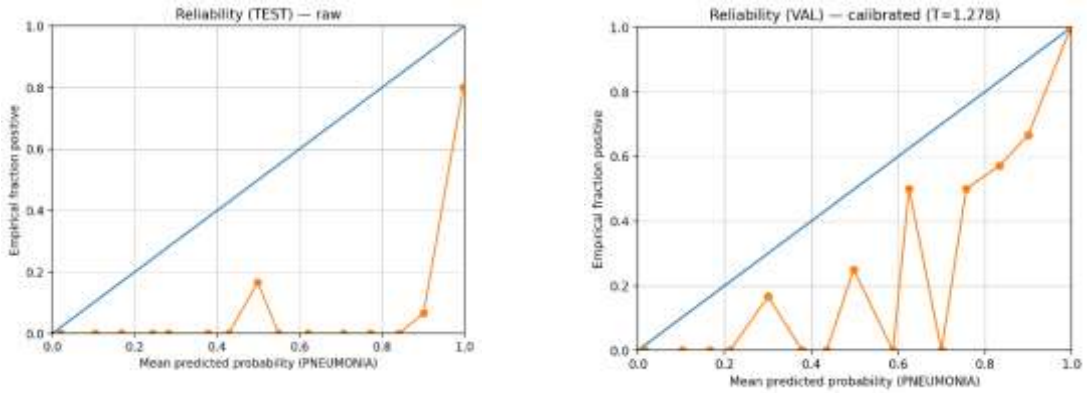

Figure 2. Reliability Diagrams (Examples): Test Raw (Left) and Validation Calibrated (Right).

Discrimination Curves

Figure 3 reports ROC and Precision–Recall curves on the test split. Because pneumonia preva lence and class balance can vary across datasets and real deployment environments, the PR curve is emphasized as it reflects performance on the positive class more directly and is often more informative under imbalance (Saito & Rehmsmeier, 2015). The ROC curve complements this view by summarizing the sensitivity–specificity tradeoff across thresholds, while the PR curve highlights how precision changes as recall increases. Together, these curves reinforce the observation from Table 1: strong ranking performance does not by itself

determine a desirable operating point, which must be selected according to the intended use and cost preferences.

If multiple models are to be compared statistically on AUROC, a standard approach is to apply a nonparametric test such as DeLong's method for correlated ROC curves (DeLong et al., 1988). In this work, our focus is primarily on the cost-aware operating point rather than on marginal AUROC differences; however, DeLong-style testing remains relevant for future extensions that compare alternative backbones or training objectives.

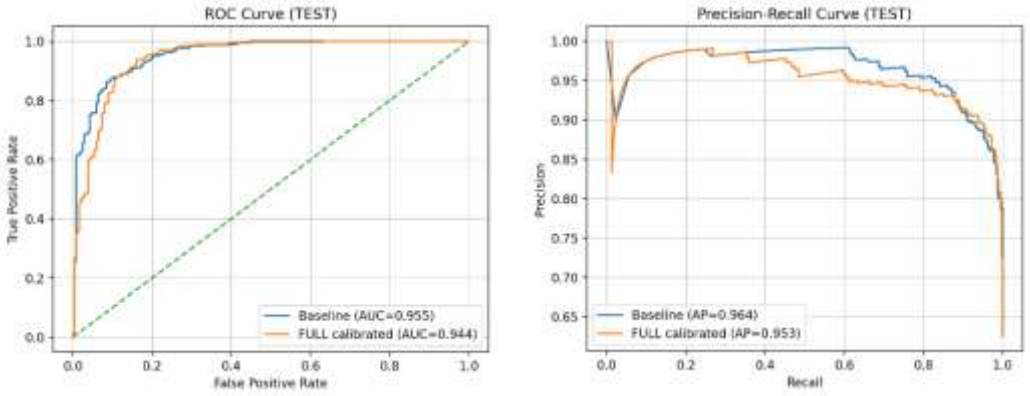

Figure 3. ROC (Left) and Precision–Recall (Right) Curves on the Test Split.

