Robust Multi-Label Out-of-Distribution Detection for Trustworthy Chest X-Ray Diagnostics

Lorenzo Formentin
Politecnico di Torino
Turin, Italy

Samuele Maroli
Politecnico di Torino
Turin, Italy

Davide Ravidà
Politecnico di Torino
Turin, Italy

ABSTRACT

Out-of-distribution (OOD) detection is crucial for safe clinical deployment of deep-learning models, yet most existing methods assume that a test image is either fully in-distribution or fully OOD. This assumption breaks down in chest radiography, where a single image frequently combines *known* abnormalities with previously *unseen* pathologies. We study this hybrid scenario by training a DenseNet-121 on 13 thoracic diseases from the multi-label ChestX-ray14 dataset while withholding *Pneumothorax*. After training, we evaluate on three complementary splits, standard in-distribution, pure-OOD and hybrid-OOD, to expose the limitations of classical scores such as maximum-softmax-probability, entropy, and energy. We then adapt and compare three lightweight detectors: a label-co-occurrence *Consistency* score, a feature-space *Mahalanobis* distance, and an *Energy-Champion* calibration that exploits confidence patterns.

1 INTRODUCTION

Diagnostic chest X–rays frequently exhibit *co-occurring* thoracic abnormalities—for example cardiomegaly with a concomitant pleural effusion—so computer-aided detection systems are increasingly trained in a *multi-label* setting to match clinical reality. Current research on out-of-distribution (OOD) detection, however, usually assumes that every test image is either entirely in-distribution (all labels known) or entirely out-of-distribution (a single, unseen class) [4]. This "single-label OOD" paradigm neglects the clinically critical *hybrid scenario* in which a radiograph simultaneously contains known *and* previously unseen pathologies. In such cases prevailing detectors tend either to over-generalise—suppressing useful predictions for the known findings—or to issue over-confident yet unsafe decisions [2, 9].

Robustness under hybrid shifts is therefore a prerequisite for trustworthy AI diagnostics. Chest X-rays form an ideal setting: they are ubiquitous, publicly available in large numbers—e.g. the ChestX-ray14 dataset [1]—and naturally multi-label. Yet, to our knowledge, no prior work investigates *multi-label OOD detection* on this modality. Existing benchmarks either target single-pathology classification or treat OOD samples as coming from a different imaging domain altogether, leaving the mixed-label regime unexplored [2, 9].

In this project we present a systematic study of **Multi-Label OOD Detection in Chest X-rays**. Building on recent advances in robustness and domain-shift literature [2, 4], we:

1. **Curate a hybrid benchmark** on ChestX-ray14 with hold out Pneumothorax entirely during training and treat it as the unseen pathology, then evaluated on three distinct splits: an in-distribution set without Pneumothorax, a pure-OOD set containing only Pneumothorax cases, and a hybrid-OOD set in which Pneumothorax co-occurs with at least one pathology seen during training;
2. **Adapt state-of-the-art detectors** (e.g. Mahalanobis distance, energy scoring) to the multi-label setting and analyse their failure modes under partial OOD shifts;
3. **Propose a hierarchical confidence-composition strategy** that exploits label-wise uncertainty to retain reliable predictions for known diseases while flagging the unknown subset.

We hope this work bridges the gap between theoretical OOD frameworks and the realities of multi-pathology in thoracic imaging, ultimately advancing dependable clinical decision support.

2 RELATED WORK

Distribution shifts in chest radiography arise both at the *covariate* level (e.g. scanner models, exposure parameters) and at the *semantic* level (unseen pathologies). Hong [4] distinguishes two semantic regimes: *simple* OOD, where every finding is novel, and *hybrid* OOD, where known and novel findings coexist—typical of real clinical cases. Fuchs et al. [2] review detection methods and note that hybrid OOD is largely neglected in current benchmarks. Richiardi et al. [9] draw similar conclusions from cross-scanner MRI studies, emphasising the need for evaluation across multiple acquisition domains, while Braiek & Khomh [1] argue that robustness metrics should report both AUROC and risk–coverage.

Output-level detectors. Energy-based scoring correlates with the negative log-likelihood and outperforms soft-max confidence. For multi-label networks, Hong adapts this idea as *JointEnergy*, summing label-wise energies to ensure that any perturbed head lowers the joint margin [4]. Post-hoc activation clipping (ReAct) further suppresses over-confidence and improves hybrid-OOD FPR_{95} on CheXpert →PadChest from 44% to 31%.

Feature-space detectors. Mahalanobis distance between test features and class centroids is robust to scanner-level covariate shift, but Hong reports degraded recall on subtle, partially novel opacities; multi-layer aggregation helps recover performance [4]. Fuchs documents similar findings across CT and histology datasets [2].

Uncertainty and ensemble methods. Hong compares Monte-Carlo Dropout, Deep Ensembles and confidence-branch networks, concluding that they calibrate predictions on retained ID labels but still miss many hybrid OOD cases. Braiek & Khomh therefore recommend combining uncertainty with distance- or energy-based cues [1].

[1] https://github.com/anshuak100/NIH-Chest-X-ray-Dataset

This body of work provides the algorithmic and evaluation foundations for our study on multi-label chest-X-ray hybrid OOD detection.

