Self-Supervised Learning for Out-of-Distribution Detection in Medical Imaging

Nikhileswara Rao Sulake[a], Likith Vinay Kumar Busam[a], Siva Teja Reddy Annapureddy[a] and Sree Harsha Alapati[a]

[a]*Department of Computer Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Nuzvid, Vijayawada, 521202, Andhra Pradesh, India*

ARTICLE INFO

Keywords:
Out-of-Distribution Detection
Chest X-ray Analysis
Self-Supervised Learning
Joint Embedding Predictive Architecture
Masked Autoencoder
Mahalanobis Distance
Model Robustness

ABSTRACT

Out-of-distribution detection represents a fundamental challenge for deploying machine learning systems in clinical environments, where models inevitably encounter data from diverse sources that differ from their training distribution. This study presents a comprehensive comparative analysis of three representation learning paradigms for chest X-ray out-of-distribution detection: Joint Embedding Predictive Architecture (JEPA), Masked Autoencoder (MAE), and traditional supervised learning. Using the TB Chest Radiography Database as in-distribution data and the Montgomery and Shenzhen tuberculosis datasets as out-of-distribution test sets, we systematically evaluate detection performance, cross-dataset generalization, calibration properties, and robustness to common image corruptions. Our experimental results demonstrate that self-supervised methods significantly outperform supervised learning on energy-based out-of-distribution detection, achieving 77-87% AUROC compared to 58-61% for supervised approaches. Furthermore, MAE exhibits exceptional robustness to contrast perturbations, maintaining 98.3% classification AUC under severe corruption where JEPA and supervised methods degrade to 67-76%. Mahalanobis distance-based detection achieves near-perfect performance across all methods, suggesting that distribution-aware scoring can compensate for representation differences. These findings provide practical guidance for deploying medical imaging AI systems with reliable out-of-distribution detection capabilities. The code is publicly available at https://github.com/Nikhil-Rao20/JEPA-Med-OOD.

1. Introduction

The deployment of deep learning models in medical imaging has achieved remarkable success across various diagnostic tasks, transforming how diseases are detected, classified, and monitored [17]. Deep neural networks have demonstrated expert-level performance in applications ranging from disease classification and lesion detection to anatomical segmentation and treatment planning. However, a fundamental challenge threatens the safe clinical deployment of these powerful systems: the inevitable encounter with data that differs from the training distribution.

Out-of-distribution samples in medical imaging arise from multiple sources. Variations in imaging equipment across different hospitals and clinics produce subtle but significant differences in image characteristics. Acquisition protocol differences, including variations in patient positioning, exposure settings, and preprocessing pipelines, further contribute to distribution shift. Patient demographic variations across institutions may introduce appearance differences related to age, body habitus, or disease prevalence. Additionally, pathological presentations not represented in the training data pose particular challenges, as models must recognize when they lack the knowledge to make reliable predictions.

When machine learning models encounter such distributional shifts, they frequently produce overconfident yet incorrect predictions [7, 16]. This failure mode poses significant risks in clinical decision-making, where incorrect diagnoses can lead to inappropriate treatment, delayed intervention, or unnecessary procedures. The consequences of such errors extend beyond individual patient outcomes to broader issues of trust in AI-assisted healthcare systems. Reliable out-of-distribution detection is therefore not merely a technical desideratum but an ethical imperative for trustworthy medical AI deployment.

Self-supervised learning has emerged as a promising paradigm for learning robust visual representations without requiring labeled data [3, 6]. By designing pretext tasks that exploit the inherent structure of visual data, self-supervised methods can learn features that capture semantic content while remaining invariant to superficial variations. Recent architectures such as the Joint Embedding Predictive Architecture and Masked Autoencoders represent two philosophically distinct approaches to representation learning [1, 5]. JEPA predicts latent representations of masked image regions, hypothesized to capture high-level semantic features relevant for downstream understanding. MAE reconstructs pixel-level content from heavily masked inputs, potentially learning more fine-grained visual patterns that generalize across domains.

Despite the theoretical promise of self-supervised representations for out-of-distribution detection, their practical effectiveness in medical imaging remains understudied. Medical imaging presents unique challenges that differ from natural image domains: datasets are typically smaller, class imbalances are common, and the distribution shifts

nikhil01446@gmail.com (N.R. Sulake)
ORCID(s):

of clinical interest may be subtle compared to the dramatic domain shifts studied in natural image benchmarks. Understanding how different pretraining paradigms affect out-of-distribution detection in this specialized domain requires systematic empirical investigation.

This work presents a systematic comparison of JEPA, MAE, and supervised pretraining for chest X-ray out-of-distribution detection. We provide a comprehensive evaluation using standardized Vision Transformer architectures and training protocols, enabling fair comparison across methods. Our evaluation encompasses multiple scoring functions including energy score and Mahalanobis distance, applied to clinically relevant out-of-distribution datasets from different hospital sources. We analyze cross-dataset generalization and robustness to common image corruptions, revealing significant differences between pretraining approaches that have direct implications for clinical deployment. Based on our findings, we identify practical recommendations for deploying medical imaging systems with reliable out-of-distribution detection capabilities.

2. Related Work

2.1. Out-of-Distribution Detection

Out-of-distribution detection has received substantial attention from the machine learning community, with methods broadly categorized into confidence-based, density-based, and distance-based approaches. Each approach offers distinct advantages and limitations that determine their suitability for different application domains.