Selective prediction

Selective prediction aims to reduce risk by abstaining on low-confidence inputs, issuing pre dictions only when the model is sufficiently certain (Chow, 1970; Geifman & El-Yaniv, 2017). In a CDS scenario, this corresponds to a policy such as "auto-triage when confident; other wise defer to clinician review or additional testing." We operationalize this idea through a confidence-based selection mechanism and evaluate the coverage–cost tradeoff on the validation set.

Figure 4 shows that as coverage decreases (i.e., the model abstains on more cases), expected cost also decreases. This behavior is desirable when abstention is effectively "free" relative to making an incorrect automated decision, as assumed here (abstain cost = 0). In practice, abstention does carry workflow costs; thus, this analysis is best interpreted as a lower bound on achievable risk reduction and as a tool for choosing a feasible operating policy. Notably, selective prediction can complement cost-sensitive thresholding: thresholding sets the decision boundary for accepted cases, while selection controls which cases are deemed safe enough to automate (Chow, 1970; Geifman & El-Yaniv, 2017).

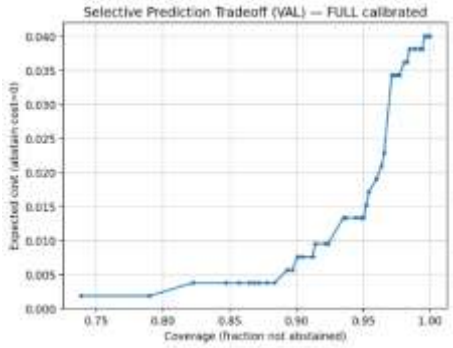

Figure 4. Selective Prediction Tradeoff on Validation (Coverage vs Expected Cost with Abstain).

Interpretability and Failure Analysis

Beyond aggregate metrics, qualitative inspection helps identify potential failure modes and spurious correlations, which is particularly important in medical imaging workflows where model errors may have safety implications (Kelly et al., 2019; Topol, 2019). We use Grad-CAM to visualize coarse spatial attributions indicating which regions most influenced model predictions (Selvaraju et al., 2017). While Grad-CAM does not guarantee causal explanations, it can reveal whether the model tends to focus on anatomically plausible lung regions versus confounding artifacts (text markers, borders, devices), enabling targeted dataset curation and robustness improvements.

Figure 5 presents a 30-sample qualitative grid to provide a broad impression of model at tention patterns across both classes. To complement this overview, we construct a structured failure-case panel (Figure 6) consisting of the top-10 high-confidence errors, separated by false positives and false negatives. This style of error analysis is aligned with interpretability-oriented tooling that aims to support user trust and debugging by presenting concrete, inspectable cases rather than relying solely on averages (Ribeiro et al., 2016; Selvaraju et al., 2017). In the context of cost-sensitive decision support, highlighting false negatives is particularly important because these cases dominate expected harm under an FN-heavy cost ratio.

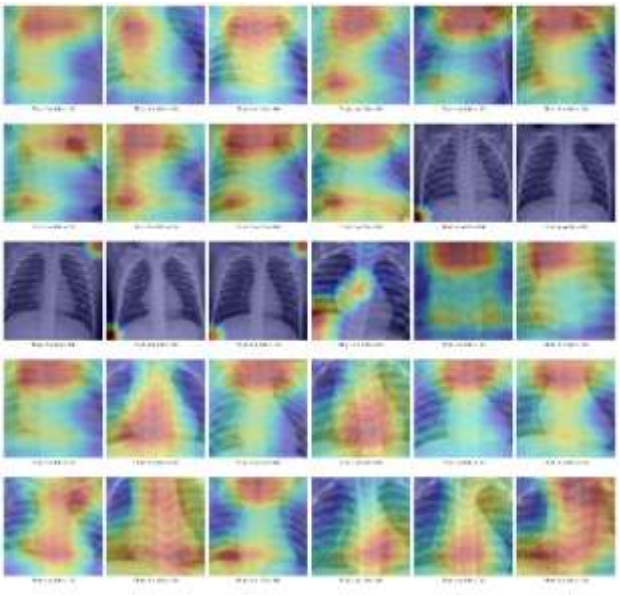

Figure 5. Grad-CAM Qualitative Grid (30 Samples).