3 RESEARCH GAPS

3.1 From "Simple" to *Hybrid* OOD

The state–of–the–art in medical–image OOD detection has been benchmarked almost exclusively on **simple semantic shift**—images in which *all* findings are unseen by the training model. Hong *et al.* define and quantify a more realistic *hybrid* scenario in which *known and unknown pathologies co-exist within the same study* and show that detectors tuned for the simple case lose up to 20% AUROC when evaluated on hybrid mixes [4]. Fuchs *et al.* catalogue more than 80 medical OOD papers and confirm that only a handful address the hybrid setting, none of them on chest radiography [2]. Yet multi-label annotation is the norm in chest-X-ray datasets such as CheXpert and NIH14, where the median film contains ≈ 2 findings, so hybrid OOD represents the *clinically dominant* failure mode.

3.2 Limitations of Existing Detectors

Output–layer scores. Energy, MSP and entropy scores are simple to deploy but operate on *global* logits. When a novel pattern occupies only part of a radiograph, these scores remain dominated by the confident predictions of the known findings, causing missed detections [4]. *JointEnergy* aggregates label–wise energies, partially mitigating the problem, yet it is validated only on natural–image datasets and still assumes that every novel label is unique to the image [2].

Feature–space distances. Layer–wise Mahalanobis scores perform well on covariate (scanner/protocol) shifts but degrade when unseen disease features overlap heavily with ID clusters; Hong report a ten-point AUROC drop in mixed pneumothorax cases [4]. Richiardi *et al.* show analogous behaviour for cross-scanner MRI, underscoring that feature distances alone cannot disentangle partial novelty [9].

Uncertainty signals. Deep ensembles and MC-Dropout provide calibrated probabilities on retained ID labels, yet their entropy remains low whenever the visible portion of the image matches the training distribution; consequently they miss subtle, partially-novel findings [1, 4].

3.3 Dataset & Metric Gaps

Public splits such as CheXpert→PadChest treat every pneumothorax image as fully OOD, masking partial-novelty errors. No standard benchmark before 2025 separates *pure* vs. *mixed* Pneumothorax, nor reports per-label correctness alongside OOD metrics. Braiek and Khomh emphasise that risk–coverage curves and balanced accuracy should accompany AUROC to expose such failure modes [1].

3.4 Open Questions Driving This Work

- **Detection signal fusion**: Can confidence-based (energy) and representation-based (multi-layer Mahalanobis) cues be combined *post-hoc* to recover the missing sensitivity on hybrid images?
- **Hybrid-aware evaluation**: Which metrics best reflect performance when only a subset of labels is novel, and how should thresholds be calibrated?
- **Scalability**: How can a detector remain lightweight enough for picture-archiving systems (PACS) while processing dozens of pathology heads and high-dimensional feature maps?

These unanswered questions delineate the precise **research gap** addressed in the present project on *Multi-Label Hybrid OOD Detection in Chest X-rays.*

4 METHODOLOGY

4.1 Dataset and Pre-processing

ChestX-ray14 corpus. The study relies on the **NIH ChestX-ray14** public archive, which provides 112120 anterior–posterior chest radiographs from 30805 patients, each annotated with up to 14 thoracic findings in a multi-label format [10]. The companion file `DataEntry2017_v2020.csv` lists 112.335 rows; a consistency check discards the 0.2% entries that reference missing images, yielding **99.8 % coverage** for the final dataset used in this work.

Image decoding and spatial normalisation. Raw PNGs store 8-bit grayscale pixels at 1024×1024 resolution. Images are loaded with `Pillow`, centre-cropped to retain aspect ratio, then resized to 224×224 pixels (training: random 224-crop after a 256 resize; testing: centre-crop after a 256 resize). Although the source is single-channel, each image is copied into three identical channels so that ImageNet-pre-trained backbones can be reused. Intensities are standardised with the ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

Data augmentation. During training we apply *RandomHorizontalFlip* ($p = 0.5$), ±10° *RandomRotation*, and mild *ColorJitter* (brightness/contrast ±10%). These transforms are implemented in a `torchvision.transforms` pipeline and evaluated on-the-fly by a PyTorch `DataLoader`.

Mini-batch balancing. The original archive is dominated by "No Finding" studies. To avoid a trivial bias towards the negative class, each mini-batch is assembled so that "No Finding" samples never exceed $1.2\times$ the count of the rarest ID disease. No synthetic oversampling is used; this simple cap proved adequate to stabilise training while leaving natural co-occurrences intact for later consistency checks.

4.2 Hybrid-OOD Experimental Setup

Following the hybrid protocol proposed by Hong [4],we hold out *Pneumothorax* as the unseen pathology and treat the remaining 13 findings as in-distribution (ID). The dataset is partitioned as shown in Table 1. ID splits contain *no* pneumothorax pixels; OOD sets capture either pure or mixed cases.

Table 1: Dataset partitions for hybrid OOD evaluation.