Confidence-based methods leverage the observation that neural networks tend to produce lower confidence predictions on out-of-distribution samples. The seminal work on baseline detection proposed using the maximum softmax probability as an out-of-distribution score, establishing a simple yet effective approach that has since been extended in numerous directions [7]. Temperature scaling improves separation between in-distribution and out-of-distribution samples by adjusting the softmax temperature during inference [15]. Input perturbation techniques further enhance detection by adding carefully crafted noise that amplifies the confidence gap [10]. Energy-based detection represents a theoretically grounded alternative that uses the energy function derived from the model logits, demonstrating superior performance compared to softmax confidence across diverse benchmarks [11].

Density-based methods attempt to model the distribution of in-distribution data explicitly, using the learned density to identify samples that fall in low-probability regions. Likelihood-based generative models and normalizing flows have been explored for this purpose, though research has demonstrated that likelihood alone is insufficient for reliable detection [18, 20]. Samples from certain out-of-distribution datasets may actually receive higher likelihood than typical in-distribution data, necessitating more sophisticated approaches such as likelihood ratios or typicality measures. Variational autoencoders with reconstruction-based scoring offer an alternative density estimation approach that has been applied to medical imaging with some success.

Distance-based methods measure the relationship between test samples and the training distribution in feature space rather than output space. The Mahalanobis distance approach fits class-conditional Gaussian distributions to features from intermediate network layers and computes the minimum distance to class centroids for each test sample [9]. This approach has proven particularly effective when combined with feature ensemble techniques that aggregate information across multiple network layers. The explicit modeling of feature distributions allows these methods to detect out-of-distribution samples even when classifier confidence remains inappropriately high.

In medical imaging specifically, out-of-distribution detection has been studied for various modalities including chest X-rays, fundus images, pathology slides, and dermoscopy images. However, most prior work has focused on supervised representations, leaving the potential of self-supervised methods largely unexplored in this critical application domain.

2.2. Self-Supervised Learning in Vision

Self-supervised learning has fundamentally transformed visual representation learning by enabling training on vast quantities of unlabeled data through carefully designed pretext tasks. The resulting representations have achieved remarkable success on downstream tasks, often matching or exceeding the performance of supervised pretraining.

Contrastive methods learn representations by maximizing agreement between differently augmented views of the same image while minimizing agreement with views from other images. The SimCLR framework demonstrated that simple contrasting of augmented views, combined with large batch sizes and appropriate projection heads, produces powerful representations [3]. Momentum-based approaches like MoCo address the computational challenges of large batch contrastive learning by maintaining a momentum-updated encoder and memory bank of negative examples [6]. Self-distillation methods such as DINO extend the contrastive framework by training student networks to match teacher predictions without negative examples, producing representations with particularly strong properties for dense prediction tasks [13, 19].

Masked image modeling represents an alternative paradigm inspired by the success of masked language modeling in natural language processing [14]. The Masked Autoencoder architecture masks a large portion of image patches, typically 75%, and trains an encoder-decoder architecture to reconstruct the missing pixel values [5]. The asymmetric design processes only visible patches through the encoder, enabling efficient training while learning representations that capture both local textures and global structure. Variations on this approach have explored different reconstruction targets including discrete visual tokens and feature maps from pretrained networks.

The Joint Embedding Predictive Architecture represents a recent paradigm that predicts representations rather than raw signals [1]. Unlike masked autoencoders which reconstruct pixel values, JEPA predicts the latent embeddings of target image regions from context regions. This approach avoids the pixel-level details that may be irrelevant for downstream understanding, instead focusing learning on semantic features that transfer across tasks. The architecture employs separate context and target encoders with exponential moving average updates for the target encoder, similar to momentum-based contrastive methods.

2.3. Self-Supervised Learning in Medical Imaging

The application of self-supervised learning to medical imaging has gained significant momentum, driven by the chronic scarcity of labeled data in clinical domains [2]. Annotation of medical images requires expert clinicians whose time is limited and expensive, making self-supervised pretraining particularly valuable for leveraging large unlabeled datasets.

Research has demonstrated that momentum contrastive pretraining on chest X-rays improves downstream classification performance compared to ImageNet pretraining, suggesting that domain-specific pretraining captures features more relevant to medical imaging tasks. Contrastive pretraining with multi-instance learning has achieved state-of-the-art results on dermatology and chest X-ray classification, demonstrating the broad applicability of self-supervised approaches across medical imaging domains. More recently, foundation models pretrained on large medical imaging datasets have shown strong transfer learning capabilities across diverse downstream tasks.

For echocardiography specifically, self-supervised methods have been applied to view classification, quality assessment, and cardiac function estimation. Video-based approaches leverage temporal consistency as a supervisory signal, though frame-level pretraining remains common for segmentation tasks where dense predictions are required.

Despite this progress, systematic comparisons of different self-supervised paradigms for medical out-of-distribution detection remain limited. Understanding how JEPA's semantic predictions and MAE's pixel reconstructions affect out-of-distribution detection provides both practical guidance for clinical deployment and theoretical insights into representation learning.

3. Methodology

3.1. Problem Formulation

We consider the problem of training an encoder that produces representations suitable for both downstream classification and out-of-distribution detection. Given an in-distribution dataset containing labeled chest X-ray images with binary classification targets indicating normal or tuberculosis presentation, we aim to learn an embedding function that maps images to a fixed-dimensional representation space. The learned representations should enable both accurate classification of in-distribution samples and reliable detection of samples drawn from out-of-distribution datasets representing different hospital sources and imaging protocols.

The out-of-distribution detection task requires distinguishing test samples that originate from the training distribution versus external distributions not seen during training. Given an out-of-distribution scoring function applied to learned embeddings or classifier outputs, we evaluate detection performance by treating in-distribution samples as positive and out-of-distribution samples as negative, measuring discrimination via standard classification metrics.