Table 2. Top-10 High-Confidence Failure Cases (Panel).

index	type	y_true	y_pred	p_cal	conf
16	FP	0	1	1.000	1.000
144	FP	0	1	1.000	1.000
129	FP	0	1	1.000	1.000
147	FP	0	1	1.000	1.000
145	FP	0	1	0.999	0.999
67	FP	0	1	0.999	0.999
146	FP	0	1	0.999	0.999
143	FP	0	1	0.999	0.999
131	FP	0	1	0.999	0.999
45	FP	0	1	0.999	0.999

Key Discussion Points

Three discussion points summarize the implications of the results. First, high AUROC/AUPRC does not guarantee a clinically useful operating point: a strong ranking model can still behave poorly at default thresholds, producing excessive false positives or failing to meet sensitivity targets. Explicit cost-aware thresholding provides a direct and interpretable mechanism for aligning model behavior with deployment preferences (Elkan, 2001; Zadrozny & Elkan, 2002). Second, cost-sensitive training can improve decision-relevant performance even when threshold free metrics slightly decrease, illustrating that optimizing for CDS objectives may yield different “best” models than optimizing AUROC alone. This is consistent with the broader message that clinical impact depends on aligning evaluation with intended use rather than on a single benchmark metric (Kelly et al., 2019; Topol, 2019).

Third, calibration improvements should primarily be assessed using reliability metrics and proper scoring rules rather than AUROC/AUPRC. Temperature scaling preserves ranking while improving probability quality as reflected by NLL and Brier score, which is consistent with cali bration theory and empirical findings on modern neural networks (Brier, 1950; Guo

et al., 2017; Niculescu-Mizil & Caruana, 2005). From a decision-support perspective, better-calibrated prob abilities enable more stable and defensible threshold-based policies, especially when combined with explicit cost ratios and (optionally) selective prediction. Overall, these results support the central thesis of this work: a lightweight CNN backbone can be made substantially more deployment-relevant through explicit cost-aware decisions, probability calibration, and struc tured qualitative analysis, without requiring architectural novelty

Comparison

Comparison to Baseline

State-of-the-art chest X-ray classification systems are typically developed and reported within a benchmarking paradigm: choose a strong backbone (commonly DenseNet- or ResNet-family CNNs), train on a large dataset, and report discrimination-oriented metrics most prominently AUROC to demonstrate improved ranking performance (He et al., 2016; Rajpurkar et al., 2017). This paradigm is well supported by the availability of large-scale datasets and curated evaluation protocols such as ChestX-ray8, CheXpert, and MIMIC-CXR (Irvin et al., 2019; John son et al., 2019; Wang et al., 2017). Within these benchmarks, incremental gains in AUROC can be meaningful for comparing architectures and training recipes. However, high benchmark AUROC does not automatically translate into a model that is ready for clinical decision sup port. In operational settings, models are not consumed as rankers; they are embedded into workflows where a specific threshold (or a small number of thresholds) triggers actions such as escalation, additional testing, or deferral to human review. Consequently, two additional prop erties become central: (i) explicit operating-point design under asymmetric clinical risk, and (ii) reliable probabilistic outputs for threshold-based policies and risk communication (Elkan, 2001; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005).

In practice, many “SOTA-style” pipelines still train with symmetric losses (standard cross entropy) and report threshold-free metrics, while leaving the choice of operating point either implicit (default $\tau = 0.5$) or tuned informally without an explicit cost model. Likewise, probabil ity calibration is often omitted from the main evaluation despite well-established evidence that modern neural networks can be miscalibrated and overconfident (Guo et al., 2017; Niculescu Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). From a CDS perspective, this omission mat ters: a threshold policy built on miscalibrated probabilities can behave unpredictably across prevalence regimes or minor shifts in data composition, which can undermine safety and clini cian trust (Kelly et al., 2019; Topol, 2019). Therefore, rather than framing our contribution as “beating” prior systems on AUROC (which is highly dependent on

dataset curation, label defi nitions, and evaluation protocol), we frame our contribution as decision-support readiness: a lightweight pipeline that adds cost-aware decision-making, calibration, and safety-oriented analysis to a standard CNN backbone.