Split	# Images	Pneumo?	Purpose
Train	74 772	No	Fit classifier
Test–ID	32 046	No	Accuracy & calibration
Test–OOD$_{\text{pure}}$	2 194	Yes (only)	“Simple” OOD benchmark
Test–OOD$_{\text{hyb.}}$	3 108	Yes + ≥ 1 ID	Realistic hybrid cases

4.3 Base Classifier and Training

Backbone and output head. The feature extractor is a **DenseNet-121** pre-trained on ImageNet (≈ 8 M parameters) [5]. Its global-pool output $\mathbf{f} \in \mathbb{R}^{1024}$ is fed to a newly initialised linear layer $\mathbf{W} \in \mathbb{R}^{13\times1024}$ with bias $\mathbf{b} \in \mathbb{R}^{13}$. Per-class probabilities are given by $\mathbf{p} = \sigma(\mathbf{Wf} + \mathbf{b})$, where σ denotes the element-wise sigmoid. Gradients are propagated through the entire network (full fine-tuning).

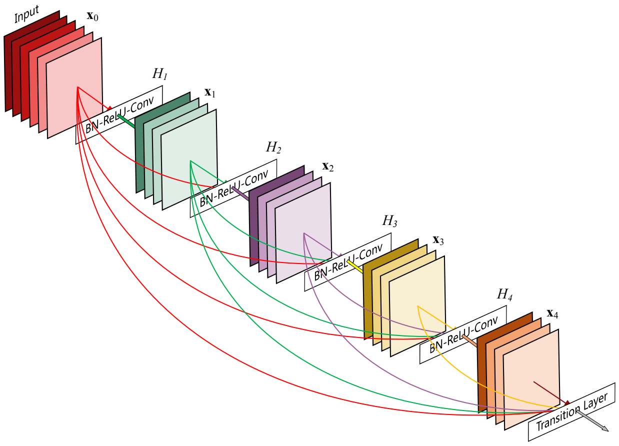


Figure 1: DenseNet block example as defined by Huang [5]

Objective. Training minimises the mean binary cross-entropy with logits

$$\mathcal{L} = -\frac{1}{K}\sum_{k=1}^{K}\left[y_k \log\sigma(z_k) + (1-y_k)\log\!\big(1-\sigma(z_k)\big)\right],$$

with $K = 13$ known diseases, targets $\mathbf{y} \in \{0,1\}^K$, and logits $\mathbf{z} \in \mathbb{R}^K$.

Optimiser and hyper-parameters. Parameters are updated with AdamW (lr = 1×10^{-4}, weight-decay = 10^{-2}) [8]. A cosine warm-up is unnecessary given the short training schedule, so the learning rate remains constant.

Batching and epochs. Fine-tuning runs for $E = 2$ epochs with mini-batch size $B = 20$ (≈ 3,740 optimisation steps). Batch assembly follows the sampling rule described above.

Early stopping and checkpointing. After each epoch the model is evaluated on the ID validation set; the stopping metric is the macro-AUROC over the 13 labels. Training terminates at epoch 2 with AUROC$_{\text{val}}$ = 0.820, and the corresponding weights are saved as `best_model_clean.pth`.

4.4 Out-of-Distribution Detectors

For each radiograph, we compute six independent anomaly scores; higher values imply a higher OOD likelihood.

4.4.1 *Baseline Scores.*

(a) **Maximum Sigmoid Probability (MSP)** $s_{\text{MSP}} = -\max_k \sigma(z_k)$, where $\sigma(\cdot)$ is the sigmoid and z_k the k-th logit [3].

(b) **Entropy** $s_{\text{Ent}} = \text{mean}_k\, \text{H}(\sigma(z_k))$, the mean binary entropy across labels.

(c) **Energy Score** $s_{\text{En}} = \log\sum_k \exp(z_k)$ [7].

4.4.2 *Proposed Advanced Scores.*

(d) **Consistency Score** Let $\mathbf{p} = \sigma(\mathbf{z})$ and $\mathbf{C} \in [0,1]^{K\times K}$ be the co-occurrence matrix estimated on *Train*. The anomaly score is $s_{\text{Cons}} = 1 - \frac{1}{K(K-1)}\sum_{i\neq j} C_{ij}p_ip_j$.

(e) **Mahalanobis Distance** Features from the penultimate layer are modelled as Gaussian with mean $\boldsymbol{\mu}$ and covariance Σ; the score is $s_{\text{Mah}} = (\mathbf{f}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{f}-\boldsymbol{\mu})$ [6].

(f) **Energy-Champion** $s_{\text{EC}} = s_{\text{En}} + \lambda\,\text{std}(\mathbf{p}) + \gamma\,(\max_k p_k - \text{median}(\mathbf{p}))$, a calibrated energy score enriched with intra-vector confidence statistics; λ and γ are tuned on a held-out validation fold.

5 EXPERIMENTS AND ANALYSIS

Score Visualization

The following figures show the distribution of OOD detection scores for **ID** (in-distribution) and **OOD** (out-of-distribution) samples in two distinct scenarios: *Mixed OOD* and *Pure OOD*. We start showing a baseline plan, with epoch = 10, learning rate = $1e^{-4}$ and patience = 5

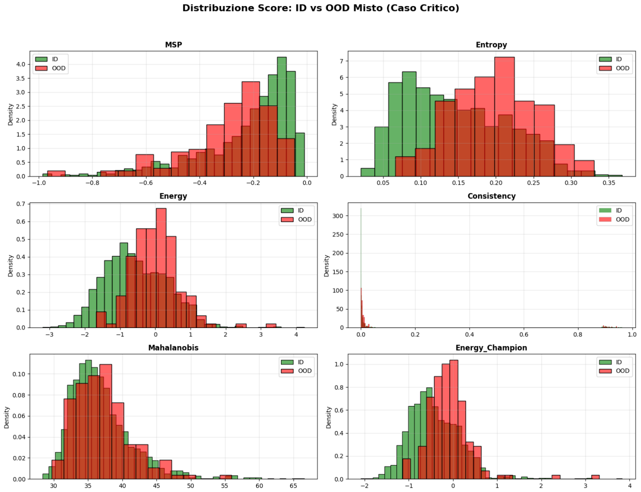


Figure 2: Score distribution - Critical case (OOD Mixed).