3.2. Architecture

All experiments use the Vision Transformer Small architecture to ensure fair comparison across pretraining methods [4]. The encoder consists of 12 transformer blocks with embedding dimension 384, 6 attention heads per layer, and MLP hidden dimension 1536, yielding approximately 21.6 million parameters. Input images are resized to 224 by 224 pixels and divided into non-overlapping 16 by 16 patches, producing 196 patch tokens that are projected to the embedding dimension and combined with a learnable classification token.

Position embeddings are added to patch embeddings to encode spatial location, and layer normalization is applied at the input to each transformer block. The classification token embedding from the final transformer block serves as the image-level representation for downstream classification and out-of-distribution detection.

3.3. Self-Supervised Pretraining Methods

JEPA learns representations by predicting target patch embeddings from context patches in latent space. The architecture consists of three components operating collaboratively: a context encoder that processes visible patches to produce context representations, a target encoder implemented as an exponential moving average copy of the context encoder that processes target patches, and a predictor network that predicts target representations from context representations and learnable mask tokens indicating target positions.

The training objective minimizes the L2 distance between predicted and actual target representations, with stop-gradient applied to the target encoder outputs to prevent representation collapse. We use aggressive context masking of 85-100% with 4 target blocks covering 15-20% each, following the original JEPA recipe. The predictor consists of 6 transformer blocks matching the embedding dimension of the main encoder.

MAE learns representations through masked image reconstruction using an asymmetric encoder-decoder design. The encoder processes only visible patches, reducing computational cost proportional to the masking ratio. With 75% random masking, only one quarter of patches are encoded during training. The decoder is a lightweight transformer that receives encoded visible patches concatenated with learnable mask tokens at target positions, producing reconstructed pixel values for all patches.

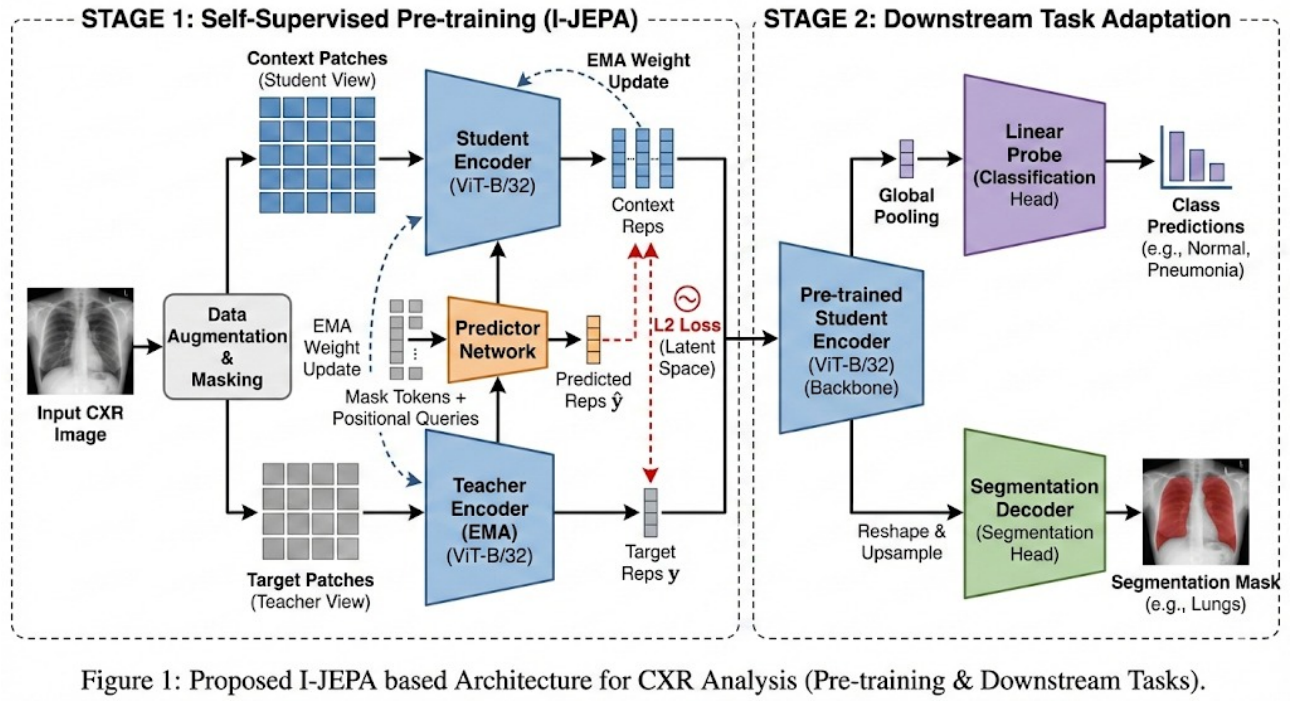


Figure 1: Proposed I-JEPA based Architecture for CXR Analysis (Pre-training & Downstream Tasks).

Figure 1: Overall pipeline of the proposed study. A shared ViT-S encoder backbone is pretrained with JEPA, MAE, and supervised objectives on in-distribution chest X-rays, then evaluated via linear probing, OOD detection with energy and Mahalanobis scores, and robustness analysis under corruption shifts.

The training objective is mean squared error computed on normalized pixel values within each patch, focusing the loss on masked regions. We use 75% random masking with a decoder of 4 transformer blocks and embedding dimension 256, substantially smaller than the encoder following the original MAE design.

Supervised pretraining trains the Vision Transformer encoder end-to-end with a classification head using cross-entropy loss. The classification head is a single linear layer applied to the classification token embedding, producing logits for each class. This approach directly optimizes features for the downstream classification task, providing a strong baseline that achieves high in-distribution accuracy.