Why this Comparison Matters

The feature comparison is not merely a checklist; it clarifies what is required for a classifier to behave sensibly in a decision-support context. Our empirical results show a concrete example of the gap between discrimination and operational utility. From Table 1, the baseline cross entropy model achieves very high AUROC/AUPRC, indicating strong ranking performance, yet behaves poorly at the default threshold with low specificity. In a triage workflow, this would correspond to an excessive number of false alarms, which can overload clinicians and dilute trust in automated recommendations even if sensitivity is excellent. Once an explicit FN:FP cost ratio is specified, threshold optimization provides a principled way to choose an operating point that reduces expected cost (Elkan, 2001; Zadrozny & Elkan, 2002). Moreover, cost sensitive training further improves the sensitivity–specificity tradeoff under the asymmetric risk preference, demonstrating that training objectives aligned with expected cost can yield more deployment-relevant behavior than optimizing symmetric error alone (Cui et al., 2019; Elkan, 2001; Lin et al., 2017).

Calibration adds a complementary dimension: even if two models have similar AUROC/AUPRC, their probability outputs can have very different reliability. Temperature scaling improves the quality of probabilistic predictions without changing ranking, which is exactly the desired be havior when the system must support threshold-based decisions and risk-aware policies (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). This matters for downstream actions such as “escalate if risk > t” or “defer uncertain cases,” where the numeric probability is treated as a meaningful quantity rather than a relative score. Selective prediction extends this logic by explicitly trading coverage for reduced risk, which can be attractive in safety-critical settings when uncertain cases can be routed for additional review (Chow, 1970; Geifman & El-Yaniv, 2017). Finally, interpretability tools and failure-case panels provide practical debugging and transparency mechanisms that help detect spurious attention patterns and systematic errors, supporting safer iteration and stakeholder trust (Ribeiro et al., 2016; Selvaraju et al., 2017).

Limitations

This study has several important limitations that constrain the scope of the conclusions. First, we evaluate on a single public dataset with a specific labeling scheme and imaging distribution; performance and calibration behavior may not transfer directly to other hospitals,

scanners, or patient populations. External validation on independent datasets and sites is essential before any clinical deployment claims can be made, consistent with broader guidance on achieving clinical impact and safe translation of AI systems (Kelly et al., 2019; Topol, 2019). Second, we adopt a fixed FN:FP cost ratio (5:1) as a reasonable illustrative setting, but real clinical costs vary across institutions, prevalence regimes, and workflow designs. In practice, costs should be elicited with domain experts and may be better modeled as a full utility function that accounts for downstream actions, resource constraints, and patient safety priorities (Elkan, 2001).

Third, our cost model is intentionally simplified: it assigns constant costs to false negatives and false positives and does not explicitly represent uncertainty labels or ambiguous findings, which are known challenges in chest X-ray datasets (Irvin et al., 2019). Fourth, while tem perature scaling improves calibration under the current experimental protocol, calibration can degrade under distribution shift and changes in prevalence; future work should study recalibra tion policies, monitoring, and robustness of probability reliability over time (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Finally, interpretability methods such as Grad-CAM provide useful qualitative signals but do not guarantee causal explanations; they should be treated as debugging and communication aids rather than definitive clinical evidence (Ribeiro et al., 2016; Selvaraju et al., 2017).

Despite these limitations, the results support a clear conclusion: making a chest X-ray classifier more relevant for decision support requires explicit attention to operating points under asymmetric cost and to the reliability of probability outputs, not only to AUROC.