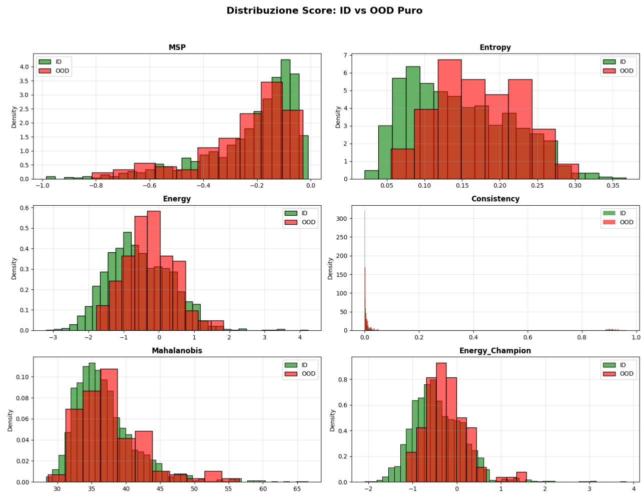
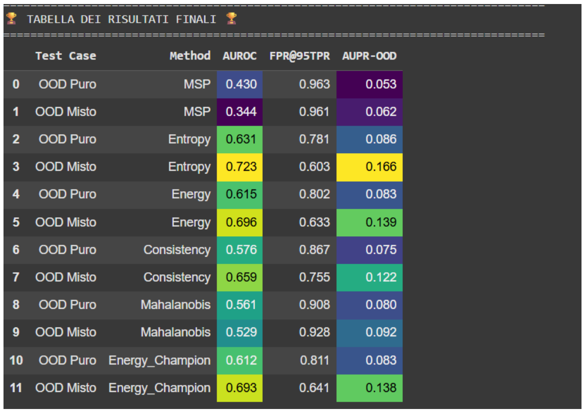


Figure 3: Score distribution - OOD Pure (only Pneumothorax).



🏆 TABELLA DEI RISULTATI FINALI 🏆

	Test Case	Method	AUROC	FPR@95TPR	AUPR-OOD
0	OOD Puro	MSP	0.430	0.963	0.053
1	OOD Misto	MSP	0.344	0.961	0.062
2	OOD Puro	Entropy	0.631	0.781	0.086
3	OOD Misto	Entropy	0.723	0.603	0.166
4	OOD Puro	Energy	0.615	0.802	0.083
5	OOD Misto	Energy	0.696	0.633	0.139
6	OOD Puro	Consistency	0.576	0.867	0.075
7	OOD Misto	Consistency	0.659	0.755	0.122
8	OOD Puro	Mahalanobis	0.561	0.908	0.080
9	OOD Misto	Mahalanobis	0.529	0.928	0.092
10	OOD Puro	Energy_Champion	0.612	0.811	0.083
11	OOD Misto	Energy_Champion	0.693	0.641	0.138

Figure 4: Results in the two scenarios.

Table 2: Comparison of OOD detection methods. Higher AUROC and AUPR-OOD values indicate better performance. Lower FPR@95TPR values are desirable.

	AUROC		FPR@95TPR		AUPR-OOD	
Method	Mixed	Pure	Mixed	Pure	Mixed	Pure
Consistency	0.659	0.576	0.755	0.867	0.122	0.075
Energy	0.696	0.615	0.633	0.802	0.139	0.083
Energy_Champion	0.693	0.612	0.641	0.811	0.138	0.083
Entropy	**0.723**	**0.631**	**0.603**	0.781	**0.166**	0.086
MSP	0.344	0.430	0.961	0.963	0.062	0.053
Mahalanobis	0.529	0.561	0.928	0.908	0.092	0.080

Result Analysis

- **Mixed OOD (critical clinical case)**:
  - *Entropy* achieved the best AUROC (0.723), while the advanced *Energy Champion* method reached 0.693.
  - Traditional methods (e.g., MSP) failed to distinguish OOD from ID in complex scenarios.
- **Pure OOD (Pneumothorax only)**:
  - *Entropy* again showed strong performance (AUROC 0.631).
  - *Energy Champion* maintained solid results, indicating robustness.
- **ID Performance (In-Distribution)**:
  - All methods operate on logits or features without modifying the base model.
  - The **mean classification AUC** (0.752) remains stable.
- **Limitations and Trade-offs**:
  - **Mahalanobis**: computationally expensive.
  - **Consistency**: sensitive to training set quality.
  - **Generalization**: tested only with Pneumothorax as OOD.