3.4. Downstream Evaluation Protocol

For all pretrained encoders, we freeze the encoder weights and train only a linear classifier on extracted features. The classification token representation from the final encoder layer serves as the 384-dimensional image embedding. A linear layer maps these embeddings to class logits, trained with cross-entropy loss on the in-distribution training set while the encoder remains fixed. This linear probing protocol enables fair comparison of representation quality across pretraining methods.

For out-of-distribution detection, we evaluate two complementary scoring functions. The energy score uses the energy function derived from classifier logits, where lower energy indicates higher likelihood of being in-distribution. This approach requires only the trained classifier without additional distribution modeling. The Mahalanobis distance fits class-conditional Gaussian distributions to the embedding space using in-distribution training data, computing the minimum Mahalanobis distance to class centroids as the out-of-distribution score. Higher distance indicates likely out-of-distribution samples. We estimate class means and shared covariance from in-distribution training embeddings.

3.5. Robustness Evaluation Protocol

We evaluate representation robustness to common image corruptions that may occur in clinical imaging settings. Gaussian noise with increasing variance simulates sensor noise from detector electronics or low-dose imaging protocols. Gaussian blur with increasing kernel size simulates defocus or motion artifacts during image acquisition. Contrast perturbation with multiplicative scaling simulates variations in exposure settings, window-level adjustments, or display calibration differences across institutions.

Each corruption is applied at six severity levels from 0 to 5, where level 0 represents clean images and level 5 represents severe corruption. We report classification area under the ROC curve at each severity level to characterize degradation patterns and identify which pretraining methods maintain robust performance under distribution shift.

4. Experiments and Results

4.1. Experimental Setup

Our in-distribution dataset is the TB Chest Radiography Database, a publicly available collection containing 4,200 chest X-ray images with binary labels [12]. The dataset

Table 1
In-Distribution Classification Performance on TB Chest Radiography Database validation set (n=840). All methods achieve strong classification performance, with supervised learning showing slight advantages due to direct optimization for the classification objective.

Method	AUROC	Accuracy	ECE	NLL
JEPA	0.939	0.917	0.029	0.206
MAE	0.993	0.971	0.013	0.085
Supervised	0.996	0.980	0.007	0.064

Table 2
Out-of-Distribution Detection with Energy Score. Self-supervised methods substantially outperform supervised learning, particularly on the Shenzhen dataset where MAE achieves 86.6% AUROC compared to 58.2% for supervised.

	Montgomery		Shenzhen	
Method	AUROC	FPR@95	AUROC	FPR@95
JEPA	0.771	0.540	0.718	0.633
MAE	0.769	0.577	0.866	0.358
Supervised	0.607	0.814	0.582	0.931

includes 700 normal images and 3,500 tuberculosis images, exhibiting the class imbalance typical of clinical screening applications. We use an 80/20 train-validation split, yielding 3,360 training and 840 validation samples for all experiments.

For out-of-distribution evaluation, we use two external tuberculosis chest X-ray datasets representing realistic distribution shifts encountered when deploying models across different healthcare institutions. The Montgomery County dataset contains 138 chest X-rays acquired from Montgomery County, Maryland, USA, using a Eureka stationary X-ray machine. The Shenzhen Hospital dataset contains 662 chest X-rays from Shenzhen No.3 People's Hospital in Guangdong Province, China, acquired on a Philips DR Digital Diagnost system [8]. These datasets differ from our training data in geographic origin, patient demographics, imaging equipment, and acquisition protocols, providing clinically relevant out-of-distribution test cases.

All models use identical Vision Transformer Small architecture and training hyperparameters to ensure fair comparison. Pretraining runs for 50 epochs with batch size 32 using the AdamW optimizer with learning rate 5e-4 and cosine annealing. Data augmentation is limited to random resized cropping with scale 0.3 to 1.0 without horizontal flipping to preserve anatomical consistency. Linear probes train for 100 epochs with learning rate 1e-3. All experiments run on a single GPU with mixed-precision training enabled.

4.2. In-Distribution Classification Performance

Table 1 presents classification performance on the in-distribution validation set across all three pretraining methods.

All three pretraining methods achieve strong in-distribution classification performance, demonstrating that the learned representations successfully capture features relevant for tuberculosis detection. Supervised learning achieves the highest AUROC of 99.6%, which is expected given the direct optimization for the classification objective. MAE follows closely at 99.3% AUROC, indicating that pixel reconstruction pretraining learns highly discriminative features despite never directly optimizing for classification during pretraining. JEPA shows somewhat lower in-distribution performance at 93.9% AUROC, suggesting that its emphasis on predicting semantic features may miss some fine-grained visual patterns relevant for tuberculosis detection.

Calibration metrics paint a consistent picture. Expected calibration error measures the difference between predicted confidence and actual accuracy, with lower values indicating better calibration. Supervised learning achieves the lowest ECE at 0.007, followed by MAE at 0.013 and JEPA at 0.029. Negative log-likelihood follows the same ordering, confirming that supervised learning produces the best-calibrated probability estimates on in-distribution data. These strong in-distribution results across all methods establish that the learned representations are suitable for downstream analysis of out-of-distribution detection capabilities.

4.3. Out-of-Distribution Detection Performance

Table 2 presents out-of-distribution detection results using the energy score, while Table 3 shows results using Mahalanobis distance.

Energy-based out-of-distribution detection reveals striking differences between self-supervised and supervised pretraining paradigms. Self-supervised methods substantially outperform supervised learning on both out-of-distribution datasets, with the performance gap being particularly pronounced on the Shenzhen dataset. On Montgomery, both JEPA and MAE achieve approximately 77% AUROC compared to 61% for supervised learning, representing a 16 percentage point advantage for self-supervised approaches. The difference becomes even more dramatic on Shenzhen, where MAE achieves 86.6% AUROC while supervised learning achieves only 58.2%, barely better than random guessing.