5. CONCLUSIONS

This paper presented a lightweight yet deployment-oriented pipeline for pneumonia classification that explicitly targets requirements commonly encountered in clinical decision support rather than optimizing only for threshold-free discrimination metrics. Using a ResNet-18 backbone as a transparent and widely adopted baseline (He et al., 2016), we incorporated cost-sensitive training through weighted cross-entropy to reflect asymmetric clinical risk under an explicit FN:FP cost ratio of 5:1, and we paired this with validation-based threshold optimization to se lect an operating point that minimizes expected decision cost (Elkan, 2001; Zadrozny & Elkan, 2002). To address the well-known issue that modern neural networks can produce miscalibrated confidence estimates, we applied post-hoc temperature scaling and evaluated reliability using ECE, reliability diagrams, and proper scoring via the Brier score (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Empirically, we observed

that cost-sensitive optimization improves the decision-relevant sensitivity–specificity balance and reduces expected cost relative to a standard cross-entropy baseline, even when AUROC changes only marginally. Calibra tion yielded improvements in probability quality as reflected by proper scoring metrics (NLL and Brier), illustrating that reliability gains can occur without changing ranking performance, consistent with calibration theory (Guo et al., 2017). Finally, we demonstrated two practi cal add-ons for safety and interpretability: selective prediction to trade coverage for reduced risk/cost in uncertain cases (Chow, 1970; Geifman & El-Yaniv, 2017), and Grad-CAM plus a structured failure-case panel to support qualitative inspection and debugging of systematic errors (Ribeiro et al., 2016; Selvaraju et al., 2017).

Implications, the primary implication is that decision-support readiness can be improved substantially with simple, well-scoped modifications to a standard chest X-ray classifier. Ex plicit cost-aware optimization and threshold selection provide a principled mechanism to align model behavior with asymmetric clinical priorities (Elkan, 2001), while probability calibration helps ensure that numeric risk estimates are meaningful for threshold-based actions and escala tion logic (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Taken together, these additions move the evaluation emphasis from "high AUROC" toward "operationally defensible decisions," which is consistent with broader clinical AI guidance that stresses workflow integration, trust worthiness, and safe translation as prerequisites for impact (Kelly et al., 2019; Topol, 2019). In practical settings, this framing can support clearer stakeholder communication: model devel opers can justify operating thresholds in terms of explicit costs, monitor calibration over time, and adopt abstention policies when automation is uncertain.

Limitations and future work, this work has limitations that motivate several concrete next steps. First, our experiments are restricted to a single public dataset and a fixed la beling scheme; external validation on independent hospital datasets and across institutions is required before drawing conclusions about general clinical utility (Irvin et al., 2019; Johnson et al., 2019; Kelly et al., 2019; Topol, 2019). Second, the FN:FP cost ratio is illustrative and context-dependent; real deployments should elicit costs with domain experts and may bene f it from richer utility models that incorporate downstream actions, resource constraints, and patient safety objectives (Elkan, 2001). Third, calibration behavior can degrade under distribu tion shift and prevalence changes; future work should investigate monitoring and recalibration strategies, and assess subgroup-specific reliability (Guo et al., 2017; Niculescu-Mizil & Caru ana, 2005). Fourth, while temperature scaling is simple and effective, alternative calibration approaches such as isotonic regression have been studied in the broader probability estimation

literature and remain a relevant baseline for comparison (Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). Finally, improving robustness against shortcut learning and spuri ous correlations remains essential for trustworthy medical imaging systems; future work should combine stronger augmentation and robustness-oriented training with structured error analy sis to better understand failure modes (Ribeiro et al., 2016; Selvaraju et al., 2017; Shorten & Khoshgoftaar, 2019).

REFERENCES

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory, 16*(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2019.00949

Danang, D., Wahyono, T., Sembiring, I., Wellem, T., & Dzulkefly, N. H. (2025, August). An adaptive framework integrating ML blockchain and TEE for cloud security. In *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICCIT65724.2025.11167152

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

Johnson, A. E. W., Pollard, T. J., Berkowitz, S., 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. *Scientific Data, 6*, 317. https://doi.org/10.1038/s41597-019-0322-0

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, 195. https://doi.org/10.1186/s12916-019-1426-2

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Lhoest, Q., del Moral, V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., et al. (2021). Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 175–184). https://doi.org/10.18653/v1/2021.emnlp-demo.21

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)* (pp. 625–632). https://doi.org/10.1145/1102351.1102430

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, *32*.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. https://arxiv.org/abs/1711.05225

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*, 60. https://doi.org/10.1186/s40537-019-0197-0

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2097–2106). https://doi.org/10.1109/CVPR.2017.369

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). https://doi.org/10.1145/775047.775151