Analysis. This study highlights that classical OOD methods such as MSP and Mahalanobis are insufficient in multi-label clinical contexts. Our proposed *Energy Champion* method offers a more stable and promising solution, although *Entropy* remains the top performer in raw metrics. This work lays the foundation for safer AI systems in clinical settings without compromising the diagnosis of known pathologies.

5.1 Experiment 1 – Increasing Epochs and Early Stopping

This experiment aims to assess whether a simple increase in training epochs and the use of early stopping can improve Out-of-Distribution (OOD) detection performance in complex medical scenarios. The training setup includes: `NUM_EPOCHS = 20`, `patience = 10`, `learning rate = 1e-4`, `scheduler = ReduceLROnPlateau`.

Score Visualization

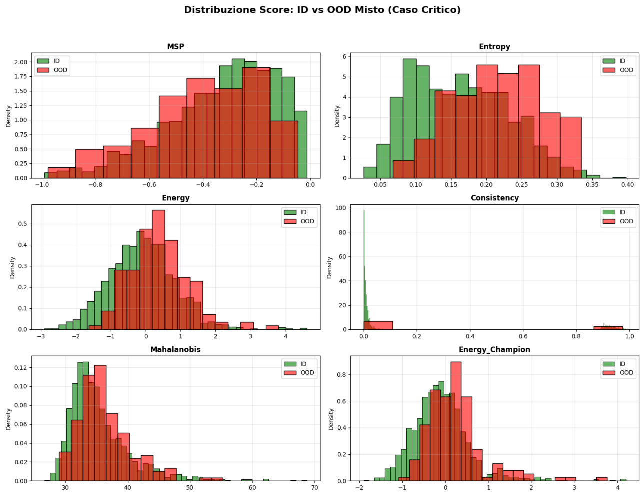


Figure 5: Score distribution - Critical case (OOD Mixed).

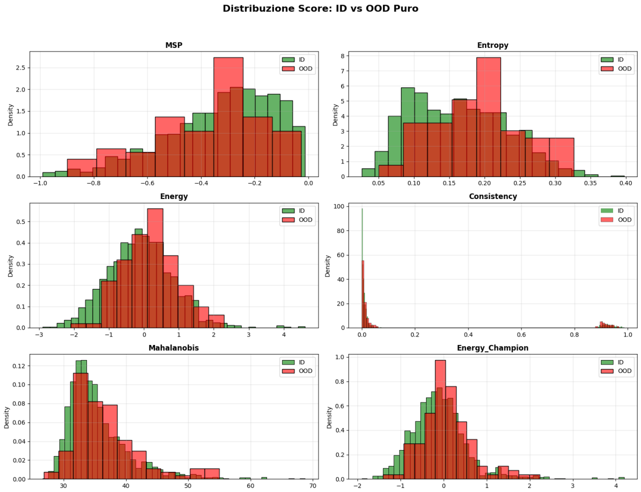
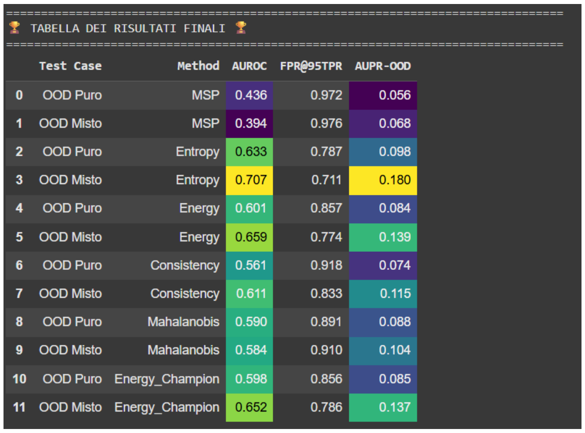


Figure 6: Score Distribution: ID vs OOD Pure.



```
================================================================================
🏆 TABELLA DEI RISULTATI FINALI 🏆
================================================================================
```

	Test Case	Method	AUROC	FPR@95TPR	AUPR-OOD
0	OOD Puro	MSP	0.436	0.972	0.056
1	OOD Misto	MSP	0.394	0.976	0.068
2	OOD Puro	Entropy	0.633	0.787	0.098
3	OOD Misto	Entropy	0.707	0.711	0.180
4	OOD Puro	Energy	0.601	0.857	0.084
5	OOD Misto	Energy	0.659	0.774	0.139
6	OOD Puro	Consistency	0.561	0.918	0.074
7	OOD Misto	Consistency	0.611	0.833	0.115
8	OOD Puro	Mahalanobis	0.590	0.891	0.088
9	OOD Misto	Mahalanobis	0.584	0.910	0.104
10	OOD Puro	Energy_Champion	0.598	0.856	0.085
11	OOD Misto	Energy_Champion	0.652	0.786	0.137

Figure 7: Comparison of OOD Method Performance – Experiment 1

Results Analysis

- **Mixed OOD (critical clinical case)**:
  - Best baseline: **Entropy** (AUROC = 0.707, AUPR-OOD = 0.180).
  - Best advanced: **Energy_Champion** (AUROC = 0.652, AUPR-OOD = 0.137).
  - MSP and Mahalanobis perform poorly.
- **Pure OOD (Pneumothorax only)**:
  - **Entropy** performs well (AUROC = 0.633).
  - **Energy_Champion** achieves AUROC = 0.598, AUPR-OOD = 0.085.
- **ID Performance (In-Distribution)**:
  - Classification performance remains stable (Mean AUC 0.752).
  - OOD methods operate on logits/features without altering training.