The false positive rate at 95% true positive rate reveals the practical implications of these performance differences. The supervised model's FPR@95 of 93.1% on Shenzhen indicates that when requiring 95% of in-distribution samples to be correctly identified, nearly all out-of-distribution samples are incorrectly classified as in-distribution. This failure mode is particularly dangerous for clinical deployment, where the system would confidently process out-of-distribution samples that should trigger human review.

This performance gap suggests that supervised representations overfit to discriminative features specific to the training distribution, resulting in unreliable confidence estimates when encountering distribution shift. The supervised model has learned to distinguish normal from tuberculosis presentations within the training distribution but has not learned features that capture the broader structure of chest X-ray imaging. In contrast, self-supervised methods, by

Table 3
Out-of-Distribution Detection with Mahalanobis Distance. Distribution-aware scoring achieves excellent performance across all methods, with AUROC exceeding 95% in most cases.

	Montgomery		Shenzhen	
Method	AUROC	FPR@95	AUROC	FPR@95
JEPA	0.956	0.186	0.985	0.077
MAE	0.994	0.030	0.984	0.085
Supervised	0.991	0.024	0.997	0.000

solving pretext tasks that do not depend on class labels, learn representations that better capture distributional differences between in-distribution and out-of-distribution data.

Mahalanobis distance-based detection presents a strikingly different picture. All methods achieve excellent performance, with AUROC exceeding 95% across all method-dataset combinations. This contrasts sharply with energy-based detection, where supervised learning performed poorly. The Mahalanobis approach models the distribution of embeddings explicitly using class-conditional Gaussians rather than relying on classifier confidence, effectively compensating for the limitations of discriminatively-trained representations.

Notably, supervised learning achieves the best Mahalanobis-based detection on Shenzhen with 99.7% AUROC and 0% FPR@95, despite its poor energy-based performance on the same data. This finding suggests that supervised representations do capture information about distribution differences in the embedding space, but this information is not reflected in classifier confidence. The learned decision boundary separates in-distribution classes effectively but does not provide meaningful confidence estimates for samples lying in different regions of the input space.

These results have important practical implications for clinical deployment. Systems using supervised representations should strongly consider Mahalanobis-based scoring rather than confidence-based detection. Self-supervised methods offer more consistent performance across both scoring approaches, providing flexibility in detection methodology. The choice between energy and Mahalanobis scoring depends on computational constraints and the availability of in-distribution feature statistics for Mahalanobis estimation.

4.4. Cross-Dataset Generalization

Table 4 presents classification performance when linear probes trained on the in-distribution dataset are directly applied to out-of-distribution datasets, measuring the generalization of learned features across distribution shift.

Cross-dataset transfer reveals pronounced differences in how well learned representations generalize beyond the training distribution. While all methods achieve similar in-distribution performance between 99.2% and 99.6% AUC, their out-of-distribution classification performance varies dramatically. MAE significantly outperforms both alternatives on out-of-distribution classification, achieving 60.7% AUC on Montgomery and 69.5% AUC on Shenzhen. These results represent 10 to 26 percentage point advantages over JEPA and 14 to 10 percentage point advantages over supervised learning.

Table 4
Cross-Dataset Classification Generalization. MAE demonstrates superior transfer to out-of-distribution datasets, suggesting reconstruction-based pretraining learns more generalizable features.

Method	ID (AUC)	Montgomery (AUC)	Shenzhen (AUC)
JEPA	0.992	0.554	0.430
MAE	0.995	0.607	0.695
Supervised	0.996	0.471	0.592

Table 5
Robustness to Image Corruptions (AUC at Severity Level 5). MAE demonstrates exceptional robustness to contrast perturbations, maintaining 98.3% AUC where other methods degrade to 67-76%.

Method	Clean	Gaussian Noise	Blur	Contrast
JEPA	0.990	0.983	0.981	0.672
MAE	0.992	0.984	0.992	0.983
Supervised	0.996	0.992	0.994	0.758

JEPA shows unexpectedly poor transfer despite strong in-distribution performance. Its out-of-distribution classification of 43.0-55.4% AUC approaches random chance, suggesting that the features learned through latent prediction do not capture disease-relevant patterns that generalize across domains. The masked prediction task may encourage learning context-target relationships specific to the training distribution's structure, encoding patterns that break down under distribution shift. This finding reveals a potential limitation of JEPA for medical imaging applications where deployment must handle domain shift.

MAE's superior transfer likely stems from its reconstruction objective, which requires learning features that capture the intrinsic visual structure of chest X-ray images rather than discriminative boundaries or distribution-specific patterns. By reconstructing masked pixel regions, the model must learn what chest radiographs look like in general, including anatomical structures, tissue densities, and pathological patterns. These general features transfer more readily to new domains where the fundamental image characteristics remain consistent despite differences in acquisition protocols.

4.5. Robustness to Image Corruptions

Table 5 presents classification AUC at maximum corruption severity, revealing how different pretraining methods handle realistic image degradations that may occur in clinical imaging pipelines.