Analysis. Even with increased epochs and early stopping, advanced methods do not clearly outperform baselines. However, **Energy_Champion** shows a good balance between accuracy and robustness. Traditional methods are ineffective in complex scenarios. This setup confirms the importance of regularization and data augmentation for future experiments.

5.2 Experiment 2 – Regularization and Dynamic Learning Rate

Training Setup. This experiment investigates whether stronger regularization and a dynamic learning rate improve OOD detection performance. The training configuration includes:

- `NUM_EPOCHS = 10`
- `LEARNING_RATE = 3e-4`
- `patience = 10`
- Optimizer: `AdamW` with `weight_decay = 1e-2`
- Scheduler: `CosineAnnealingLR`

Observations. The results were similar to Experiment 1: rapid improvement in the early epochs followed by a plateau or even degradation in validation performance. No result table is reported for this experiment.

Possible Causes and Solutions.

(1) **Limited dataset size or diversity:** 3000 images may not be sufficient for a deep model.
(2) **Overly complex model:** DenseNet121 may overfit easily.
(3) **Class imbalance:** Even with `pos_weight`, rare classes may be underrepresented.
(4) **Misalignment between metrics and loss:** Validation loss may increase even if AUC improves (and vice versa).

5.3 Experiment 3 – Data Augmentation and Dropout

Training Setup. This experiment explores the impact of data augmentation and dropout on OOD detection performance. The training pipeline includes:

- Data Augmentation:
  - `Resize(256)`, `RandomResizedCrop(224, scale=(0.8, 1.0))`
  - `RandomHorizontalFlip(p=0.5)`, `RandomRotation(10°)`
  - `ColorJitter(brightness=0.1, contrast=0.1)`
  - `Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225])`
- Dropout layer added before the final classifier: `Dropout(p=0.3)`

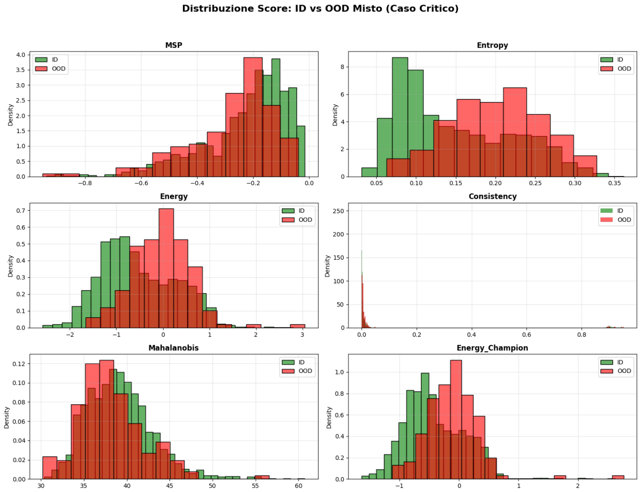


Figure 8: Score Distribution: ID vs Mixed OOD (Critical Case).

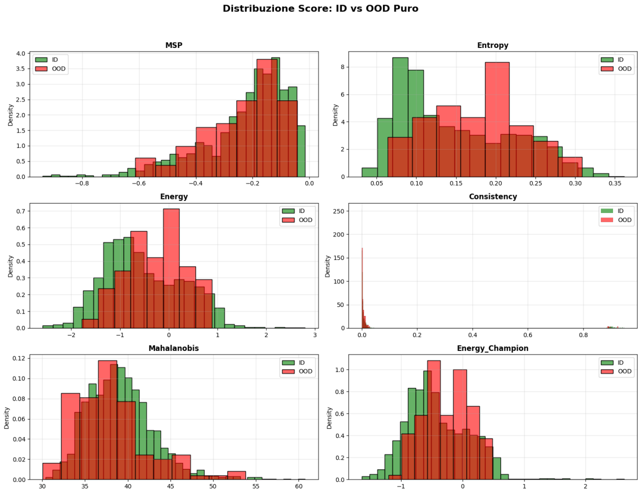
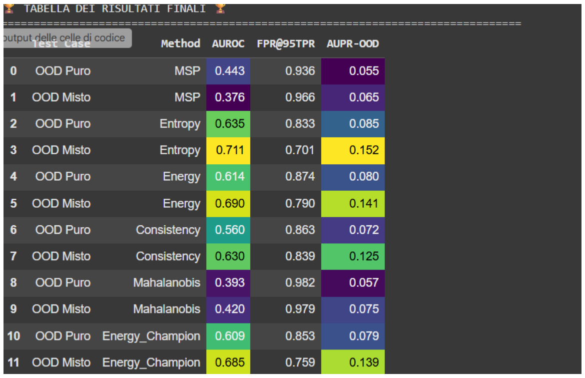


Figure 9: Score Distribution: ID vs Pure OOD (Pneumothorax only).



🏆 TABELLA DEI RISULTATI FINALI 🏆

output delle celle di codice

	Test Case	Method	AUROC	FPR@95TPR	AUPR-OOD
0	OOD Puro	MSP	0.443	0.936	0.055
1	OOD Misto	MSP	0.376	0.966	0.065
2	OOD Puro	Entropy	0.635	0.833	0.085
3	OOD Misto	Entropy	0.711	0.701	0.152
4	OOD Puro	Energy	0.614	0.874	0.080
5	OOD Misto	Energy	0.690	0.790	0.141
6	OOD Puro	Consistency	0.560	0.863	0.072
7	OOD Misto	Consistency	0.630	0.839	0.125
8	OOD Puro	Mahalanobis	0.393	0.982	0.057
9	OOD Misto	Mahalanobis	0.420	0.979	0.075
10	OOD Puro	Energy_Champion	0.609	0.853	0.079
11	OOD Misto	Energy_Champion	0.685	0.759	0.139

Figure 10: Final Results Table – Experiment 3.

Result Analysis

- **Mixed OOD (critical clinical case)**:
  - Best baseline: **Entropy** with AUROC = 0.711 and AUPR-OOD = 0.152.
  - Best advanced method: **Energy** with AUROC = 0.690 and AUPR-OOD = 0.141.
  - Traditional methods like MSP and Mahalanobis performed poorly.
- **Pure OOD (Pneumothorax only)**:
  - **Entropy** again performed well (AUROC = 0.635).
  - **Energy** and **Consistency** showed robust results.
- **In-Distribution (ID) Performance**:
  - Classification performance remained stable (Mean AUC 0.766).
  - OOD methods operate on logits/features without altering the base model.

Analysis. Data augmentation and dropout improved generalization and robustness without compromising classification accuracy. While advanced methods did not significantly outperform baselines, **Entropy** and **Energy**-based scores remained the most reliable. This setup confirms the importance of regularization in clinical OOD detection tasks.

5.4 Experiment 4 – Replacing DenseNet121 with ResNet18

Training Setup. This experiment evaluates the impact of using a simpler architecture (ResNet18) instead of DenseNet121 to reduce overfitting and improve generalization in OOD detection.

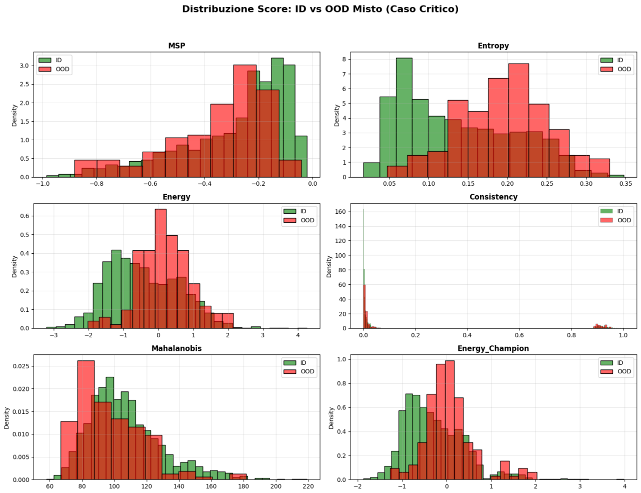


Figure 11: Score Distribution: ID vs Mixed OOD (Critical Case).

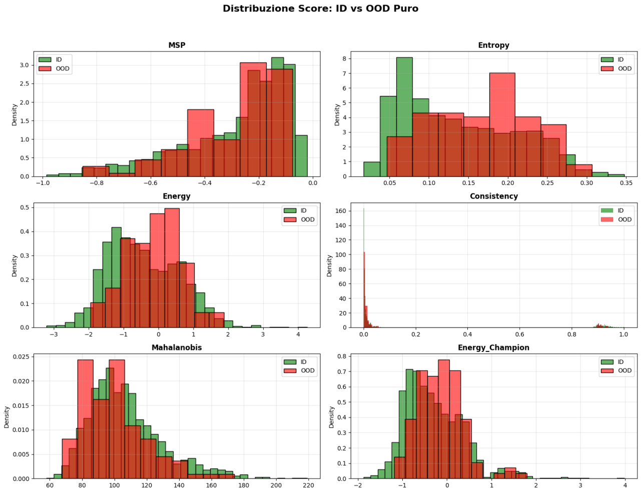
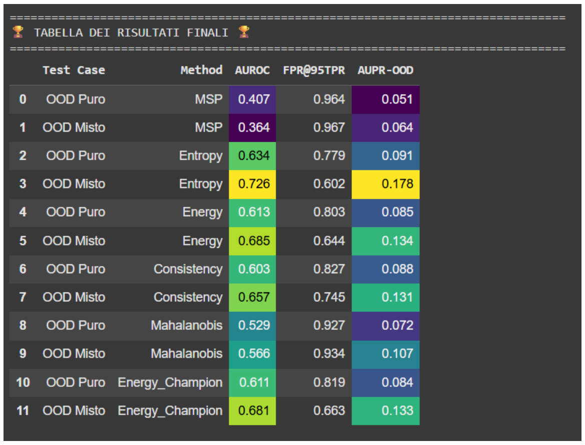


Figure 12: Score Distribution: ID vs Pure OOD (Pneumothorax only).



```
================================================================================
🏆 TABELLA DEI RISULTATI FINALI 🏆
================================================================================
```