All pretraining methods maintain near-baseline performance under Gaussian noise and blur corruptions, with less than 2% degradation in classification AUC at maximum

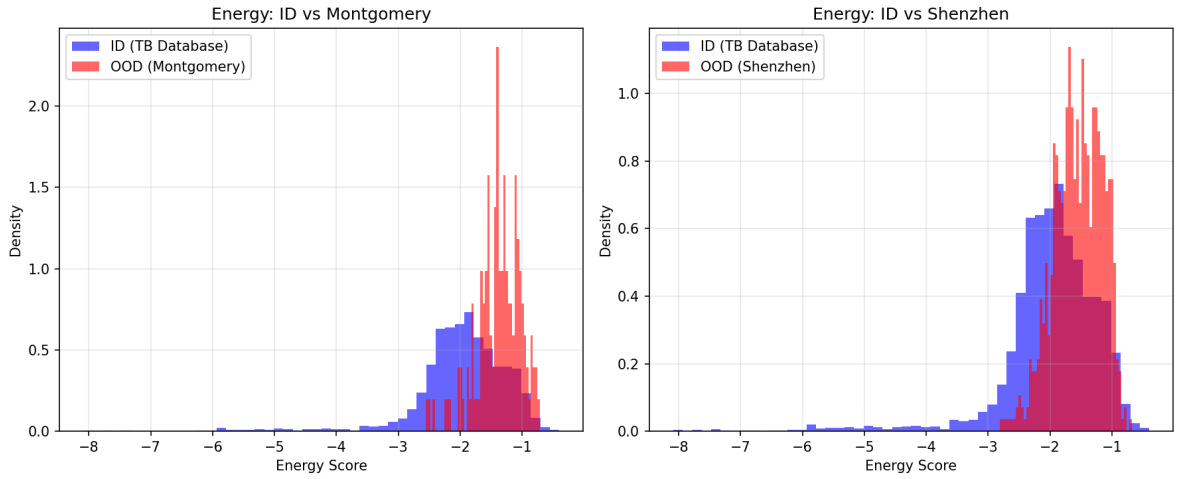


Figure 2: Energy score distributions for in-distribution and out-of-distribution samples across all pretraining methods. Self-supervised methods show better separation between distributions, enabling more reliable detection.

Cross-Dataset Classification (Linear Probe trained on ID only, evaluated on OOD)

Method	ID Val AUC ↑	ID Val Acc ↑	Montgomery AUC ↑	Montgomery Acc ↑	Shenzhen AUC ↑	Shenzhen Acc ↑
JEPA	0.992	0.977	0.554	0.580	0.430	0.494
MAE	0.995	0.979	0.607	0.572	0.695	0.569
Supervised	0.996	0.982	0.471	0.572	0.592	0.483

Figure 3: Cross-dataset classification performance across pretraining methods. MAE demonstrates consistent performance advantages when transferring to out-of-distribution datasets from different hospital sources.

severity. This resilience suggests that the learned representations are naturally robust to additive noise and defocus artifacts, likely because these corruptions do not fundamentally alter the spatial structure or diagnostic content of chest X-ray images.

However, contrast perturbations reveal a dramatic vulnerability in JEPA and supervised representations. Under severe contrast shifts at severity level 5, JEPA degrades from 99.0% to 67.2% AUC, representing a 22 percentage point performance collapse. Supervised learning similarly degrades from 99.6% to 75.8% AUC. These substantial performance drops indicate that these representations encode contrast-sensitive features that become unreliable when the intensity distribution shifts.

In stark contrast, MAE maintains 98.3% AUC under the same severe contrast perturbation, representing less than one percentage point decline from clean image performance. This exceptional contrast robustness represents a significant practical advantage for clinical deployment, where imaging protocol variations across institutions commonly manifest as differences in exposure, window-level settings, and display calibration.

The contrast robustness advantage likely stems directly from MAE's reconstruction objective. Reconstructing masked image regions requires learning the intensity distributions and appearance statistics of training images across all brightness levels. The model must predict pixel values in both bright and dark regions, implicitly learning features that are robust to global intensity transformations. JEPA's latent prediction and supervised learning's discriminative training do not impose such constraints, allowing the learned features to encode contrast-sensitive patterns that provide discriminative value on in-distribution data but break down under contrast shift.

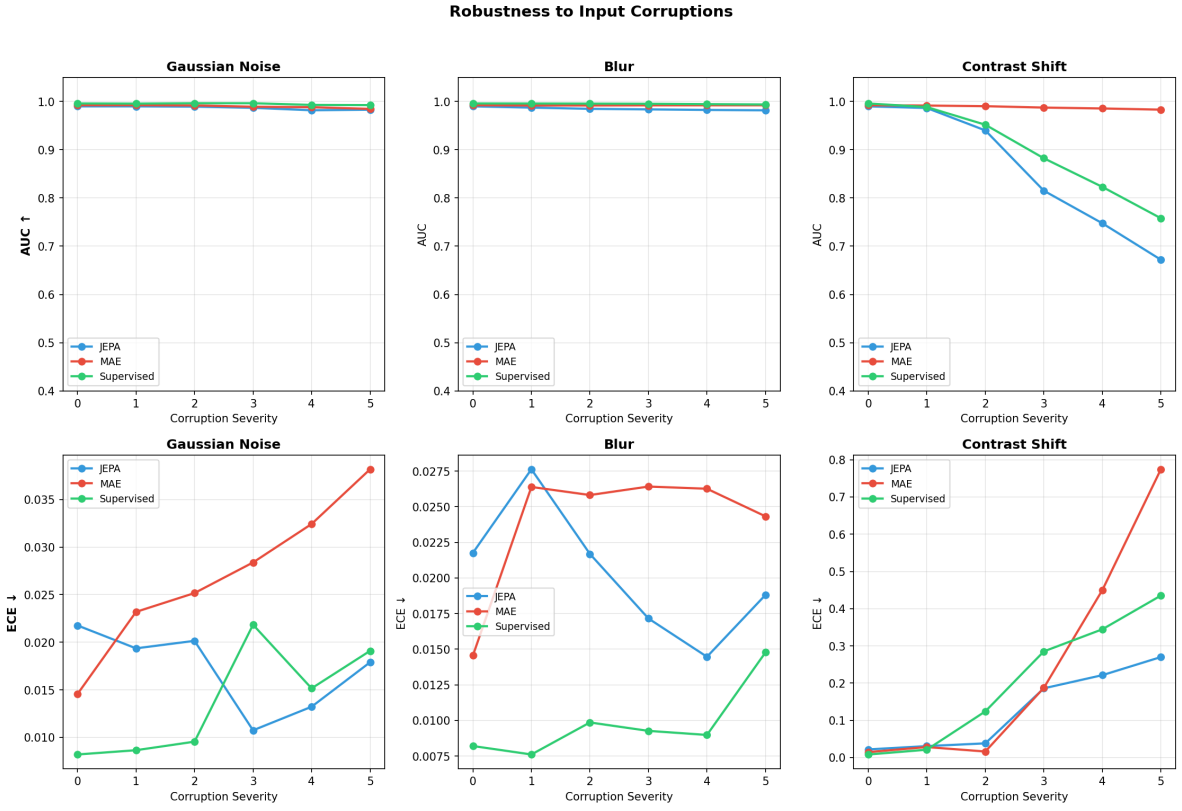
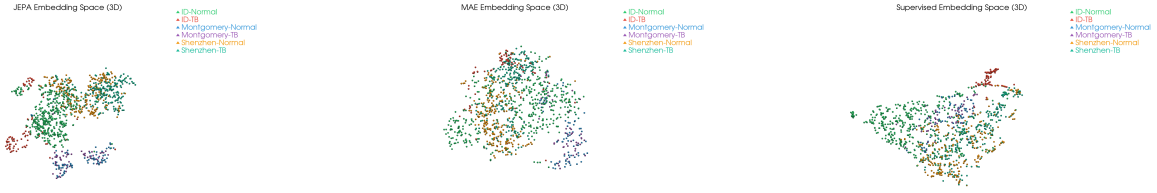


Figure 4: Classification AUC versus corruption severity for all corruption types. MAE maintains stable performance across all severity levels, while JEPA and Supervised methods show progressive degradation under contrast perturbations.



(a) JEPA (b) MAE (c) Supervised

Figure 5: Three-dimensional t-SNE visualizations of embedding spaces for each pretraining method. Points are colored by dataset source and disease label, revealing different geometric organizations learned by each approach.

4.6. Embedding Space Analysis

Figure 5 visualizes the learned embedding spaces using t-SNE dimensionality reduction, revealing qualitative differences in how each pretraining method organizes representations.

The embedding visualizations reveal qualitatively different representations learned by each pretraining method. Supervised learning produces the most compact and well-separated clusters for in-distribution data, consistent with its superior in-distribution classification accuracy. However, the out-of-distribution samples from Montgomery and Shenzhen datasets do not form distinct clusters in the supervised embedding space. Instead, they appear mixed with in-distribution samples, explaining the poor energy-based out-of-distribution detection: the classifier cannot distinguish between in-distribution and shifted samples because they occupy similar regions of embedding space.

MAE's embedding space shows more complex structure. While class separation exists for in-distribution data, out-of-distribution samples tend to occupy systematically different regions of the space. This geometric organization enables effective Mahalanobis-based detection, as the centroid-based distance metric captures the displacement of out-of-distribution points from in-distribution clusters. The reconstruction objective appears to encode information about image source and acquisition characteristics that manifests as geometric separation in embedding space.

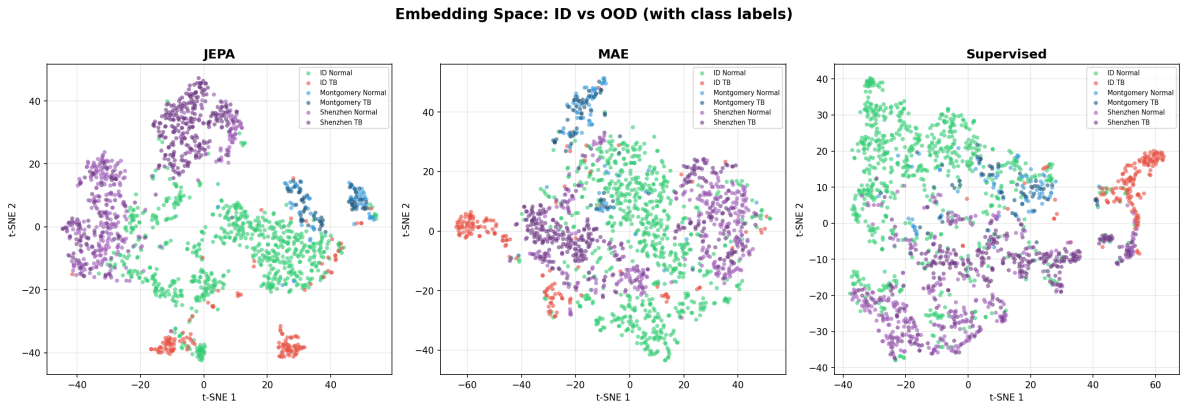


Figure 6: Detailed embedding visualization showing the distribution of samples from in-distribution and out-of-distribution datasets across the learned representation spaces.

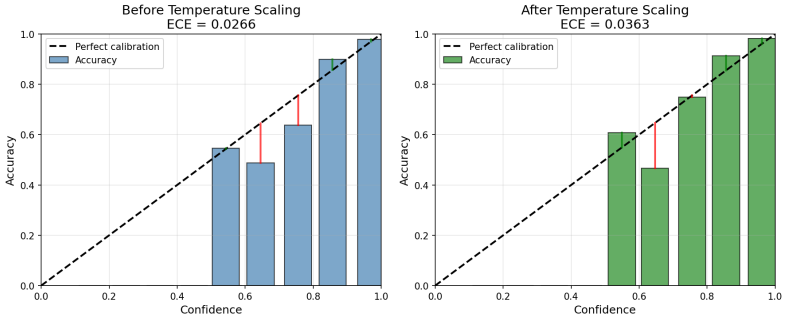


Figure 7: Reliability diagram showing calibration of probability estimates. Perfect calibration would follow the diagonal line. Supervised learning shows the best calibration on in-distribution data.