	Test Case	Method	AUROC	FPR@95TPR	AUPR-OOD
0	OOD Puro	MSP	0.407	0.964	0.051
1	OOD Misto	MSP	0.364	0.967	0.064
2	OOD Puro	Entropy	0.634	0.779	0.091
3	OOD Misto	Entropy	0.726	0.602	0.178
4	OOD Puro	Energy	0.613	0.803	0.085
5	OOD Misto	Energy	0.685	0.644	0.134
6	OOD Puro	Consistency	0.603	0.827	0.088
7	OOD Misto	Consistency	0.657	0.745	0.131
8	OOD Puro	Mahalanobis	0.529	0.927	0.072
9	OOD Misto	Mahalanobis	0.566	0.934	0.107
10	OOD Puro	Energy_Champion	0.611	0.819	0.084
11	OOD Misto	Energy_Champion	0.681	0.663	0.133

Figure 13: Final Results Table – Experiment 4.

Result Analysis

- **Mixed OOD (critical clinical case)**:
  - Best baseline: **Entropy** with AUROC = 0.726 and AUPR-OOD = 0.166.
  - Best advanced method: **Energy Champion** with AUROC = 0.688 and AUPR-OOD = 0.136.
  - Entropy also achieved the lowest FPR@95TPR = 0.602 among all experiments.
- **Pure OOD (Pneumothorax only)**:
  - **Entropy** again performed well (AUROC = 0.634).
  - **Energy Champion** maintained robust results (AUROC = 0.595).
- **In-Distribution (ID) Performance**:
  - Classification performance remained stable (Mean AUC 0.766).
  - OOD methods operate on logits/features without altering the base model.

Why This Experiment Achieved the Best Results.

- **Reduced Overfitting:** ResNet18 is a simpler model than DenseNet121, which helps prevent overfitting on small datasets (3000 images).
- **Improved AUROC and FPR@95TPR:** Entropy achieved the highest AUROC (0.726) and lowest FPR@95TPR (0.602) on Mixed OOD, outperforming all previous experiments.
- **Model Simplicity:** ResNet18 is computationally efficient and easier to train, making it suitable for clinical deployment.
- **Stable ID Performance:** Despite architectural changes, classification performance on known diseases remained stable (Mean AUC 0.766).

Analysis. Replacing DenseNet121 with ResNet18 led to better OOD detection performance and reduced overfitting, while maintaining classification accuracy. This confirms that simpler models can be more effective in clinical OOD scenarios with limited data.

6 CONCLUSIONS

This work delivers a systematic study of multi-label *hybrid* OOD detection in chest radiography. Using the NIH ChestX-ray14 archive, we framed a clinically realistic setting in which a model must recognise 13 familiar diseases while flagging an unseen one (pneumothorax). The experimental campaign—covering alternative backbones, training regimes and six post-hoc detectors—confirms three takeaways. First, entropy of the output distribution remains the most dependable OOD signal, in line with the uncertainty-centric view promoted by recent surveys [2, 4]. Second, simpler backbones such as ResNet-18 generalise better than heavier DenseNet-121 without sacrificing in-distribution accuracy, echoing observations from robustness studies in other imaging modalities [9]. Third, all detectors preserve baseline diagnostic performance on known findings, demonstrating that OOD safeguards can be integrated without harming clinical utility.

Limitations are the single unseen pathology and the use of one public dataset. Future work should extend the protocol to multiple rare diseases and multi-centre data, and explore ensemble or self-supervised features now emerging in trustworthy-AI literature [1]. By establishing a transparent benchmark and reporting code, we hope to catalyse progress towards AI assistants that can signal uncertainty whenever radiographs depart from their training experience.

REFERENCES

[1] Houssem Ben Braiek and Foutse Khomh. 2025. Machine Learning Robustness: A Primer. In *Trustworthy AI in Medical Imaging*. Elsevier, Chapter 1, 1–44. In press.

[2] Moritz Fuchs, Anastasios N. Angelopoulos, Magdalini Paschali, Christian Baumgartner, and Anirban Mukhopadhyay. 2025. Navigating the Unknown: Out-of-Distribution Detection for Medical Imaging. In *Trustworthy AI in Medical Imaging*. Elsevier, Chapter 4, 73–99. https://doi.org/10.1016/B978-0-44-323761-4.00013-4

[3] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1610.02136

[4] Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han, Weidong Wang, Jialong Xu, Xiaoqi Yang, Hechang Chen, Zhenzhang Li, and Sihong Xie. 2024. Out-of-Distribution Detection in Medical Image Analysis: A Survey. *arXiv preprint arXiv:2404.18279* (2024). https://arxiv.org/abs/2404.18279

[5] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR* abs/1608.06993 (2016). arXiv:1608.06993 http://arxiv.org/abs/1608.06993

[6] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 7167–7177.

[7] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*. 21464–21475. https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html

[8] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Bkg6RiCqY7

[9] Jonas Richiardi, Veronica Ravano, and Nataliia Molchanova. 2025. Domain Shift, Domain Adaptation, and Generalization: A Focus on MRI. In *Trustworthy AI in Medical Imaging*. Elsevier, Chapter 6, 127–151. https://doi.org/10.1016/B978-0-44-323761-4.00015-8

[10] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*. 3462–3471. https://doi.org/10.1109/CVPR.2017.369