JEPA's embeddings show intermediate characteristics with reasonable class separation but less distinct organization of out-of-distribution samples. The embedding geometry provides some discrimination but lacks the clear structural separation seen in MAE, which may explain JEPA's moderate out-of-distribution detection performance between self-supervised MAE and supervised learning.

5. Limitations

While this study provides comprehensive insights into self-supervised learning for medical out-of-distribution detection, several limitations should be acknowledged when interpreting the results and considering their applicability to clinical deployment.

Our experiments focus on a single medical imaging modality with a specific clinical task, namely chest X-ray imaging for tuberculosis detection. The relative performance of JEPA, MAE, and supervised learning may differ for other medical imaging domains such as CT, MRI, ultrasound, or pathology, where image characteristics and relevant features differ substantially. The TB Chest Radiography Database exhibits severe class imbalance with 700 normal samples versus 3,500 tuberculosis samples, which may affect classifier calibration and performance estimates. Additionally, the out-of-distribution test datasets are relatively small with 138 and 662 samples respectively, limiting statistical power for detecting subtle performance differences between methods.

The distribution shifts we evaluate represent a specific type of domain shift arising from differences in hospital source and imaging equipment. Other clinically relevant distribution shifts including the appearance of novel pathologies not seen during training, protocol variations within the same institution, and demographic differences in patient populations may exhibit different detection characteristics. The relationship between method performance on equipment-based domain shift versus other shift types remains unexplored.

All experiments use Vision Transformer Small architecture to enable fair comparison, but the effects of model scale on representation learning and out-of-distribution detection are not investigated. Larger architectures such as ViT-Base and ViT-Large may exhibit different dynamics, potentially amplifying or mitigating the performance differences we observe. Foundation models pretrained on millions of diverse medical images may also show different behavior than models pretrained on our limited in-distribution dataset.

The self-supervised models in our study are pretrained on the same data used for downstream evaluation, reflecting a realistic scenario where out-of-distribution data is unavailable during training. However, larger pretraining datasets from diverse sources could improve representation quality

and potentially change the relative performance ordering of methods. Access to more diverse pretraining data might particularly benefit methods like JEPA that appear to learn distribution-specific features.

6. Conclusion

This work presents a systematic comparison of self-supervised learning methods for chest X-ray out-of-distribution detection, evaluating JEPA, MAE, and supervised pretraining across multiple dimensions relevant to clinical deployment. Our comprehensive evaluation across out-of-distribution datasets, scoring methods, generalization tests, and robustness analysis yields several findings with direct practical implications.

Self-supervised pretraining provides substantial advantages for confidence-based out-of-distribution detection. Using energy scoring, JEPA and MAE outperform supervised learning by 15 to 25 percentage points in AUROC, achieving detection rates of 77-87% compared to 58-61% for supervised approaches. This advantage stems from representations that capture distributional characteristics beyond discriminative class boundaries. For clinical deployments requiring confidence-based out-of-distribution flagging, self-supervised pretraining should be strongly preferred over supervised learning.

Mahalanobis distance-based detection achieves excellent performance regardless of pretraining paradigm, with all methods exceeding 95% AUROC. This finding suggests that explicit distribution modeling in embedding space can compensate for representation limitations, providing a reliable detection path even for supervised models. The practical implication is that appropriate detection methodology may be as important as representation choice for reliable deployment.

MAE demonstrates exceptional robustness advantages over both JEPA and supervised learning. Under severe contrast perturbations, MAE maintains 98% classification AUC while other methods degrade to 67-76%. MAE also shows superior cross-dataset generalization, outperforming alternatives by 10-26 percentage points when classifying out-of-distribution samples. These robustness properties are particularly valuable for medical imaging, where protocol variations across institutions commonly involve contrast and exposure differences.

Based on these findings, we recommend MAE pretraining combined with Mahalanobis-based out-of-distribution detection as a robust baseline for medical imaging deployment when distribution shift is expected. JEPA shows promise for energy-based detection but exhibits sensitivity to distribution-specific features that limits its transfer capabilities. Supervised learning achieves the strongest in-distribution performance but requires careful attention to detection methodology to avoid overconfident predictions on shifted data.

Our immediate next step extends this framework to echocardiogram segmentation using EchoNet-Dynamic for in-distribution training and EchoNet-Pediatric for out-of-distribution evaluation. This extension will test whether the patterns observed in chest X-ray classification generalize to video-based medical imaging and dense prediction tasks. Additional future directions include scaling to larger Vision Transformer architectures, investigating temporal self-supervised methods for video understanding, combining multiple scoring methods for improved detection, and validation on larger multi-site clinical datasets with diverse distribution shifts.

References

[1] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629.

[2] Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488.

[3] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[5] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

[6] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

[7] Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

[8] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., and Thoma, G. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475.

[9] Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177.

[10] Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

[11] Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475.

[12] Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. B., et al. (2020). Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601.

[13] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

[14] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[15] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

[16] Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.

[17] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

[18] Nalisnick, E., Matsukawa, A., Teh, Y. W., Goro, D., and Lakshminarayanan, B. (2019). Do deep generative models know what they don't know? In *International Conference on Learning Representations*.

[19] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

[20] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M., Dillon, J. V., and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14707–14718.